



Coarse-to-fine visual representation learning for medical images via class activation maps

Boon Peng Yap*, Beng Koon Ng

School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore
Centre for OptoElectronics and Biophotonics, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore

ARTICLE INFO

Keywords:

Weakly supervised learning
Contrastive learning
Class activation map
Fundus
X-ray

ABSTRACT

The value of coarsely labeled datasets in learning transferable representations for medical images is investigated in this work. Compared to fine labels which require meticulous effort to annotate, coarse labels can be acquired at a significantly lower cost and can provide useful training signals for data-hungry deep neural networks. We consider coarse labels in the form of binary labels differentiating a normal (healthy) image from an abnormal (diseased) image and propose CAMContrast, a two-stage representation learning framework for medical images. Using class activation maps, CAMContrast makes use of the binary labels to generate heatmaps as positive views for contrastive representation learning. Specifically, the learning objective is optimized to maximize the agreement within fixed crops of image-heatmap pair to learn fine-grained representations that are generalizable to different downstream tasks. We empirically validate the transfer learning performance of CAMContrast on several public datasets, covering classification and segmentation tasks on fundus photographs and chest X-ray images. The experimental results showed that our method outperforms other self-supervised and supervised pretrain methods in terms of data efficiency and downstream performance.

1. Introduction

Transfer learning via the pretraining and fine-tuning paradigm is one of the most popular approaches in training deep neural networks for medical imaging tasks. In a typical transfer learning setup, a neural network is first initialized with weights pretrained on a large generic dataset before fine-tuning on downstream tasks with smaller datasets. Beyond the natural image domain, ResNet [1] models trained on the ImageNet dataset [2] in a fully supervised manner have been widely adopted as pretrained weights for a diverse range of medical imaging tasks, such as diabetic retinopathy grading [3] and thorax diseases classification [4]. Compared to random initialization, these pretrained weights can often improve convergence speed and reduce the number of manual annotations required in downstream tasks; the latter advantage is especially important for the medical imaging domain, as annotations are prohibitively expensive to acquire due to the requirement of domain-specific knowledge.

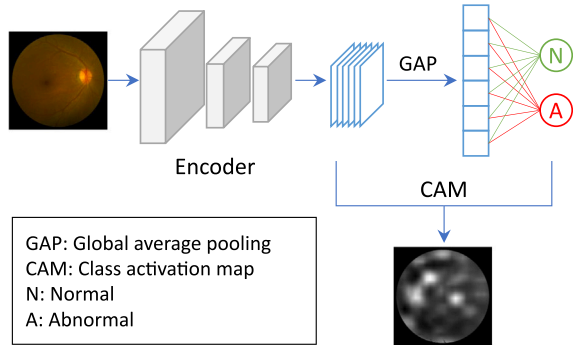
Apart from fully supervised pretraining, there is a recent push for self-supervised pretraining methods which seeks to extract generalizable representations without requiring human-annotated samples. The appeal of learning from potentially unlimited annotation-free data has attracted a lot of interest from both natural and medical imaging community. Notable methods include image restoration [5–9],

rotation angle prediction [10,11], contrastive learning [12–14], and other consistency-based methods [15–17]. SimCLR [12] popularizes the contrastive learning objective for self-supervised pretraining on image data. In medical imaging, restorative methods are among the most popular approaches for self-supervised learning, and were shown to be complementary to discriminative and adversarial learning [6].

In scenarios where coarsely-annotated data are available, the annotations could be further exploited to guide the representation learning process to produce pretrained weights that can be transferred to downstream tasks with limited fine-grained annotations. In this work, a two-stage coarse-to-fine representation learning framework for medical imaging tasks is investigated. Specifically, the proposed learning framework utilizes coarse labels in the form of image-level annotations distinguishing between normal (health) and abnormal (diseased) images. Compared to fine-grained labels such as specific type or severity of diseases or pixel-wise delineations of diseased regions, coarse labels require significantly less effort to acquire, and could also be automatically mined from radiological reports [4]. Intuitively, coarse labels can provide training signals to enforce certain structures in the representation space (e.g., one that linearly separates normal and abnormal images) that is beneficial for the downstream tasks. Thus, when transferred to downstream tasks, neural networks pretrained

* Corresponding author at: School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore.
E-mail addresses: boonpeng001@e.ntu.edu.sg (B.P. Yap), ebkng@ntu.edu.sg (B.K. Ng).

Stage 1: Train heatmap generator



Stage 2: Learn fine-grained representation

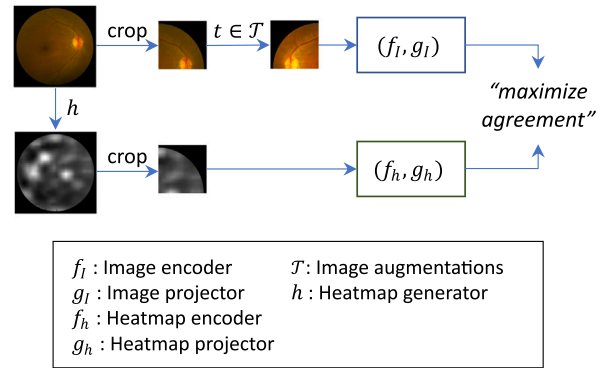


Fig. 1. Overview of the proposed CAMContrast framework. In stage 1, a classifier is trained on the coarse-grained abnormality labels and CAM is used to extract the spatial information of the abnormal regions. The trained classifier is then frozen and used as heatmap generator in the next stage. In stage 2, an image encoder (f_I) is trained to maximize the agreement within image-heatmap pairs in the representation space. After pretraining, f_I is used to initialize the encoder for downstream tasks.

with coarse labels should outperform their self-supervised counterparts [18,19]. Although supervised contrastive objective was shown to be effective in datasets like ImageNet with 1000 classes, it might perform poorly when applied to coarse datasets with only two classes (normal vs. abnormal). Supervised contrastive objective treats different images with the same coarse labels as positive pairs and will try to pull their representations close to each other during pretraining. This causes the network to be invariant towards fine-grained details that are required to distinguish between different types of diseases and lesions. The issue is further exacerbated by the commonly used random crop augmentation strategy, where some of the crops are not in agreement with their image-level labels. For example, crops sampled from images labeled as abnormal might not contain any abnormal regions. This will result in a substantial amount of false positive pairs that are detrimental to representation learning.

To preserve fine-grained details while learning from coarse labels, a novel representation learning framework called CAMContrast is proposed in this work. As shown in Fig. 1, CAMContrast consists of two training stages. In stage one, a convolutional neural network (CNN) is trained to classify normal and abnormal images. The trained CNN is then repurposed as a view generator for stage two, in which grayscale heatmaps are derived from class activation maps (CAM) [20]. These heatmaps (with values ranging from 0 to 1) provide spatial information of how much a subregion contributes to the discrimination of the abnormal class, and thus can serve as fine-grained views for representation learning. By keeping the view generator fixed and training a new network via contrastive learning within image-heatmap pairs, the issues encountered in coarse-grained supervised contrastive learning (i.e., loss of fine-grained details and false positive pairs) are circumvented. Stage two is where the actual representation learning takes place: a new network is trained to maximize the agreement between an augmented view of an image crop and the corresponding heatmap via supervised contrastive learning. This is in marked contrast to the typical contrastive learning setup [12,21], which maximizes agreement between two augmented views of the same image. In addition, a fixed crop training strategy is introduced to exploit inter-image similarities of subregions from the same relative position. The efficacy of CAMContrast is demonstrated in a series of transfer learning experiments on multiple datasets of fundus photographs and chest X-ray images. The experimental results show that our proposed framework learns generalizable representations from coarse-grained datasets and improves performance in a wide range of downstream tasks. We release the source codes and pretrained models at <https://github.com/BPYap/CAMContrast>.

The main contributions of this work include:

1. Introduction of a two-stage framework for learning fine-grained representation from coarse labels.
2. Formulation of fixed crop contrastive learning objective using image-heatmap as positive training pairs.
3. Extensive empirical studies on fundus and chest X-ray datasets to verify the effectiveness of CAMContrast.

2. Related work

2.1. Weakly supervised learning

Weakly supervised learning alleviates the cost of acquiring precise annotations by incorporating weaker annotations into the learning routine. Precise annotations such as object masks require tremendous effort to collect [22] as each pixel needs to be individually classified. Consequently, previous works have explored ways to utilize weaker annotations as additional supervision for segmentation tasks. The most popular approach involves creating pseudo masks from bounding boxes [23, 24] or image-level labels [25–27]. When multiple types of labels exist, deep supervision [28] together with mean-teacher [29] were shown to be able to significantly cut down the requirement for precise labels [30]. Outside of segmentation tasks, weak annotations in the form of hash tags [31], hierarchical labels [32] and attribute labels [19] have also been used to improve performance on downstream classification tasks with fine-grained labels. When evaluating downstream performance using weights pretrained on weak labels, many recent approaches have only considered the dataset-specific case, in which the downstream tasks share the same training data as the pretrain task. In this work, we explore whether abnormality labels can be used to learn generic representations for medical imaging tasks, and consider a more comprehensive transfer learning setup where the downstream tasks may have different distribution compared to the pretraining data.

2.2. Contrastive learning

Recent advances in visual representation learning methods are largely based on the idea of contrastive learning [33,34], which seeks to minimize the distance between two semantically similar images (positive pairs) in a shared projection space while maximizing distances for semantically different images (negative pairs). In self-supervised learning where labels are not available, the positive pairs usually consist of different augmented views of the same image and the learning objective is to discriminate each image (instance) from other training images [12,13]. If label information is available, the positive pairs

can be defined as instances that shared the same labels. Some examples of label information used in contrastive learning include object category [21], image metadata [14,35,36] and image caption [37]. Closely related to our work are multi-modal self-supervised learning for fundus photographs [38] and lesion-based contrastive learning [18]. The former synthesizes complementary views for fundus images by training a CycleGAN [39] model on a dataset consisting of fundus images with matching fluorescein angiography images, while the latter sample positive pairs using an object detector trained on bounding box annotations. By contrast, our proposed framework generates heatmaps as positive views from a classification model trained only on coarse image-level labels.

2.3. Class activation map

First proposed as a technique to localize class-specific image regions for classification-trained CNN, class activation maps (CAM) [20] have been widely used to visually explain the decision process of a CNN and to generate starting labels for weakly supervised dense prediction tasks [25–27,40–43]. In essence, CAM is generated by taking the class-weighted average of feature maps extracted at the last convolutional layer of a CNN that precedes a global average pooling (GAP) layer. Other variants such as Grad-CAM [44] has been introduced to allow CAM generation from any CNN-based network. In this work, we propose to exploit the localization capability of CAMs by using the CAM-derived heatmaps as alternative views for the training images in a multiview contrastive learning setup [45]. By maximizing agreement within the image-heatmap pairs, we posit that the continuous values of heatmaps can help guide the contrastive learning process to focus on the fine-grained details and learn representations that are generalizable to different downstream tasks.

2.4. Coarse-to-fine learning

Prior work in coarse-to-fine learning has largely focused on a specific task setting. For example, coarse-to-fine learning has been studied in the problem of image retrieval [46,47], where learning objectives catered to image retrieval are proposed to learn fine-grained information from coarse-grained labels. Beyond image retrieval, specialized algorithms have been introduced to build a semantic tree from flatten fine-grained labels [48], or to learn visio-linguistic compositionality from coarse textual labels [49]. In this work, we focus on representation learning from coarse labels, where existing methods for image retrieval or compositionality learning are not applicable for image-level visual representation learning. Our work is most related to recent work in weakly supervised learning, namely Cl-InfoNCE [19] and CCL-K [50]. Cl-InfoNCE uses clustering to group images with similar set of coarse labels while CCL-K used similarity kernels to compute loss weightages. Similar to CAMContrast, both Cl-InfoNCE and CCL-K are based on the contrastive learning framework. However, unlike CAMContrast, Cl-InfoNCE and CCL-K address the issue of sparse positive pairs when there is a large set of coarse labels. In contrast, we study the other extreme, i.e., coarse labels with only two classes. Under this setting, Cl-InfoNCE and CCL-K are equivalent to the supervised contrastive learning objective.

3. Method

The proposed CAMContrast framework involves training two neural networks in a two-stage learning process. The first stage trains a heatmap generator on coarse labels which outputs positive views for contrastive representation learning in the second stage.

3.1. Heatmap generator

The purpose of a heatmap generator is to generate grayscale images (heatmaps) that emphasize the discriminative regions of the abnormal class. They can be derived from CAMs [20] which are computed as the class-weighted average of the feature maps at the last convolutional layer (that precedes a GAP layer):

$$M^c(x) = \sum_{i=1}^K w_i^c F_i(x), \quad (1)$$

where K is the number of feature maps, w_i^c is the classifier weight connecting the i th GAP unit to the output neuron of class c , and F_i is the i th feature map. Let $c = 0$ be the normal class and $c = 1$ be the abnormal class, the CNN classifier is trained to minimize the following loss function, \mathcal{L} :

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B [\mathcal{L}_{ce}(x_i, y_i) + \lambda \mathcal{L}_{attn}(x_i, y_i)] \quad (2)$$

$$\mathcal{L}_{ce}(x, y) = - \sum_{c=0}^1 \mathbb{1}(y = c) \log p^c(x) \quad (3)$$

$$\mathcal{L}_{attn}(x, y) = \frac{\mathbb{1}(p^0(x) \geq 0.5, y = 0)}{|M^0(x)|} \sum_{i,j} (1 - M_{i,j}^0(x) + M_{i,j}^1(x)), \quad (4)$$

where B is the batch size, x and y are the input images and ground truth labels, respectively, λ is a weighting parameter which is set to 0.01 in all experiments, $\mathbb{1}$ is an indicator function, $p^c(x)$ is the predicted probability for class c , and $M_{i,j}^c(x)$ is the (i, j) element of the CAM extracted for class c . \mathcal{L}_{ce} is a standard cross-entropy function while \mathcal{L}_{attn} is the complementary guided attention loss [51]. The latter provides regularization on the CAMs of normal images by simultaneously encouraging M^0 to expand the whole images and minimizing the areas covered by M^1 .

To generate views for pretraining, the CAMs are converted into heatmaps via min-max normalization. The heatmaps would normally be upscaled and overlaid on top of the input images to visualize the discriminative regions. In this work, however, the pretraining heatmaps are kept at their original size (same size as the feature maps) to save memory and computation time since they are not used for visualization purpose. The crop-level dataset was pre-generated before stage 2 pretraining. First, the image-level heatmaps ($W_h \times H_h$ pixels) are obtained via Eq. (1) and min-max normalization from the input images ($W_i \times H_i$ pixels). Then, each image and its corresponding heatmap are divided into four non-overlapping crops (top-left, top-right, bottom-left, bottom-right) and one center crop. After cropping, the image crops and heatmap crops have a resolution of $\frac{W_i}{2} \times \frac{H_i}{2}$ pixels and $\frac{W_h}{2} \times \frac{H_h}{2}$ pixels, respectively.

3.2. Contrastive representation learning

The heatmaps generated from stage one contains spatial information on regions that are likely to be abnormal and can be regarded as alternative views of the original images. They can therefore be utilized as a guidance for learning abnormality-aware representations. Under the contrastive learning framework, image-heatmap pairs are treated as positive pairs and their representations are encouraged to be close to each other in a shared latent space. As the images and heatmaps have different sizes and channels, vector representations of the images and heatmaps are computed via two separate CNN encoders, f_I and f_h , and projected to a shared latent space via two MLP projectors, g_I and g_h , where the contrastive learning objective, \mathcal{L}_{con} , is optimized:

$$\mathcal{L}_{con}(z_i, z_j) = -\log \frac{\exp(z_i^T z_j / \tau)}{\sum_{k=1}^N \mathbb{1}(i \neq k) \exp(z_i^T z_k / \tau)}, \quad (5)$$

where (i, j) and (i, k) are a pair of positive and negative examples, respectively, N is the number of projection vectors in a mini-batch, z

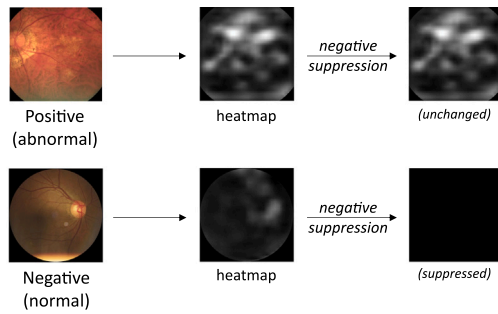


Fig. 2. Illustration of the negative suppression technique. After applying negative suppression, heatmaps derived from positive (abnormal) images remained unchanged while heatmaps derived from negative (normal) images are zeroed out.

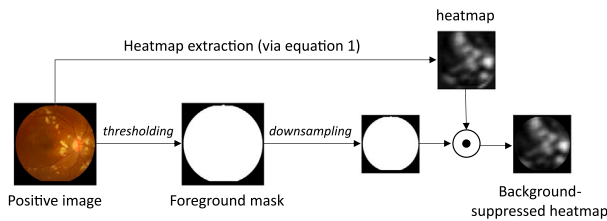


Fig. 3. Illustration of the background suppression process on positive heatmaps. The threshold value is set to 15 for both retinal fundus photographs and chest X-rays.

is the normalized projection vectors and τ is a temperature parameter. The loss objective has the same form as the self-supervised formulation [12,13], the main difference being how the views are constructed. Self-supervised methods typically require two augmented views of the same image as positive views, while the proposed CAMContrast method uses one augmented view and one heatmap view, which means the projection vectors (z) may contain vectors from either the image or heatmap projector.

Negative suppression. While the attention-based regularization (Eq. (4)) has largely suppressed the activations of heatmaps in the negative (normal) class, the actual heatmaps generated by the heatmap generator might still consist of some highly activated regions, which is not ideal since they should be completely empty for normal images. As shown in Fig. 2, we introduce a post-processing step for the normal heatmaps by simply resetting their activations to zero before computing the projection vectors. Consequently, all normal images will be pulled towards a prototypical vector of an empty heatmap when Eq. (5) is being optimized. Our experimental results (Section 4.3) show that this simple negative suppression technique is beneficial for downstream performance. Additionally, the background activations (i.e., values outside the bodily regions) in the abnormal heatmaps are also masked out using foreground masks obtained by thresholding the input images. Concretely, the foreground masks are obtained via thresholding the input images before cropping. As shown in Fig. 3, the obtained foreground mask is downsampled to the resolution of the heatmap and multiplied with the heatmap (element-wise) to obtain the background-suppressed heatmap.

Fixed crop training. Compared to images from the natural image domain, medical images are largely consistent in terms of overall appearance. They share some common structures among crops sampled from the same relative position, as shown in Fig. 4. Based on this observation, we propose to treat each crop as a distinct instance by dividing each image into five crops consisting of four non-overlapping corner crops and one center crop. Instead of sampling two random crops within each image as positive pairs, augmented views are generated at

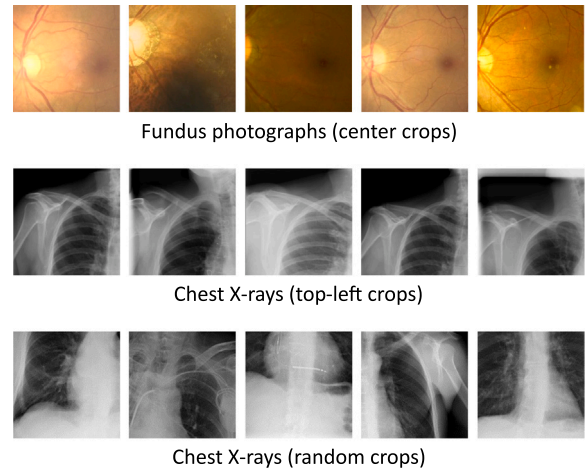


Fig. 4. Visual comparison of fixed crops (first two rows) with random crops (third row). The fixed crops are sampled at the same relative position from the original image.

the crop-level such that the views of two crops sampled at different positions (e.g., top-left vs. center) relative to the whole image are treated as negative pairs even if they belong to the same image. Following the standard practice of contrastive learning, standard augmentations including random cropping, random horizontal flipping, rotation, color jittering and random grayscaleing are applied to each cropped image, while no random augmentation is applied to the heatmap crops. By maximizing the agreement between the representations of the randomly augmented image crop with the heatmap representations, the desirable augmentation invariance property is induced in the trained image encoder, making it more robust for downstream transfer learning. In other words, the criteria of positive pairs in fixed crop training require each pair of image-heatmap crops to have the same relative position, on top of being originated from the same source image (i.e., the image where the heatmap is derived from). This will allow the network to better exploit the inter-instance similarities found within the same relative position. The cropping strategy forces the network to pick up fine-grained details that distinguishes among different instances with similar appearance, also known as hard negatives pairs [52,53]. This results in a network that is invariant towards redundant features shared across the negative pairs. To reduce computation overhead, the image-heatmap crops are pre-generated before pretraining. This produces a crop-level dataset that is five times larger than the image dataset. Within each minibatch, each image crop is sampled together with its heatmap crop to ensure that every instance has a matching positive pair.

3.3. Practical considerations of two-stage design

A pseudocode of CAMContrast is given in Algorithm 1. Our two-stage design of CAMContrast is motivated by three factors: training stability, hyperparameter complexity, and computation overhead. A single-stage design which jointly optimizes the heatmap and contrastive loss would introduce additional hyperparameters, such as a warmup term to ensure training does not diverge at earlier epochs, and multiple balancing terms to balance the contribution of individual training objective. This drastically increases the search space and complicates the pretraining process, as multiple hyperparameters now have to be tuned from scratch for each imaging modalities. In contrast, our two-stage design decouples the training of heatmap generator and image encoder. By training the heatmap generator to convergence in stage 1, the generated heatmaps will be consistent in approximating the locations of the abnormal regions for stage 2 training. Furthermore,

Table 1

Details of the datasets used in transfer learning experiments. Abbreviations: DR - diabetic retinopathy; DME - diabetic macular edema; OD - optic disc; OC - optic cup.

Dataset	Task	# train	# val	# test	Metrics
IDRiD-cls [54]	Joint DR-DME classification	331	82	103	Joint accuracy ^b
IDRiD-seg [54]	Retinal lesions segmentation	44	10	27	Avg. AUC-PR
REFUGE-cls [55]	Galucoma classification	400	400	400	F1-score
REFUGE-seg [55]	Joint OD-OC segmentation	400	400	400	Avg. F1-score
Vessel-seg ^a	Retinal vessel segmentation	40	10	38	F1-score
ChestX-ray14 [4]	Thorax disease classification	69 044	17 480	25 596	Avg. AUC-ROC
SIIM-ACR [56]	Pneumothorax segmentation	8540	2 135	1 372	F1-score

^a Consists of samples from three datasets: DRIVE [57], STARE [58], CHASE_DB1 [59].

^b Measures the accuracy when all predictions of a test sample are correct.

once the heatmap generator has been trained, the heatmaps can be pre-generated at the start of stage 2 training, without introducing any overhead in heatmap computation during the representation learning process.

4. Experiments

4.1. Datasets

Pretrain datasets. Two public datasets are selected as the pretrain dataset for each medical imaging modality (fundus photographs and

chest X-rays). The first dataset is OIA-ODID [60], consisting of 10,000 fundus photographs annotated with labels from eight categories. The second dataset, ChestX-ray14 [4], consists of 112,120 chest X-rays annotated with labels from fourteen categories. Each dataset is converted into a binary coarse-grained dataset by assigning all abnormal images (regardless of disease type) to a generic abnormal class and the rest of the images to the normal class. For OIA-ODIR, all 10,000 samples are used as training samples. For ChestX-ray14, the training samples consist of 86,524 samples from the official training split. All images are resized to 448 pixels along the shorter side and fixed crops of 224×224 pixels are taken during pretraining in stage two. In addition, standard augmentations including cropping, horizontal flipping, rotation, color jittering and grayscaling are randomly applied to each training sample.

Downstream datasets. Each pretrained model is benchmarked against seven diverse downstream tasks, including five tasks in the fundus photograph modality and two tasks in the chest X-ray modality. The details of the downstream datasets and data augmentation schemes are given in Tables 1 and 2, respectively. For datasets without official validation split, 20% of the training split is selected for validation. During training, images in the fundus classification and segmentation tasks are resized to 350 and 514 pixels along the shorter side, respectively, whereas the X-ray images are resized to 224×224 pixels. The primary evaluation metrics used in prior works is reported for each downstream task.

4.2. Experiment setups

Following prior work in medical image pretraining [6,9,14], U-Net [61] is used as the network architecture in all experiments. Representations of the image crops are obtained by the encoder of U-Net while the representations of the heatmap crops are extracted using a small CNN consisting of three convolutional layers, each with a set of 3×3 kernels followed by a ReLU activation function. The projection networks follow the design of the SimCLR framework [12], which consists of a 2-layer MLP projection head to project the encoder representation to a 128-dimensional latent space. During pretraining, the networks are optimized using the SGD optimizer with a momentum of 0.95, a base learning rate of 0.001 and a weight decay of $1e-4$. The learning rate is linearly warmed-up to the base value for the first 30% of the optimization steps before gradually reduced via a cosine scheduler. The batch size is set to 120, and the total optimization steps for OIA-ODIR and ChestX-ray14 are 25k and 72k, respectively.

For the transfer learning experiments, the U-Net encoder is initialized with the pretrained weights (the projectors and heatmap encoder are discarded), and a randomly initialized linear layer or decoder is attached for classification and segmentation task, respectively. Grid search is employed to search for the optimal task-specific learning rate and weight decay value for each initialization method to ensure fair comparisons. Details of the search range and the training setup for downstream optimizations are provided in the Appendix. Up to two NVIDIA V100 GPUs, each with 16 GB of memory, were used to optimize the networks. Each experiment is repeated three times with different random seeds.

Algorithm 1 CAMContrast's learning algorithm.

Input: image dataset D_I , crop dataset D_c , batch size B , transformation functions \mathcal{T} , normal-abnormal classifier w , image encoder f_I , image projector g_I , heatmap encoder f_h , heatmap projector g_h

Output: pretrained encoder f_I

```

1: # stage 1
2: Initialize  $w$  with random weights
3: # sample from image-level dataset
4: for minibatch  $\{(x_i, y_i)\}_{i=1}^B \subset D_I$  do
5:   Update  $w$  to minimize Equation 2
6: end for
7: Convert  $w$  to heatmap generator  $h$  via Equation 1 and min-max
   normalization
8: # stage 2
9: Initialize  $f_I$ ,  $g_I$ ,  $f_h$  and  $g_h$  with random weights
10: # sample from crop-level dataset
11: for minibatch  $\{x_i\}_{i=1}^B \subset D_c$  do
12:   for  $i \in \{1, \dots, B\}$  do
13:     # construct views
14:     Sample augmentation function  $t \sim \mathcal{T}$ .
15:      $\tilde{x}_i^1 = t(x_i)$ 
16:      $\tilde{x}_i^2 = h(x_i)$ 
17:     # compute projections
18:      $z_i^1 = g_I(f_I(\tilde{x}_i^1))$ 
19:      $z_i^2 = g_h(f_h(\tilde{x}_i^2))$ 
20:     # normalize projections
21:      $z_i^1 = z_i^1 / \|z_i^1\|$ 
22:      $z_i^2 = z_i^2 / \|z_i^2\|$ 
23:   end for
24:   Let  $\ell = 0$ 
25:   for  $i \in \{1, \dots, B\}$  do
26:     # accumulate loss with Equation 5
27:      $\ell = \ell + \mathcal{L}_{con}(z_i^1, z_i^2) + \mathcal{L}_{con}(z_i^2, z_i^1)$ 
28:   end for
29:    $\ell = \frac{1}{2B} \ell$ 
30:   Update  $f_I$ ,  $g_I$ ,  $f_h$  and  $g_h$  to minimize  $\ell$ 
31: end for
32: return  $f_I$ 

```

Table 2

Parameters of image augmentation used in the downstream tasks. Images are transformed using the functions implemented in the torchvision library (<https://pytorch.org/vision/stable/index.html>).

Parameter	Fundus cls.	Fundus seg.	Chest X-ray cls.	Chest X-ray seg.
Crop size	320 × 320	514 × 514	224 × 224	224 × 224
Random crop – scale	[0.8, 1.0]	1.0	[0.8, 1.0]	1.0
Random crop – aspect ratio	[0.75, 1.33]	1.0	[0.75, 1.33]	1.0
Random horizontal flip – probability	0.5	0.5	0.5	0.5
Random rotation – degree	0.0	[-90, 90]	[-30, 30]	[-30, 30]

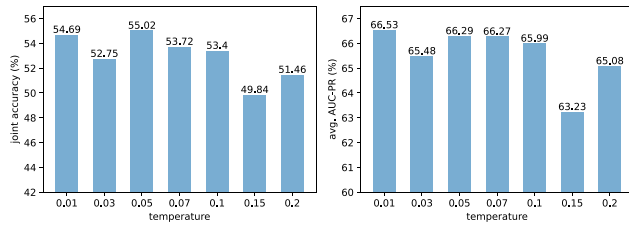


Fig. 5. Effect of temperature parameter on downstream performance in IDRiD. (Left: joint accuracy of the classification task. Right: average AUC-PR score of the segmentation task.)

Table 3

Effect of different cropping strategy on downstream performance in IDRiD. “No cropping” uses the whole image resized to 224 × 224 pixels.

Cropping strategy	Joint acc. (%)	AUC-PR (%)
No cropping	52.10	66.66
Random crop	48.54	66.27
Four-crop	53.40	64.42
Five-crop	55.02	66.29

Table 4

Effect of different CAM-related components on downstream performance in IDRiD.

Ablation	Joint acc. (%)	AUC-PR (%)
No attention loss	49.84	66.94
No negative suppression	49.51	64.97
No background suppression	52.43	65.84
Discard all negatives	47.57	65.90
Unweighted heatmaps	49.51	65.54
CAM consistency	46.28	64.02
CAMContrast	55.02	66.29

4.3. Ablation studies

To quantify the importance of each component introduced in the CAMContrast framework, ablation studies were conducted on the IDRiD benchmark [62], including a fine-grained classification task for severity level of diabetic retinopathy and diabetic macular edema, and a segmentation task for four types of retinal lesions. The performance for the classification and segmentation task are evaluated via joint accuracy and average AUC-PR score, respectively.

Temperature parameter. Fig. 5 shows the effect of the temperature parameter on downstream tasks. Temperature parameter is used to weight the cosine similarity scores of the positive and negative pairs in contrastive pretraining (Eq. (5)). The performance is generally stable across both tasks when the temperature is in the range from 0.05 to 0.1. Unless otherwise specified, 0.05 is the default temperature used in all experiments.

Cropping strategy. The ablation results for different cropping strategy used during the pretraining stage is tabulated in Table 3. The commonly used random cropping strategy achieves the lowest classification performance and comparable segmentation performance. Four-crop, which is a variant of five-crop without the center crops, attains a competitive performance in terms of classification accuracy but performs worse in the segmentation task. This suggests that the representations learnt

Table 5

Effect of different crop size on downstream performance in IDRiD.

Crop size (pixels)	Joint acc. (%)	AUC-PR (%)
112 × 112	51.95	65.67
224 × 224 (default)	55.02	66.29
336 × 336	55.83	67.03

from the center crops play an important role in dense prediction tasks. The proposed five-crop strategy achieves the best trade-off between classification and segmentation performance.

CAM-related components. In Table 4, the impact of different components related to the CAMs are investigated. Compared to the proposed CAMContrast framework, the alternatives produce representations that significantly degrade the downstream classification performance. Without the attention-based regularization in stage one (row 1), the generated heatmaps tend to focus on broader image regions and lose the ability to locate diseased regions effectively (Fig. 6), leading to degradation in classification performance. Both classification and segmentation performance drop when the heatmaps of the negative (normal) images or backgrounds are not suppressed during pretraining in stage two (row 2 and row 3). Furthermore, if the normal image-heatmap pairs are excluded from pretraining (row 4), it also results in poor downstream performance, which suggests that maximizing agreement between normal images and the prototypical vector representing an empty heatmap (via negative suppression) provides useful training signals for contrastive representation learning. To demonstrate the importance of class-specific weightings, the CAMs are replaced with the summation of feature maps at the last convolutional layer, resulting in unweighted heatmaps (row 5). Training with unweighted heatmaps as contrasting views causes performance drop in both downstream classification and segmentation tasks. Lastly, an alternative one-stage approach based on maximizing consistency between CAMs of different augmented views [63] is also explored (row 6); it achieves lower performance in both classification and segmentation tasks.

Spatial resolution of image crops. The impact of the spatial resolution of pretraining images is investigated in Table 5. The downstream performance is proportional to the crop size, with 336 × 336 pixels achieving the highest classification and segmentation performance at the cost of longer pretraining time and higher memory footprint.

Class imbalance in pretraining dataset. To study how class imbalance in the pretraining dataset will affect the downstream performance, pretraining datasets with different class ratio were constructed. Concretely, the class ratio is defined as the ratio of training samples from the minority class over training samples from the majority class. For example, a class ratio of 0.1 indicates that the majority class has 10 times more samples than the minority class. The default class ratio of the pretraining dataset for retinal fundus photographs is 0.8, which is relatively balanced. Table 6 tabulates the impact of different class ratios on the downstream performance. The performance is relatively robust until the class ratio reaches 0.1, where large drops in classification and segmentation performance are observed, suggesting that severe class imbalance will negatively impact the transfer learning performance of CAMContrast.

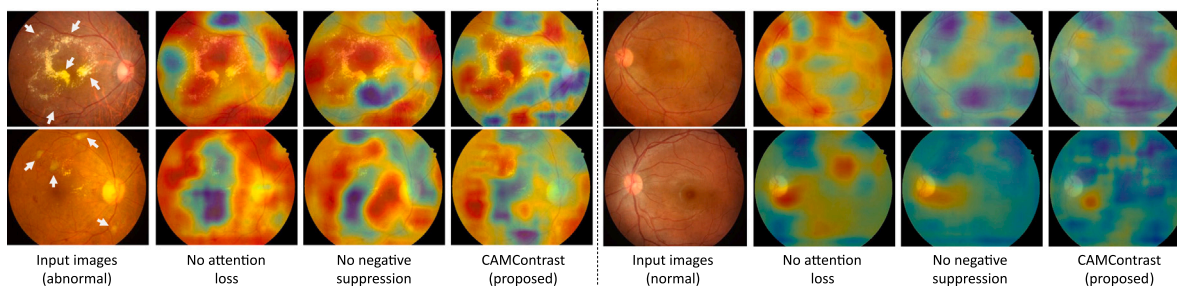


Fig. 6. Visual examples of the heatmaps. For visualization purpose, the heatmaps are upsampled to the size of 448×448 pixels and overlaid on top of the images. The images labeled as abnormal are marked with white arrows indicating the abnormal regions. Notably, inclusion of attention loss helps restricts the high activation regions to the actual abnormal regions, while negative suppression helps further suppress activation in normal images.

Table 6
Effect of different class ratio on downstream performance in IDRiD.

Class ratio	Joint acc. (%)	AUC-PR (%)
0.1	51.46	65.04
0.3	54.37	66.62
0.5	54.69	66.69
0.8 (default)	55.02	66.29

4.4. Main results

Baselines. The proposed CAMContrast framework is compared against four baselines: (i) random initialization (no pretraining), (ii) SimCLR [12], (iii) supervised pretraining with cross-entropy (SupCE), and (iv) supervised pretraining with contrastive loss (SupCon). Random initialization and SimCLR pretraining baselines give the lower bound performance for no pretraining and self-supervised pretraining, respectively. The rest of the baselines uses the coarse-grained abnormality labels as supervisory signals. In addition to the conventional random crop strategy, fixed crop strategy utilizing the pre-generated image crop dataset (used in CAMContrast) is also implemented for comparison (i.e., different fixed crops from the same image are treated as independent instances). Baselines pretrained with this strategy are denoted with the “-crop” suffix.

In the result tables, state-of-the-art results are also reported for reference. We note that these results reflect the recent performance upper-bounds achievable under different experiment settings (e.g., different input resolutions or augmentations) and were not meant to be a direct comparison with the main reported results. Recent work has pushed the state-of-the-art boundary through specialized architecture [64–66,68], multimodal pretraining [69], or task-specific adaptation [66,67]. By contrast, this work focuses on transferable weakly-supervised pretraining from the perspective of injecting positional information from binary image-level labels.

Results on fundus photograph datasets. Models pretrained on the OIA-ODIR dataset are fine-tuned on five downstream tasks with limited annotations (less than 1000 annotations). The results are shown in Table 7. Our proposed framework achieves the best performance in four out of five tasks, with significant improvement in the classification tasks. In general, models trained using the fixed crop strategy are found to generalized better than their random cropping counterparts, suggesting that fixed crop might be a better alternative for pretraining with medical images. It is also worth noting that the self-supervised baselines were able to outperform the supervised baselines in three tasks (REFUGE-cls, IDRiD-seg, REFUGE-seg) despite not having access to the disease labels. In addition, supervised contrastive learning objective struggles to outperform both self-supervised and supervised cross-entropy objectives in most benchmark datasets. These results highlight the shortcomings of pretraining directly with coarse labels

using either the cross-entropy or supervised contrastive learning objective. The shortcomings are addressed by the proposed two-stage pretraining framework through replacing the second views of SimCLR with CAM-derived heatmaps to facilitate coarse-to-fine representation learning.

Results on chest X-ray datasets. For models pretrained on the ChestX-ray14 dataset, two transfer learning experiments were conducted. The first is a linear evaluation experiment which assesses the quality of learned representations by training a linear classifier on top of a frozen encoder. The linear classifier is trained to classify each X-ray image in ChestX-ray14 into fourteen thorax disease categories. In the second experiment, the models are fine-tuned on the SIIM-ACR dataset for pneumothorax segmentation. To simulate low data-regime setting, multiple models are trained on different subsets of the original training data with varying amount of labeled samples. The proposed method outperforms other methods across all labeled proportions as shown in Tables 8 and 9. Furthermore, CAMContrast is able to achieve competitive performance in the linear classification task using only 25% of the training data. For the segmentation task, the performance gain is substantial especially at the 10% setting where labels are scarce. Qualitative segmentation results are shown in Fig. 7. The results obtained from model pretrained on CAMContrast exhibit a higher degree of overlap with the ground truth when compared to the baselines.

4.5. UMAP visualizations

In Fig. 8, the learnt representations for both pretrain datasets are visualized using UMAP [70]. Each data point represents an image crop and is labeled as either normal or abnormal according to the image-level label. In most cases, the data points are clustered by their relative position. Within each cluster, there is a noticeable separation between normal and abnormal crops in the representations extracted from the supervised models. However, this does not necessarily imply a better downstream performance, as SimCLR-crop was able to outperform SupCE-crop on several tasks in the fundus photographs benchmark (Table 7) despite the representations from SimCLR-crop not being visually separable in terms of the abnormality labels. It is interesting to note that the pretrain method with a more compact representations correlates to better performance in the downstream tasks. Clusters with high intra-cluster compactness often characterizes a good clustering algorithm [71]. In the case of representation learning, a pretrain method that can cluster its pretraining data into compact and well-separated clusters might be a key for improving transfer learning performance. With the guidance of abnormality-informed heatmaps, CAMContrast learns representations that are both compact and locally separable, which translates to a higher downstream performance.

Table 7
Comparisons of transfer learning performance on different fundus photograph datasets. Evaluation metrics are shown below the task names. Higher values indicate better performance.

Method	Classification tasks		Segmentation tasks		
	IDRiD-cls (joint accuracy)	REFUGE-cls (F1-score)	IDRiD-seg (avg. AUC-PR)	REFUGE-seg (avg. F1-score)	Vessel-seg (F1-score)
<i>SOTA</i> ^a	68.0	80.05	71.11	93.10	81.41
Random (no pretrain)	34.63 ± 0.46	63.36 ± 3.30	63.03 ± 0.05	91.32 ± 0.17	78.49 ± 0.18
SimCLR	43.37 ± 0.46	67.72 ± 1.74	64.51 ± 0.16	90.89 ± 0.16	79.71 ± 0.12
SimCLR-crop	40.78 ± 0.79	77.78 ± 1.43	64.14 ± 0.47	91.67 ± 0.14	79.76 ± 0.08
SupCE	43.37 ± 1.21	74.91 ± 4.37	62.72 ± 0.27	91.53 ± 0.04	80.04 ± 0.13
SupCE-crop	49.52 ± 4.82	75.70 ± 2.62	62.40 ± 0.89	91.74 ± 0.10	79.55 ± 0.07
SupCon	41.43 ± 0.46	66.32 ± 0.49	64.23 ± 0.45	91.46 ± 0.10	78.79 ± 0.14
SupCon-crop	42.40 ± 1.65	68.94 ± 1.66	63.48 ± 0.53	91.55 ± 0.19	78.52 ± 0.07
CAMContrast (this work)	55.02 ± 2.42	81.09 ± 0.89	66.29 ± 1.18	91.66 ± 0.06	80.82 ± 0.05

^a Recent state-of-the-art results for IDRiD-cls [64], REFUGE-cls [65], IDRiD-seg [66], REFUGE-seg [67] and Vessel-seg [68] are provided for reference.

Table 8
Comparisons of linear evaluation performance on the test set of ChestX-ray14, measured in terms of mean AUC-ROC scores (%) across 14 disease categories. The percentages represent different labeled proportions of the complete training set.

Method	1%	5%	25%	50%	100%
<i>SOTA</i> ^a	–	–	–	–	84.8
SimCLR-crop	58.43 ± 0.57	61.80 ± 0.35	64.18 ± 0.18	64.67 ± 0.13	65.04 ± 0.06
SupCE-crop	61.55 ± 0.52	66.06 ± 0.36	69.75 ± 0.10	70.30 ± 0.01	70.57 ± 0.05
SupCon-crop	60.69 ± 0.59	63.82 ± 0.32	66.55 ± 0.19	66.98 ± 0.14	67.21 ± 0.13
CAMContrast (this work)	63.74 ± 0.57	68.72 ± 0.06	72.06 ± 0.13	72.62 ± 0.08	72.85 ± 0.01

^a Recent state-of-the-art result [69] is provided for reference.

Table 9
Comparisons of fine-tuning performance on the test set of SIIM-ACR, measured in terms of F1 scores (%). The percentages represent different labeled proportions of the complete training set.

Method	10%	25%	50%	75%	100%
<i>SOTA</i> ^a	–	–	–	–	83.1
Random (no pretrain)	70.29 ± 3.51	73.21 ± 0.52	75.58 ± 0.49	77.34 ± 0.36	78.16 ± 0.09
SimCLR-crop	70.31 ± 1.23	72.39 ± 0.52	74.79 ± 0.12	76.74 ± 0.20	77.50 ± 0.51
SupCE-crop	74.33 ± 3.31	75.75 ± 0.26	77.18 ± 0.91	77.99 ± 0.28	78.32 ± 0.29
SupCon-crop	71.58 ± 1.79	73.02 ± 0.78	75.20 ± 0.31	77.49 ± 0.23	77.93 ± 0.14
CAMContrast (this work)	75.88 ± 0.67	76.50 ± 0.53	77.29 ± 0.64	78.51 ± 0.42	79.39 ± 0.12

^a Recent state-of-the-art result [69] is provided for reference.

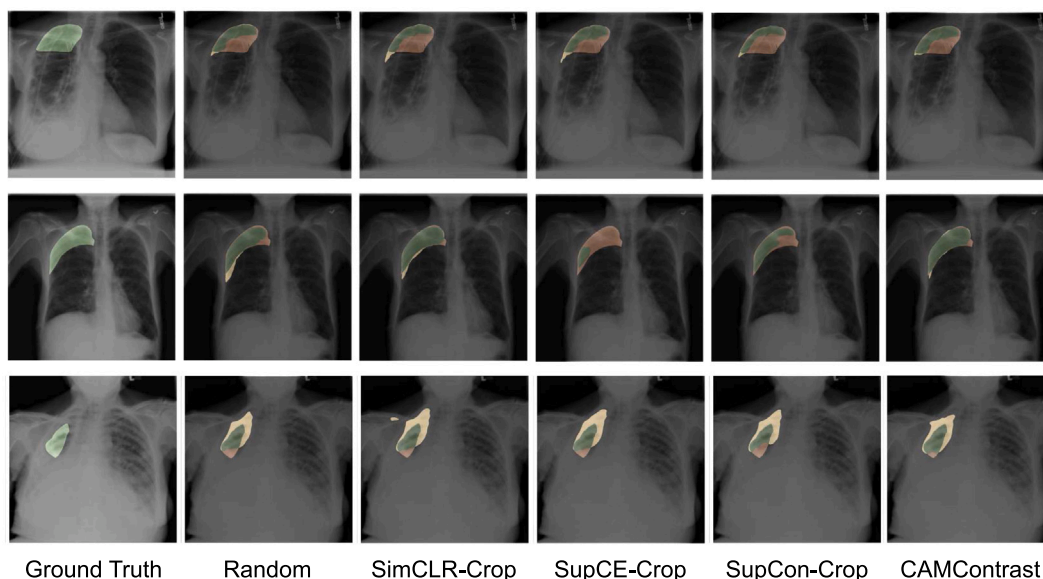


Fig. 7. Segmentation results on the test images of SIIM-ACR dataset. Segmentation masks are overlaid on top of the input images. Orange and yellow region indicates under- and over-segmentation, respectively.

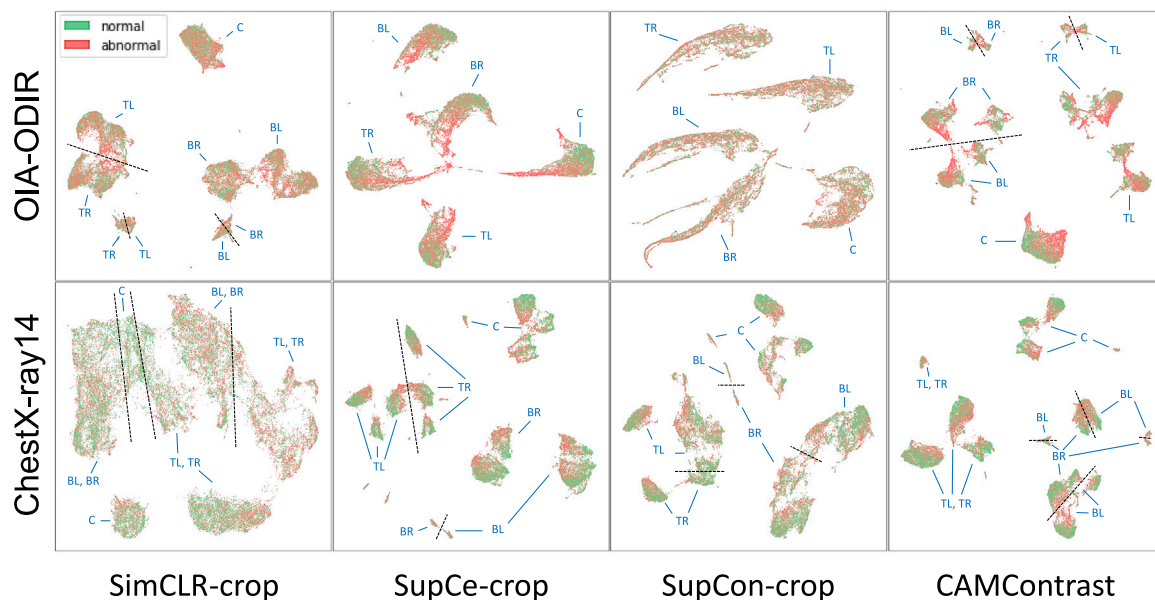


Fig. 8. UMAP visualizations of the encoder representations for the different combinations of pretrain method and dataset. Each cluster is also annotated with a relative position label indicating the cropping location of the data points in the cluster (TL: top-left, TR: top-right, C: center, BL: bottom-left, BR: bottom-right).

Table 10

Linear evaluation results (AUC-ROC) on the multilabel subset of ChestX-ray14, averaged across five validation folds.

Method	Atelectasis	Cardiomegaly	Effusion	Pneumothorax	Consolidation	Average
Random	66.44 ± 1.00	65.08 ± 2.80	70.51 ± 1.34	64.89 ± 1.81	67.87 ± 1.28	66.96 ± 0.97
SupCE-crop	72.01 ± 1.19	76.47 ± 0.54	82.18 ± 0.54	72.68 ± 0.48	72.70 ± 1.27	75.21 ± 0.31
CAMContrast	73.38 ± 1.08	81.17 ± 1.15	84.22 ± 0.53	74.42 ± 1.21	73.54 ± 1.41	77.35 ± 0.42

4.6. Generalization to multilabel setting

The versatility of the proposed CAMContrast framework in a multilabel scenario is explored in this section. Specifically, we consider a subset of the ChestX-ray14 dataset and select 5 disease classes with similar manifestations, namely Atelectasis, Cardiomegaly, Effusion, Pneumothorax, and Consolidation. These diseases are often manifested as abnormal spatial changes to the lung cavity. In total, the multilabel subset consists of 30 000 images. During pretraining, stage 1 of CAMContrast is adapted to the multilabel setting by computing the CAMs for each positive class and taking their average as the heatmap. For example, if an image is labeled positive for both Atelectasis and Effusion, the CAMs corresponding to Atelectasis and Effusion are averaged into a single heatmap. For stage 2, the same image-heatmap contrastive learning objective is applied. The quality of the learnt representation is then evaluated via 5-fold cross-validation study using the linear evaluation protocol. Table 10 shows the linear evaluation results, averaged across 5 validation folds at patient level. From the quantitative results, CAMContrast outperforms the conventional supervised pretraining method (SupCE-crop) across all disease classes, suggesting its generalization beyond the binary coarse-grained label setting.

4.7. Generalization to 3D dataset

To validate the efficacy of CAMContrast in 3D medical images, additional experiments are conducted on brain MRI scans from the Alzheimer's Disease Neuroimage Initiative (ADNI)¹ [72]. Specifically, we consider the ADNI-1 image collections, which consists of 784

T1-weighted MRI scans, each labeled with one of the three classes: Alzheimer's disease (AD), mild cognitive impairment (MCI) and cognitively normal (CN). To simulate weakly-supervised pretraining with binary labels, AD and MCI are treated as the abnormal label while CN is treated as the normal label. The Clinica software suite [73] is used to preprocess the raw data into 3D tensors of size $80 \times 96 \times 80$ for training and evaluation. It is also used to split the samples into a development set and a holdout testing set. The development set consists of 628 samples while the holdout testing set consists of 156 samples. Among the 628 samples, 5-fold cross-validation at the patient level is conducted and the best model in each fold is evaluated on the 156 holdout samples. The image encoder used in this experiment is a variant of ResNet encoder introduced by Yang et al. [74] which adapts the architecture to handle 3D volumetric inputs and to produce 3D CAMs. For crop-based training, each volume is divided into 8 non-overlapping sub-volumes and 2 center sub-volumes, each with a size of $40 \times 48 \times 40$. Other than architectural changes, the pretraining and fine-tuning pipeline for the baselines and CAMContrast remains unchanged. During pretraining, the network is trained for 120 epochs using the SGD optimizer with a learning rate of $1e-3$, a weight decay of $1e-5$, and a batch size of 16. For the transfer learning experiments, the encoder is fine-tuned on the 3-way classification task (AD, MCI, CN) for 100 epochs with a weight decay of $1e-5$ and a batch size of 8. The search range for the learning rate include 5 logarithmically spaced values in the interval $[1e-3, 1e-1]$.

Table 11 summarizes the transfer learning results. Overall, CAMContrast achieved the highest mean average AUC-ROC compared to

¹ Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The

primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Table 11

Transfer learning results (AUC-ROC) on ADNI-1 dataset. The results are averaged across five validation folds. The model that achieves the best validation performance in each fold is used to generate predictions for the holdout testing set.

Method	Validation				Test			
	CN	MCI	AD	Mean	CN	MCI	AD	Mean
Random	70.28 ± 5.28	55.28 ± 4.07	72.21 ± 4.17	65.92 ± 3.99	67.17 ± 1.46	52.65 ± 1.7	70.51 ± 2.32	63.44 ± 1.76
SupCE-crop	73.56 ± 4.63	56.24 ± 4.31	74.92 ± 5.12	68.24 ± 4.26	71.16 ± 3.98	53.13 ± 2.67	73.84 ± 0.77	66.04 ± 1.77
SupCon-crop	70.98 ± 5.83	54.97 ± 5.21	73.19 ± 4.66	66.38 ± 4.02	68.03 ± 2.83	52.9 ± 0.63	71.91 ± 3.1	64.28 ± 1.84
CAMContrast	74.21 ± 5.4	58.33 ± 5.62	75.51 ± 4.21	69.35 ± 4.65	72.85 ± 3.53	55.11 ± 1.97	72.62 ± 1.85	66.86 ± 1.93

Table 12

Comparisons of downstream performance with different backbone architectures in ChestX-ray14 (linear evaluation) and SIIM-ACR (fine-tuning). U-Net encoder is the default architecture used in the main experiments.

Architecture	Method	ChestX-ray14 (mean AUC)	SIIM-ACR (F1-score)
U-Net encoder (default)	Random	61.62 ± 0.24	78.16 ± 0.09
	SupCE-crop	70.57 ± 0.05	78.32 ± 0.29
	CAMContrast	72.85 ± 0.01	79.39 ± 0.12
ResNet-34	Random	59.54 ± 0.2	77.85 ± 0.89
	SupCE-crop	66.42 ± 0.05	77.85 ± 0.69
	CAMContrast	70.06 ± 0.01	78.91 ± 0.31
DenseNet-121	Random	63.94 ± 0.29	76.22 ± 1.77
	SupCE-crop	68.03 ± 0.04	77.21 ± 0.66
	CAMContrast	71.31 ± 0.05	78.82 ± 0.24
MiT-b1	Random	59.28 ± 0.04	74.93 ± 3.45
	SupCE-crop	63.99 ± 0.08	72.75 ± 1.85
	CAMContrast	68.04 ± 0.05	74.32 ± 0.7

Table 13

Downstream performance of different pretraining method in ChestX-ray14 (linear evaluation) and SIIM-ACR (fine-tuning). Self-supervised baselines (SSL) are provided for reference. During pretraining, all methods used the fixed crop strategy introduced in Section 3.2.

Framework	ChestX-ray14 (mean AUC)	SIIM-ACR (F1-score)
SimCLR (SSL)	65.04	77.50
SimCLR (CAMContrast)	72.85	79.39
DiRA (SSL)	72.26	79.99
DiRA (CAMContrast)	73.59	80.76

alternative pretraining methods, suggesting that it can be used to improve downstream performance in 3D setting without the need to modify the algorithm other than the backbone architecture.

4.8. Generalization to other backbone architectures

To validate the efficacy of CAMContrast in other backbone architectures, additional experiments were conducted on other notable architectures, including ResNet-34 [1], DenseNet-121 [75], and MiT-b1 [76]. Table 12 shows the transfer learning results on the X-ray datasets. Across different architectures, the proposed CAMContrast pretraining framework consistently outperforms the supervised cross-entropy training in both classification (ChestX-ray14) and segmentation (SIIM-ACR) tasks. However, CAMContrast slightly underperforms random initialization in segmentation task using the Transformer-based MiT-b1 backbone. Furthermore, the downstream performance for this backbone architecture is the lowest compared to other convolution-based architectures. We posit that this is due the lack of inductive biases in the Transformer architecture [77,78]. Without built-in inductive biases provided by the convolution operation, Transformer requires a much larger pretraining dataset to extract transferable representations, and this limits its use in medical imaging tasks with limited data [78].

4.9. Generalization to other pretraining framework

CAMContrast describes a general framework of extracting fine-grained representations from coarse annotations, and Section 3 provides a simple instantiation based on the CAM visualization and the

SimCLR objective. In general, the objective of maximizing similarities within image-heatmap pairs can be integrated into any pairwise learning framework. To demonstrate the generalizability of the proposed method, CAMContrast is integrated into DiRA [6], a self-supervised learning framework recently proposed for medical imaging tasks. Specifically, a heatmap projector and the contrastive objective on image-heatmap pairs are added to the MoCo-v2 [79] variant of DiRA. Table 13 shows the transfer learning results on the X-ray datasets. With the addition of a momentum encoder and an auxiliary decoder for reconstruction and adversarial learning, self-supervised DiRA attains similar performance as the SimCLR implementation of CAMContrast. However, this comes at a cost of larger memory footprint due to the decoder parameters as well as additional effort in hyperparameter tuning to balance the weighing factor of each loss term. This suggests a trade-off between the availability of pretrain labels and the complexity of learning objectives. Nevertheless, we show that CAMContrast is complementary to the DiRA framework as it is able to improve upon the downstream performance when the heatmap views are introduced.

5. Discussion

CAMContrast aims to learn a general-purpose image feature extractor for medical images using only coarsely labeled images. The image feature extractor can then be used for different kinds of downstream medical image analysis applications with limited fine-grained labels, thereby reducing the data acquisition cost. Beyond single modality use cases, the high-dimensional image features extracted by our deep learning-based extractor can also be potentially combined with radiomic features [80,81] or clinical features [82] in a hybrid system to improve predictive performance while remains interpretable to clinical practitioners.

The proposed fixed crop training strategy is particularly effective in medical images that exhibit high inter-image similarity, such as fundus photographs, chest radiographs and brain MRI scans. For image modality such as skin images commonly used by dermatologists to diagnose skin diseases, the images may be taken from an ordinary RGB camera at different angles and focused on different parts of the body. At such, applying fixed crop training to skin images would be akin to random cropping, as there is little to no structural similarity between different skin images. This diminishes the beneficial effect of hard negative pairs formed by structurally similar image crops, which limits the applicability of CAMContrast in this scenario.

6. Conclusion

A coarse-to-fine representation learning framework for medical images is proposed in this work. Utilizing class activation maps, novel views in the form of heatmaps are extracted from a classification-trained neural network. These heatmaps, which contain localized information of the abnormal regions, enable fine-grained representation learning through maximizing the similarities of image-heatmap pairs in a shared projection space. Furthermore, a fixed crop training strategy is introduced to promote learning of fine-grained details by forcing the neural network to distinguish between crops of different images sampled from the same position. Compared to other supervised pretraining methods, CAMContrast is more data-efficient and achieves better transfer learning performance in benchmarks comprising of fundus photographs and chest radiographs.

Table A.1
Search range and hyperparameters used in the downstream tasks.

Parameter	IDRiD-cls	REFUGE-cls	Chest X-ray14	IDRiD-seg	REFUGE-seg	Vessel-seg	SIIM-ACR
Learning rate	Grid search ^a	Grid search ^a	Grid search ^a	Grid search ^b	Grid search ^c	Grid search ^c	Grid search ^c
Weight decay	Grid search ^d	Grid search ^d	0.0	1e-5	1e-5	1e-5	1e-5
LR scheduler	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine
Optimizer	AdamW	SGD	SGD	AdamW	SGD	SGD	SGD
Momentum	–	0.9	0.9	–	0.9	0.9	0.9
Max epochs	300	150	60	300	90	90	60
Batch size	40	40	40	5	5	5	5

^a LR $\in \{1e-4, 3e-4, 1e-3, 3e-3, 1e-2, 3e-2, 1e-1\}$.

^b LR $\in \{3e-5, 1e-4, 3e-4, 1e-3, 3e-3\}$.

^c LR $\in \{1e-3, 3e-3, 1e-2, 3e-2, 1e-1\}$.

^d WD/LR, WD $\in \{0, 1e-6, 1e-5, 3e-5, 1e-4, 3e-4, 1e-3, 3e-3\}$.

Table B.1
Disease-specific accuracies (%) on the test set of IDRiD-cls.

Method	Diabetic retinopathy	Macular edema	Average accuracy	Joint accuracy
Random (no pretrain)	43.04 \pm 2.42	74.76 \pm 1.37	58.90 \pm 1.27	34.63 \pm 0.46
SimCLR	51.78 \pm 0.46	76.70 \pm 0.79	64.24 \pm 0.46	43.37 \pm 0.46
SimCLR-crop	48.22 \pm 2.00	77.02 \pm 1.21	62.62 \pm 0.69	40.78 \pm 0.79
SupCE	51.13 \pm 2.79	76.05 \pm 1.21	63.59 \pm 1.73	43.37 \pm 1.21
SupCE-crop	61.49 \pm 3.75	75.73 \pm 1.37	68.61 \pm 2.55	49.52 \pm 4.82
SupCon	48.22 \pm 1.65	76.38 \pm 0.46	62.30 \pm 0.61	41.43 \pm 0.46
SupCon-crop	51.78 \pm 3.21	79.61 \pm 1.37	65.69 \pm 1.39	42.40 \pm 1.65
CAMContrast (this work)	60.84 \pm 3.21	82.20 \pm 0.46	71.52 \pm 1.79	55.02 \pm 2.42

Table B.2
Disease-specific AUC-ROC (%) on the test set of ChestX-ray14.

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass
SimCLR-crop	65.19	66.43	69.61	65.08	59.22
SupCE-crop	69.79	75.46	76.98	66.88	69.53
SupCon-crop	67.36	71.37	73.12	65.88	63.65
CAMContrast (this work)	72.19	79.08	79.51	67.38	73.69
Method	Nodule	Pneumonia	Pneumothorax	Consolidation	Edema
SimCLR-crop	59.09	60.00	69.62	66.86	77.99
SupCE-crop	65.32	62.77	73.64	69.48	79.57
SupCon-crop	60.51	61.74	72.41	68.23	78.09
CAMContrast (this work)	67.80	63.78	77.54	70.51	81.25
Method	Emphysema	Fibrosis	Pleural thickening	Hernia	Average
SimCLR-crop	62.79	70.93	63.47	54.82	65.08
SupCE-crop	70.58	74.40	69.49	64.86	70.63
SupCon-crop	63.63	72.80	66.29	55.83	67.21
CAMContrast (this work)	75.33	76.10	70.13	65.63	72.85

CRedit authorship contribution statement

Boon Peng Yap: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation. **Beng Koon Ng:** Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the support from the National Supercomputing Centre of Singapore. The computational work for this article was fully performed on resources of the National Supercomputing Centre, Singapore (<https://www.nscg.sg>).

Appendix A. Hyperparameter search range

The transfer learning setup largely follows the protocols described in SimCLR [12]. It involves grid search of task-specific parameters for

each weight initialization method. The search range and other hyperparameters are summarized in Table A.1. After selecting the optimal parameters on the validation set, the networks are retrained using all samples from the training and validation sets. Final results are reported on hold-out test sets.

Appendix B. Additional quantitative results

The disease-specific results for the downstream multiclass classification tasks are provided in Tables B.1 and B.2. Notably, CAMContrast achieves the best performance in macular edema classification in IDRiD-cls, and consistently outperforms other baselines across all 14 diseases in ChestX-ray14, with up to 4.75% improvement in emphysema classification. The results further highlight the effectiveness of CAMContrast in learning fine-grained information from coarsely labeled datasets.

Appendix C. Additional results on low data regime setting

Tables C.1 and C.2 shows additional experiment results for the DiRA framework [6] under low data regime settings. In this setup, the performance gap between self-supervised learning and CAMContrast becomes larger, indicating that coarse labels are able to provide informative signals for learning the encoder representations.

Table C.1

Comparisons of linear evaluation performance on the test set of ChestX-ray14, measured in terms of mean AUC-ROC scores (%) across 14 disease categories. The percentages represent different labeled proportions of the complete training set.

Method	1%	5%	25%	50%
DiRA (SSL)	63.36 ± 0.73	68.39 ± 0.04	71.36 ± 0.14	71.94 ± 0.10
DiRA (CAMContrast)	65.00 ± 0.51	69.96 ± 0.16	72.82 ± 0.13	73.25 ± 0.04

Table C.2

Comparisons of fine-tuning performance on the test set of SIIM-ACR, measured in terms of F1 scores (%). The percentages represent different labeled proportions of the complete training set.

Method	10%	25%	50%	75%
DiRA (SSL)	74.30 ± 1.10	77.03 ± 0.57	77.58 ± 0.59	79.34 ± 0.09
DiRA (CAMContrast)	76.65 ± 0.92	78.16 ± 0.53	77.91 ± 0.25	79.50 ± 0.17

References

- [1] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *CVPR, IEEE*, 2016, pp. 770–778.
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: *CVPR, IEEE*, 2009.
- [3] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, *JAMA* 316 (22) (2016) 2402–2410.
- [4] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: *CVPR, IEEE*, 2017, pp. 2097–2106.
- [5] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, D. Rueckert, Self-supervised learning for medical image analysis using image context restoration, *Med. Image Anal.* 58 (2019) 101539.
- [6] F. Haghghi, M.R.H. Taher, M.B. Gotway, J. Liang, Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis, in: *CVPR, IEEE*, 2022, pp. 20824–20834.
- [7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: feature learning by inpainting, in: *CVPR, IEEE*, 2016, pp. 2536–2544.
- [8] H.-Y. Zhou, C.-K. Lu, S. Yang, X. Han, Y. Yu, Preservational learning improves self-supervised medical image models by reconstructing diverse contexts, *ICCV* (2021) 3479–3489.
- [9] Z. Zhou, V. Sodha, J. Pang, M.B. Gotway, J. Liang, Models genesis, *Med. Image Anal.* 67 (2021) 101840, <http://dx.doi.org/10.1016/j.media.2020.101840>.
- [10] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, in: *ICLR*, 2018.
- [11] X. Li, X. Hu, X. Qi, L. Yu, W. Zhao, P. Heng, L. Xing, Rotation-oriented collaborative self-supervised learning for retinal disease diagnosis, *IEEE Trans. Med. Imaging* PP (2021).
- [12] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: *ICML, PMLR*, 2020, pp. 1597–1607.
- [13] K. He, H. Fan, Y. Wu, S. Xie, R.B. Girshick, Momentum contrast for unsupervised visual representation learning, in: *CVPR, IEEE*, 2020, pp. 9726–9735.
- [14] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu, Contrastive learning of global and local features for medical image segmentation with limited annotations, in: *NeurIPS*, Vol. 33, Curran Associates, Inc., 2020.
- [15] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: *ICCV*, 2021, pp. 9650–9660.
- [16] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent—a new approach to self-supervised learning, in: *NeurIPS*, Vol. 33, Curran Associates, Inc., 2020.
- [17] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: *ICML, PMLR*, 2021, pp. 12310–12320.
- [18] Y. Huang, L. Lin, P. Cheng, J. Lyu, X. Tang, Lesion-based contrastive learning for diabetic retinopathy grading from fundus images, in: *MICCAI*, Springer, 2021.
- [19] Y.-H.H. Tsai, T. Li, W. Liu, P. Liao, R. Salakhutdinov, L.-P. Morency, Learning weakly-supervised contrastive representations, in: *ICLR*, 2022.
- [20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: *CVPR, IEEE*, 2016, pp. 2921–2929.
- [21] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: *NeurIPS*, Vol. 33, Curran Associates, Inc., 2020, pp. 18661–18673.
- [22] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, *CVPR* (2016) 3213–3223.
- [23] J. Dai, K. He, J. Sun, BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, *ICCV* (2015) 1635–1643.
- [24] M. Rajchl, M.J. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, B. Kainz, D. Rueckert, DeepCut: Object segmentation from bounding box annotations using convolutional neural networks, *IEEE Trans. Med. Imaging* 36 (2017) 674–683.
- [25] Z. Huang, X. Wang, J. Wang, W. Liu, J. Wang, Weakly-supervised semantic segmentation network with deep seeded region growing, in: *CVPR, IEEE*, 2018, pp. 7014–7023.
- [26] A. Kolesnikov, C.H. Lampert, Seed, expand and constrain: Three principles for weakly-supervised image segmentation, in: *ECCV*, Springer, 2016, pp. 695–711.
- [27] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: *CVPR, IEEE*, 2020, pp. 12275–12284.
- [28] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: *AISTATS, PMLR*, 2015, pp. 562–570.
- [29] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: *Advances in Neural Information Processing Systems*, 2017, pp. 1195–1204.
- [30] S. Reiß, C. Seibold, A. Freytag, E. Rodner, R. Stiefelhagen, Every annotation counts: Multi-label deep supervision for medical image segmentation, *CVPR* (2021) 9527–9537.
- [31] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Barambe, L. van der Maaten, Exploring the limits of weakly supervised pretraining, in: *ECCV*, 2018, pp. 181–196.
- [32] S. Zhang, R. Xu, C. Xiong, C. Ramaiah, Use all the labels: A hierarchical multi-label contrastive learning framework, *CVPR* (2022) 16639–16648.
- [33] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: *CVPR*, Vol. 1, IEEE, 2005, pp. 539–546.
- [34] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv:1807.03748.
- [35] B. Dufumier, P. Gori, J. Victor, A. Grigis, M. Wessa, P. Brambilla, P. Favre, M. Polosan, C. McDonald, C. Piguat, E. Duchesnay, Contrastive learning with continuous proxy meta-data for 3D MRI classification, in: *MICCAI*, Springer, 2021, pp. 58–68.
- [36] J. Peng, P. Wang, C. Desrosiers, M. Pedersoli, Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels, in: *NeurIPS*, Curran Associates, Inc., 2021.
- [37] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, 2020, arXiv: 2010.00747.
- [38] X. Li, M. Jia, M.T. Islam, L. Yu, L. Xing, Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis, *IEEE Trans. Med. Imaging* 39 (2020) 4023–4033.
- [39] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: *ICCV*, 2017, pp. 2223–2232.
- [40] S.J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, B. Schiele, Exploiting saliency for object segmentation from image level labels, in: *CVPR, IEEE*, 2017, pp. 5038–5047.
- [41] X. Zhang, Y. Wei, J. Feng, Y. Yang, T.S. Huang, Adversarial complementary learning for weakly supervised object localization, in: *CVPR, IEEE*, 2018, pp. 1325–1334.
- [42] X. Zhang, Y. Wei, G. Kang, Y. Yang, T. Huang, Self-produced guidance for weakly-supervised object localization, in: *ECCV*, Springer, 2018, pp. 597–613.
- [43] Y. Zhong, J. Wang, L. Wang, J. Peng, Y.-X. Wang, L. Zhang, DAP: Detection-aware pre-training with weak supervision, *CVPR* (2021) 4535–4544.
- [44] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: *ICCV*, 2017, pp. 618–626.
- [45] Y. Tian, D. Krishnan, P. Isola, Contrastive multiview coding, in: *ECCV*, 2020.
- [46] H. Touvron, A. Sablayrolles, M. Douze, M. Cord, H. Jégou, Graft: Learning fine-grained image representations with coarse labels, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 874–884.
- [47] Y. Zhu, X. Gao, B. Ke, R. Qiao, X. Sun, Coarse-to-fine: Learning compact discriminative representation for single-stage image retrieval, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11260–11269.
- [48] J. Zhang, K. Mei, Y. Zheng, J. Fan, Learning multi-layer coarse-to-fine representations for large-scale image classification, *Pattern Recognit.* 91 (2019) 175–189.
- [49] H. Singh, P. Zhang, Q. Wang, M.M. Wang, W. Xiong, J. Du, Y. Chen, Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality, in: *Conference on Empirical Methods in Natural Language Processing*, 2023.
- [50] Y.-H.H. Tsai, T. Li, M.Q. Ma, H. Zhao, K. Zhang, L.-P. Morency, R. Salakhutdinov, Conditional contrastive learning with kernel, in: *International Conference on Learning Representations*, 2022.
- [51] S. Venkataramanan, K.-C. Peng, R.V. Singh, A. Mahalanobis, Attention guided anomaly localization in images, in: *ECCV*, Springer, 2020.

- [52] H. Xuan, A. Stylianou, X. Liu, R. Pless, Hard negative examples are hard, but useful, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, Springer, 2020, pp. 126–142.
- [53] Y. Kalantidis, M.B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus, Hard negative mixing for contrastive learning, *Adv. Neural Inf. Process. Syst.* 33 (2020).
- [54] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabudhe, F. Meriaudeau, Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research, *Data* 3 (3) (2018) 25.
- [55] J.I. Orlando, H. Fu, J.B. Breda, K. Van Keer, D.R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, et al., Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs, *Med. Image Anal.* 59 (2020) 101570.
- [56] SIIM, SIIM-ACR pneumothorax segmentation, 2019, <https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation>.
- [57] J. Staal, M.D. Abràmoff, M. Niemeijer, M.A. Viergever, B. Van Ginneken, Ridge-based vessel segmentation in color images of the retina, *IEEE Trans. Med. Imaging* 23 (4) (2004) 501–509.
- [58] A. Hoover, V. Kouznetsova, M.H. Goldbaum, Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response, *IEEE Trans. Med. Imaging* 19 (2000) 203–210.
- [59] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanovara, A.R. Rudnicka, C.G. Owen, S.A. Barman, An ensemble classification-based approach applied to retinal blood vessel segmentation, *IEEE Trans. Biomed. Eng.* 59 (9) (2012) 2538–2548.
- [60] N. Li, T. Li, C. Hu, K. Wang, H. Kang, A benchmark of ocular disease intelligent recognition: one shot for multi-disease detection, in: *Bench*, 2020.
- [61] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *MICCAI*, Springer, 2015, pp. 234–241.
- [62] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, et al., IdriD: Diabetic retinopathy–segmentation and grading challenge, *Med. Image Anal.* 59 (2020) 101561.
- [63] G. Patel, J. Dolz, Weakly supervised segmentation with cross-modality equivariant constraints, *Med. Image Anal.* 77 (2021) 102374.
- [64] W. Tang, Z. Yang, Y. Song, Disease-grading networks with ordinal regularization for medical imaging, *Neurocomputing* 545 (2023) 126245.
- [65] A. Haider, M. Arsalan, C. Park, H. Sultan, K.R. Park, Exploring deep feature-blending capabilities to assist glaucoma screening, *Appl. Soft Comput.* 133 (2022) 109918.
- [66] H. Wang, Y. Zhou, J. Zhang, J. Lei, D. Sun, F. Xu, X. Xu, Anomaly segmentation in retinal images with poisson-blending data augmentation, *Med. Image Anal.* 81 (2022) 102534.
- [67] Z. Li, C. Zhao, Z. Han, C. Hong, TUNet and domain adaptation based learning for joint optic disc and cup segmentation, *Comput. Biol. Med.* 163 (2023) 107209.
- [68] A. Galdran, A. Anjos, J. Dolz, H. Chakor, H. Lombaert, I.B. Ayed, State-of-the-art retinal vessel segmentation with minimalistic models, *Sci. Rep.* 12 (1) (2022) 6174.
- [69] R. Wang, Q. Yao, H. Lai, Z. He, X. Tao, Z. Jiang, S.K. Zhou, ECAMP: Entity-centered context-aware medical vision language pre-training, 2023, arXiv preprint arXiv:2312.13316.
- [70] L. McInnes, J. Healy, N. Saul, L. Grossberger, UMAP: Uniform manifold approximation and projection, *J. Open Source Softw.* 3 (29) (2018) 861.
- [71] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Clustering validity checking methods: part II, *SIGMOD Rec.* 31 (2002) 19–27.
- [72] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, L. Beckett, The Alzheimer’s disease neuroimaging initiative, *Neuroimaging Clin.* 15 (4) (2005) 869–877.
- [73] A. Routier, N. Burgos, M. Díaz, M. Bacci, S. Bottani, O. El-Rifai, S. Fontanella, P. Gori, J. Guillon, A. Guyot, et al., Clinica: An open-source software platform for reproducible clinical neuroscience studies, *Front. Neuroinform.* 15 (2021) 689675.
- [74] C. Yang, A. Rangarajan, S. Ranka, Visual explanations from deep 3D convolutional neural networks for Alzheimer’s disease classification, in: *AMIA Annual Symposium Proceedings, Vol. 2018*, American Medical Informatics Association, 2018, p. 1571.
- [75] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [76] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Process. Syst.* 34 (2021) 12077–12090.
- [77] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: *International Conference on Learning Representations*, 2021.
- [78] C. Matsoukas, J.F. Haslum, M.P. Soderberg, K. Smith, Is it time to replace CNNs with transformers for medical images?, 2021, ArXiv abs/2108.09038.
- [79] X. Chen, H. Fan, R.B. Girshick, K. He, Improved baselines with momentum contrastive learning, 2020, ArXiv abs/2003.04297.
- [80] S.M. Rezaeijo, N. Chegeni, F. Baghaei Naeini, D. Makris, S. Bakas, Within-modality synthesis and novel radiomic evaluation of brain MRI scans, *Cancers* 15 (14) (2023) 3565.
- [81] M.R. Salmanpour, M. Hosseinzadeh, S.M. Rezaeijo, A. Rahmim, Fusion-based tensor radiomics using reproducible features: Application to survival prediction in head and neck cancer, *Comput. Methods Programs Biomed.* 240 (2023) 107714.
- [82] M. Hosseinzadeh, A. Gorji, A. Fathi Jouzdani, S.M. Rezaeijo, A. Rahmim, M.R. Salmanpour, Prediction of cognitive decline in Parkinson’s disease using clinical and DAT SPECT imaging features, and hybrid machine learning systems, *Diagnostics* 13 (10) (2023) 1691.