CDNet: Contrastive Disentangled Network for Fine-Grained Image Categorization of Ocular B-Scan Ultrasound

Ruiling Dan, Student Member, IEEE, Yunxiang Li, Yijie Wang, Xiaodiao Chen, Gangyong Jia, Shuai Wang, Ruiquan Ge, Member, IEEE, Guiping Qian, Jun Jin, Senior Member, IEEE, Juan Ye, and Yaqi Wang

Abstract—Precise and rapid categorization of images in the B-scan ultrasound modality is vital for diagnosing ocular diseases. Nevertheless, distinguishing various diseases in ultrasound still challenges experienced ophthalmologists. Thus a novel contrastive disentangled network (CDNet) is developed in this work, aiming to tackle the fine-grained image categorization (FGIC) challenges of ocular abnormalities in ultrasound images, including intraocular tumor (IOT), retinal detachment (RD), posterior scleral staphyloma (PSS), and vitreous hemorrhage (VH). Three essential components of CDNet are the weakly-supervised lesion localization module (WSLL), contrastive multi-zoom (CMZ) strategy, and hyperspherical contrastive disentangled loss (HCD-Loss), respectively. These components facilitate feature disentanglement for fine-grained recognition in both the input and output aspects. The proposed CDNet is validated on our ZJU Ocular Ultrasound Dataset (ZJU-OUSD), consisting of 5213 samples. Furthermore, the generalization ability of CDNet is validated on two public and widely-used chest X-ray FGIC benchmarks. Quantitative and qualitative results demonstrate the efficacy of our proposed CDNet, which achieves state-of-the-art performance in the FGIC task.

Index Terms—Fine-grained categorization, ocular B-scan ultrasound, disentangled, contrastive learning.

I. INTRODUCTION

OCULAR abnormalities, such as intraocular tumor (IOT), retinal detachment (RD), posterior scleral staphyloma (PSS), and vitreous hemorrhage (VH), can develop a severe visual disability and even blindness without treatment [1], [2]. Global assessments in 2020 show that 1.1 billion individuals have ocular diseases, which is expected to rise to 1.8 billion by 2050 [3]. Untimely treatment will increase eye disease and treatment costs, leading to more severe economic losses to patients and families. A recent global financial assessment in 2020 indicates a US$411 billion loss concerning ocular diseases, negatively impacting global medical systems and economic growth [4]. Hence, early diagnosis and timely treatment are crucial for enhancing visual prognosis and decreasing the financial burden [5].

A proper diagnostic instrument is a key to early and quick detection and following cure of ocular anomalies [6]. Commonly used imaging methods include optical coherence tomography (OCT) and digital retinal imaging. However, OCT and digital retinal imaging may fail to obtain information about ocular posterior segment diseases when facing the condition of light-conducting media opacification, such as cataracts or vitreous hemorrhage [7]. In contrast, ocular ultrasound (OUS) can effectively provide posterior segment information. Moreover, the OUS is cost-effective and widely available, which is vital for most grassroots hospitals and patients in third-world countries. As one of the non-invasive measures, OUS is preferred over biopsy regarding patient comfort and has been widely applied to obtain clinical information for patients with PSS, VH, RD, and vision loss [5], [8]. Besides, OUS can be utilized in ocular diagnosis to obtain features for detecting an intraocular tumor via the location and lesion extent [6].

Nonetheless, ultrasound requires the expertise of ophthalmologists for the accurate and rapid diagnosis of ocular abnormalities. Moreover, ophthalmologists may dispute the diagnosis of ocular abnormalities in clinical practice since explicit or implicit human factors, such as lack of experience and tiredness, may
causing the diagnosis to be subjective and prone to error [9]. Therefore, non-biased and automated recognition algorithms are desirable to provide a precise and efficient diagnosis of ocular ultrasound images.

However, the diagnosis of ocular abnormalities is a challenging task for existing works due to many intrinsic complexities: 1) As shown in Fig. 1, the classification performance of models is likely to be severely affected by the irrelevant background, similar spatial features, subtle differences, and contrast diversity between OUS samples, which means it is a challenging fine-grained categorization task. 2) There are often multiple abnormal ocular regions in an image, and they are often concentrated in one area, as depicted in Fig. 4. Thus, the model may learn all lesion features when extracting classification features. It may induce the model to focus on the most salient features, resulting in some small pathological changes that tend to be ignored. 3) Extracting only the entangled lesion features will deteriorate models’ generalization performance, neglecting tiny and partial lesions and even wrong categorization.

Recently, plenty of network backbones have been proposed for coarse-grained categorization tasks, including VGG [10], Inception v3 [11], ResNet [12], DenseNet [13], and EfficientNet [14]. However, they are struggling with fine-grained tasks due to the above challenges. Therefore, it is necessary to design a suitable network for ocular tasks. The research on ocular abnormalities is relatively limited, although many studies are devoted to fine-grained categorization tasks [15], [16]. The existing attention-based and two-stage approaches [15], [17] have achieved promising results in fine-grained categorization tasks. These methods are, however, not suitable for the fine-grained categorization of ocular abnormalities, as the lesion feature disentanglement is far under-explored in these frameworks. The feature disentangled-based approaches are more appropriate than those without considering disentanglement for addressing this classification problem, as the former can distinguish the images with only part lesion features. However, classic deep models trained by cross-entropy loss (CE-Loss) to extract coupled features are sub-optimal for the fine-grained OUS recognition task, where tiny lesions are readily neglected. The CE-Loss stimulates models to concentrate on texture correlations instead of semantic and robustly informative features, impeding the feature disentanglement [18].

To handle these challenges, we propose a contrastive disentangled network that locates lesions in a weakly-supervised manner and extracts disentangled features for ocular ultrasound fine-grained image categorization. Specifically, the first step is to locate the lesion regions to reduce irrelevant backgrounds and capture tiny differences. Previous works [19], [20], [21] highly rely on segmentation labels or bounding boxes for lesion localization and further feature disentanglement. Unfortunately, such substantial annotations of lesions are far more expensive and unavailable in our dataset, where merely category labels are accessible. Thus, during the training, CDNet first locates the lesion regions via the WSLL module without relying on strong labels but only on the category labels. We devise a novel CMZ training strategy based on these weak lesion locations to generate multiple semantic consistency and challenging views containing tiny lesions to capture more subtle lesions and facilitate input feature disentanglement. Moreover, results in Section IV demonstrate that decoupling only inputs causes sub-optimal performance for fine-grained recognition of the OUS task due to the absence of output disentanglement. Therefore, we propose a novel loss, namely HCD-Loss, based on the similarity-measuring capacity of the widely-used matrix operation for the output feature disentanglement. To our knowledge, our work is the first to investigate the fine-grained image categorization (FGIC) challenge via input and output aspects.

More details of our CDNet are depicted in Fig. 2. The main contributions of our work can be summarized as follows:

1) A novel dual-branch CDNet is designed to solve the FGIC task of ocular ultrasound. The CDNet tackles the FGIC task in a novel feature disentanglement paradigm that consists of input and output disentanglement.
2) We propose a contrastive multi-zoom (CMZ) strategy to promote input feature disentanglement. The CMZ strategy can generate challenging views of lesion regions for the training.
3) We devise a hyperspherical contrastive disentangled loss (HCD-Loss) to boost output feature disentanglement. The HCD-Loss can facilitate the CDNet to extract discriminative and decoupled features for the FGIC task.
4) We build the ZJUOUSD, a large-scale ocular abnormal images benchmark in ultrasound, which can advance fine-grained OUS abnormalities categorization in bedside practice.

II. RELATED WORKS

We propose a contrastive disentangled network to categorize OUS images in this work. Our method mainly involves the four fields: ocular ultrasound recognition, attention mechanism, contrastive representation learning, and feature disentanglement. Thus, the most related works in these fields are reviewed in this section.

A. Ocular Ultrasound Abnormalities Recognition

1) Traditional Methods for OUS Task: An accurate and timely diagnosis of ocular abnormalities is vital for ocular therapy in the B-scan ultrasound modality. Traditional methods in literature [5], [22] categorize ocular abnormalities via machine learning (ML) and morphology digital image processing. However, the ocular ultrasound diagnosis systems based on traditional methods rely heavily on prior knowledge and doctor-handcrafted features, which hinders the improvement of the ocular diagnosis.

2) Deep Learning-Based Methods for OUS Task: Recent deep learning-based methods have improved ocular disease recognition tasks, such as cataract detection and retinal abnormality classification. Wang et al. [23] propose a collaborative monitoring deep learning method to detect cataracts, eliminating
the background’s influence and enhancing the detection performance. Zheng et al. [24] develop an ensemble attention method (EAM) for cataract recognition, consisting of three classification networks and an object detection network. Like the EAM, Wu et al. [25] propose a cataract detection paradigm based on eyeball detection and a multi-task learning network.

Nevertheless, the above approaches face dilemmas when the location labels of lesions and eyeballs are unavailable since they greatly rely on supervised detection for further categorization. Moreover, these works do not comprehensively explore fine-grained ocular abnormalities, while our work mainly focuses on categorizing tiny ocular lesion problems in a feature-disentanglement paradigm. Many works have investigated weakly-supervised deep models for medical image analysis to alleviate the reliance on strong annotations.

B. Attention Mechanism and Zoom-In Strategy

Plenty of works have explored the weakly-supervised mechanism for recognition, such as Grad-Cam [26], spatial attention mechanism variants [15, 27], and Zoom-in-based architecture [28, 29]. The attention-based Zoom-in [28, 29] models first locate the saliency map and crop and then zoom into it for diagnosis. The key idea to zoom is to generate robust attention to guide the zoom-in process and consequently assist models in capturing more detailed features for subtle lesions. Following similar ideas, Xing et al. [17] constructed an Attention Guided Deformation Network (AGDN) for wireless capsule endoscopy recognition. To generate valuable augmentations and compress background noise, Li et al. [15] propose a multi-scale attention-guided augmentation framework, namely MAG-SD, for the chest X-ray image categorization.

Nonetheless, the above methods neglect that the network is apt to learn coupled features from local regions and significantly degrade, for instance, when merely subtle and partial lesions are visible, as shown in the first row of Fig. 1. Contrary to the above works, we utilize the attention mechanism to reduce the reliance on strong supervision when decoupling features.

C. Contrastive Representation Learning in Medical Visual Recognition

Contrastive representation learning has been widely explored recently. It achieves superior performance in self- or semi-supervised training of deep models for medical images [30, 31]. However, contrastive learning without considering relationships among samples impedes the ability of contrastive loss. Combining supervised learning and contrastive learning, Khosla et al. [32] proposed a SupCon Loss for natural image categorization. Hu et al. [33] developed a supervised local contrastive loss to leverage limited pixel-wise annotation for cardiac segmentation following the SupCon [32]. Chartsias et al. [34] proposed using contrastive learning to mitigate the labeling bottleneck for view classification of echocardiograms. The SupCon demonstrates the potential to facilitate training. However, directly implementing SupCon in medical FGIC tasks fails to work well, as shown in Table VI due to the intrinsic complexities of the FGIC dataset, which is consistent with the discovery in [35]. One of the solutions to handle the FGIC task is to expand further the positive and the negative sets for contrastive learning. However, the SupCon highly relies on task-specific and strong augmentations, impeding its capacity in the medical domain, as proven in [32, 33, 34]. Moreover, none of the above works focus on the feature disentanglement in the ocular FGIC problem.

D. Feature Disentanglement in Medical Visual Recognition

By dividing independent features into different components, the feature disentanglement mechanism aims at reducing feature coupling with each other for visual tasks [18, 19, 36, 37]. Previous works try to capture features in latent space utilizing adversarial learning [38] or auto-encoder mechanism [39], which are apt to fall into mode collapse and sensitive to hyper-parameters [40]. These works lack the explicit feature decomposition (e.g., lesion regions), leading to a sub-optimal disentangled feature for the downstream task (e.g., FGIC task). Explicitly decomposing inputs, Zhao et al. [19] devised a local and global feature disentangled network to identify the benign and malignant thyroid nodules. Feng et al. [20] employs the triplet loss [41] and proposed a two-stream framework consisting of a recognition stream and a classification stream for the diagnosis of vertebral compression fracture. To tackle the training instability due to misleading correlation between noises and lesions, Huang et al. [21] proposed decoupling the correlation using supervision from the bounding box. However, the above works highly rely on the supervision of segmentation labels that are unavailable in our dataset.
In this study, inducing feature disentanglement via the CMZ strategy and the HCD-Loss is essential to advance diagnostic performance. The CMZ strategy caters to offering challenging views for feature disentanglement without expensive annotations, which is distinct from the previous works. The HCD-Loss devotes itself to the output feature disentanglement for better lesion recognition collaborating with the CMZ strategy. Our CDNet simultaneously considers both the input and the output disentanglement that distinguishes itself from previous works while achieving a promising result for the OUS task.

III. METHODOLOGY

A. Overview of Our Method

As shown in Fig. 2, our framework consists of four parts, including the (a) global branch, (b) local branch, (c) aligning and decoupling output, and (d) fusion and prediction. The MSFE in (a) first extracts multi-scale features of global views during training and then marks saliency regions (e.g., lesions) via the WSLL in a weakly-supervised manner. Subsequently, the CMZ strategy considering the location of saliency regions in the local branch generates multi-zoom local views of the global views to reduce the coupling among views. The MSFE in the local branch takes these disentangled views as the second inputs and outputs of lesion-specific features. In part (c), three shared-weights multi-layer perceptrons (MLP) armed with L2-normalization serve as the project network for embedding these features in a hypersphere space. The HCD-Loss promotes the global and local branches to pull, push, and decouple varied views to obtain aligned and disentangled features. In part (d), a fully-connected categorization head (FC Head) is trained on top of the frozen backbone (i.e., the MSFE with WSLL) via a CE-Loss. Concatenated features are fed to the FC head to make image-level predictions. Notably, in the test time, the local branch only zooms into the attention center of the global view. It extracts local features for fusion and prediction, which is not depicted in Fig. 2 for simplicity.

We will introduce the main modules of our method in the following paragraphs, including the WSLL module, CMZ strategy, and HCD-Loss.

B. Weakly-Supervised Lesion Localization Module

In image categorization tasks, mere category annotations are available, and lots of works have explored the locating mechanism [15, 26, 27, 28, 29]. Aiming to utilize feature maps from hierarchical scales, we propose the WSLL module operated on multi-stages of the backbone, including the feature maps $F_{Text}$ and $F_{Sem}$ with sizes of $14 \times 14$ and $7 \times 7$. However, a direct concatenating of features is not an optimal strategy to capture semantic features, as the literature proves [42]. Feature maps from different scales are selected as Texture Adaptor and Semantic Adaptor inputs, respectively, to extract hybrid spatial attention maps. Moreover, we propose a Multi-Channel Spatial Pooling (MCSP) for weighting and squeezing the features, as illustrated in Fig. 3. The MCSP takes the semantic features as input, weights via the multi-stage spatial attention $Att_{MS}$ then pool them. The texture spatial attention $Att_{Text}$, the semantic spatial attention $Att_{Sem}$, and multi-stage spatial attention $Att_{MS}$ are calculated by (1), (2), and (3), respectively. Thus, the output feature $F_{out}$ of our WSLL is summarised as (4).

The output feature $F_{out}$ of our WSLL is summarised as (4).

\[
\begin{align*}
Att_{Text} &= Pool(\sigma(BN(Conv_{1 \times 1}(F_{Text})))) \\
Att_{Sem} &= \sigma(BN(Conv_{1 \times 1}(F_{Sem}))) \\
Att_{MS} &= Att_{Text} + Att_{Sem} \\
F_{out} &= MCSP(F_{Sem}, Att_{MS})
\end{align*}
\]

where $Pool$, $\sigma$, $BN$, and $MCSP$ represent average pooling, sigmoid activation, batch normalization, and MCSP operator respectively. Formally, the MCSP is calculated by (5).

\[
MCSP(F, A) = C1D_{1}(Concat_a(\{GAP(F \otimes A_{MS})\}_{a=1}^{MS}))
\]

where $C1D_{1}$ and $Concat_a$ represent 1D convolution with kernel size 1 and the function that concatenates spatial-pooled features along the channel axis, respectively. The $MS$ denotes the channel number of the multi-stage spatial attention map (i.e., $Att_{MS}$), which is set to 32 in this paper. The WSLL can locate lesions in a weakly-supervised manner for further lesion decomposition with the CMZ strategy.

C. Contrastive Multi-Zoom Strategy

1) Motivation of the CMZ Strategy: Extracting disentangled features can enhance recognition as not all discriminative regions (e.g., lesions) appear [19, 43, 44]. Intuitively, the various lesion scope induces significant intra-class variance. For instance, the VH images in the first and second row are apt to disturb the training of coarse-grained networks, as depicted in Fig. 1. Our insight is that the categorization capacity of networks can be enhanced if partial lesion regions are recognized well, as it is a frequent scene in the OUS task. More specifically, low-overlapping and disentangled inputs are more suitable for partial lesion recognition than the entangled ones. Nonetheless, the widely-used attention-guided Center Zoom mechanism only zooms in on the whole suspicious area, which increases the difficulties of feature decomposition and the probability of ignoring subtle regions [19, 44]. We delicately designed our CMZ strategy for reducing the intersection of the local-branch inputs inspired by Contrastive Crop [45], which is proposed first in self-supervised learning (SSL) for data augmentation.

2) Entangled and Disentangled Lesions: As depicted in Fig. 4, the entangled views can be disentangled into varied components, miming the circumstance when only partial lesions are visible. Therefore, the MSFE is forced to recognize tiny and

![Fig. 3. Architecture of the proposed WSLL. The $B$, $C_1$, and $C_2$ are the batch size, the channel number of the $F_{Text}$ and $F_{Sem}$, respectively. In practice, the $F_{out}$ will be squeezed to $B \times C_2$ for output.](Image)
Algorithm 1: Contrastive Multi-Zoom Strategy.

Input: \( T \): Input images
\( \text{Att}_{MS} \): Multi-stage spatial attention maps
\( Z_n \): Zooming times
\( k \): Binarization threshold of the attention
\( \alpha \): Parameter of Beta distribution
\( C_s, C_r \): The scale and aspect ratio for cropping

Output: \( \{e_z\}_{z=1}^{Z_n} \): \( Z_n \) contrastive zoomed views

1. Average the multi-stage spatial attention maps along the channel dimension, which thus will be squeezed into one:

\[
\text{Att}_S = \text{Average}(\text{Att}_{MS})
\]

2. Generate bounding boxes via convex hull algorithm for the attention region larger than the threshold \( k \):

\[
B = \text{GenBBBox}(\text{Att}_S, k)
\]

3. for \( z \leftarrow 1 \) to \( Z_n \) do

4. Zoom in \( Z_n \) times via Contrastive Crop, \( \alpha \)

controls how far the cropping position deviates from the bounding box center:

\[
\{x_z, y_z, h_z, w_z\} = \text{ContrastiveCrop}(T, \text{Att}_S, B, \alpha)
\]

end

6. Set the left-top start point position for zoom-in. \( X \) and \( Y \) denote the coordinate sets of the \( Z_n \) points:

\[
X = \{x_z\}_{z=1}^{Z_n}, Y = \{y_z\}_{z=1}^{Z_n}
\]

7. Set the height and width of cropping rectangle. \( H \) and \( W \) represent the height and width sets:

\[
H = \{h_z\}_{z=1}^{Z_n}, W = \{w_z\}_{z=1}^{Z_n}
\]

8. Zoom to the coordinates and scales, then resize the views:

\[
\{e_z\}_{z=1}^{Z_n} = \text{ZoomAndResize}(T, X, Y, H, W)
\]

return \( \{e_z\}_{z=1}^{Z_n} \)

Fig. 4. Visualization of the entangled lesions and the disentangled lesions. The regions masked by varied colors represent different parts of lesions.

partial regions while extracting decoupled features from varied views.

Specifically, the CMZ strategy generates multi-zoom views away from the center of the attention while maintaining a little focus on the most discriminative region. More details are described in the Algorithm 1. Our CMZ strategy can significantly reduce the overlapping between cropped views, simulating the condition when only tiny lesions are present in the OUS image.

For another, such simulation can facilitate the decomposition of features in the input aspect, thus promoting recognition performance. Center Zoom replaces the Contrastive Crop for the test time since it is more important to diagnose the most salient regions for the final diagnosis. Notably, only the CMZ strategy is not enough for an optimal feature disentanglement of the FGIC task, as proven in Section IV.

D. Hyperspherical Contrastive Disentangled Loss

1) Motivation of the HCD-Loss: Matrix operation is one of the cornerstones of current deep learning-based methods. It has been widely employed in many visual tasks due to its depicting capacity of relationships between entities [32], [46], [47], [48], [49]. Based on these observations, we propose utilizing the matrix operation’s fantastic similarity-measuring capacity for pulling, pushing, and decoupling the different terms. Specifically, to stimulate robust and decoupled feature representation, we devise a novel loss function, HCD-Loss. The HCD-Loss aims to pull together clusters of points from the same category, push apart clusters of points from different categories, and decouple features from the same global views. As depicted in Fig. 5, views (e.g., \( A_{ij}^c \) and \( A_{ij}^s \)) from the same categories are pulled together and views (e.g., \( B_{ij}^c \) and \( C_{ij}^c \)) from varied categories are pushed apart, while views (e.g., \( B_{ij}^c \) and \( B_{ij}^v \)) from the same image are forced to decouple with each other.

We denote our multi-scale feature extractor as \( f(. \) and the MLP with \( L2Norm \) as \( g(.) \). Our goal for the FGIC task is to learn a fine feature embedding network from the training dataset \( D = \{(x_i, y_i)\}_{i=1}^{N} \), where \( N \) represents the number of our training set samples. The \( f \) extracts a high-dimension (e.g., 2048) feature from the input \( x_i \). Then \( g \) projects the feature to an L2-normalized d-dimensional (e.g., 128) embedding \( r_z \), such that \( r_z = g(f(x_i)) \in \mathbb{R}^d \) and falls on the unit hypersphere. Let \( \mathcal{I} \) denote the randomly sampled batch images during training from \( D, \mathcal{P}(i) \) denote the index set of all positive samples of the anchor sample \( i \) (i.e., samples with the same category as \( i \)). The negative
pairs in the batch are images with distinct labels to the anchor sample \(i\), the indices of which can be denoted as \(N(i)\). Formally, \(P(i)\) and \(N(i)\) are denoted as (6) and (7). One naive approach to utilizing contrastive learning for our task is directly conducting image augmentations to a batch and applying contrastive loss. However, such a procedure orienting self-supervised manners neglects the multi-zoomed views (e.g., row (c1) and (c2) in Fig. 5) from the batch and increases the augmentation processing burden.

\[
P(i) = \{ j \mid \forall j \in I, y_j = y_i, \text{and } j \neq i \} \tag{6}
\]

\[
N(i) = \{ j \mid \forall j \in I, y_j \neq y_i \} \tag{7}
\]

2) Extending the Positive and the Negative Samples: To extend samples and facilitate aligning embeddings, we propose to utilize the global views and the corresponding multi-zoomed views. Specifically, the \(C_P(i)\), denoted in (8), represents the contrastive multi-zoomed views from \(i\). The \(N_N(i)\), denoted in (9), represents contrastive multi-zoomed views from each \(N(i)\). Thus the extended positive samples set \(P(i)\) and negative samples set \(N(i)\) can be denoted in (10) and (11). This way, the positive and negative samples are substantially extended for robust feature representation.

\[
C_P(i) = \{ c \mid c = CMZ(i, Att_i, B(i), z) \} \mid \text{if } z = 1 \tag{8}
\]

\[
C_N(i) = \{ c \mid c \in C_P(j), \forall j \in N(i) \} \tag{9}
\]

\[
P'(i) = P(i) \cup C_P(i) \tag{10}
\]

\[
N'(i) = N(i) \cup C_N(i) \tag{11}
\]

3) Pulling and Pushing for Discriminative Features: To extract closely aligned and discriminative representations to all entries from varied classes, we propose the \(L_{pp}(P', N')\). Formally, the \(L_{pp}\) orienting the OUS task is expressed as (12).

\[
L_{pp}(P', N') = \sum_{i \in I} \frac{1}{|P'(i)|} \sum_{p \in P'} \mathcal{M}(i, p) \tag{12}
\]

where the \(\mathcal{M}(i, p)\) represents the similarity between \(i\) and \(p\). The \(\mathcal{M}(i, p)\) is denoted in (13).

\[
\mathcal{M}(i, p) = \log \frac{\exp(s_t \cdot \cos(\theta_i, \theta_p))}{\sum_{a \in P \cup N} \exp(s_t \cdot \cos(\theta_i, \theta_a))} \tag{13}
\]

where the \(s_t\) denotes the temperature scaling factor, and the \(\cos(\theta_i, \theta_j) = \theta_i \cdot \theta_j\) denotes cosine similarity, since the \(\theta_i\) is normalized. Therefore, the negatives summation in the denominator of \(\mathcal{M}\) highly contributes to the loss when negatives are continually wrongly categorized, spurring a closer feature embedding.

4) Decoupling Outputs Features: The \(L_{pp}(P', N')\) results in a more robust clustering of the embedding space than that generated from \(L_{pp}(P, N)\), as Section IV proves. Nonetheless, merely \(L_{pp}\) for pulling and pushing will lead to a sub-optimal representation due to neglecting feature disentanglement. Therefore, the matrix between the representation of multi-zoom views is calculated to further decouple different dimensions of features in the aspect of outputs. Formally, the decoupling loss \(L_{dc}\) is defined as (14).

\[
L_{dc}(C_P) = d \times \sum_{x} \sum_{y \neq x} \cos^2(V(c_1, x), V(c_2, y)) \tag{14}
\]

where the \(d \in [0, 1]\) represents the balancing hyperparameter, the \(c_n \in C_P\) represents the contrastive local views, and the \(V(c_n, x) \in \mathbb{R}^{h \times 1}\) represents the dimension vector composed of the \(x\)th dimension features of a batch of zoomed views. The \(V(c_n, x)\) is \(L_2\)-normalized before calculation, and the \(b\) denotes the batch size. Intuitively, the \(\cos^2(V(c_1, x), V(c_2, y))\) is limited to zero (i.e., orthogonality) for decoupling features in varied dimensions. The \(\cos(V(c_1, x), V(c_2, y))\) is forced to 1, as the specific \(x\)-dimension features from the same image are expected to be parallel. We only employ the \(L_{dc}\) between contrastive multi-zoom views (i.e., \(c_n\)) instead of global images since the multi-zoom views are highly entangled with the global while decoupling them tends to overfit background noises.

5) Joint Loss Function: Specifically, the outputs of the MLP are used as inputs to the HCD-Loss for training the \(f\), as illustrated in Fig. 2. Therefore, our joint HCD-Loss is denoted as (15).

\[
L_{HCD} = L_{pp}(P', N') + \lambda L_{dc}(C_P) \tag{15}
\]

where \(\lambda\) is the trade-off hyperparameter. We further investigate the selection of \(\lambda\) in Section IV-C4. Such that the HCD-Loss can contribute to aligning and decoupling features.

### IV. Experiments and Discussions

#### A. Data and Preprocessing

1) Materials: Although there are some related studies on the categorization task of the OUS images, there is no large-scale, fine-grained labeled public dataset with a standardized collection method available. In this work, ZJUOUSD was built as a large-scale dataset based on the OUS images collected from The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. In terms of image annotation, two ophthalmologists with more than five years of clinical experience initially labeled the ocular ultrasound images. Another senior ophthalmologist with more than 15 years of clinical experience examined the label sets. If there are disagreements in label datasets, the senior ophthalmologist should review and unite them. Moreover, patients’ medical notes and follow-up information have been taken into consideration to confirm the diseases further. This study is conducted in compliance with the Declaration of Helsinki. The research protocols are approved by the ethics committee of the Second Affiliated Hospital, Zhejiang University School of Medicine (IRB: 2020-741). Our proposed CDNet is validated on our ZJU Ocular Ultrasound Dataset (ZJUOUSD), whose more details are depicted in Table I.

Furthermore, the generalization ability of CDNet is evaluated on two public and widely-used chest X-ray FGIC datasets: Chest X-Ray Images for Classification Dataset (CXR) [50] with 5856 cases and COVID-19 RADIOGRAPHY DATABASE...
(COVID-RD) [51], [52] with 4035 cases, whose more details are shown in Table II. The CXRD is taken from patients with normal lungs (normal), bacterial pneumonia (BP), and viral pneumonia (VP). Following the previous work [53] for chest x-ray categorization, the initial test set is sub-divided into a valuation subset and a test subset in this work. The COVID-RD is initially composed of 3616 COVID-19 positive cases (C19), 6012 lung opacity cases, 10192 normal cases (Normal), and 1345 viral pneumonia (VP) CXR images. Following the division protocol of literature [54], we select C19, Normal, and VP for analysis, as depicted in Table II.

2) Preprocessing and Augmentation: The original ultrasound images have small but misleading icons, including details about the patients and the imaging devices’ parameters, confusing the model’s optimization. Thus we remove background noises, crop the central region from original ultrasound images via setting a specific cropping location for each imaging device and resize all of them to 224 × 224 for reasoning, same for CXRD and COVID-RD. Besides, data augmentation is well-known for remitting the impact of overfitting, thus improving the robustness of deep networks since the labeled data is precious in medical FGIC tasks. Specifically, we implement the common data augmentation mechanism via a widely-used tool [55]. Augmentation transformations for each training phase are randomly selected from a pre-defined sequence of transformation lists: such as Rotation, Flip, Affine transformation, and CLAHE.

B. Performance Comparison

1) Network Training and Evaluation Metrics: Our experiments were based on the PyTorch framework and NVIDIA RTX 2080Ti GPUs. We trained our network from scratch for 128 epochs, and the Adam algorithm (momentum = 0.97, weight decay = 5 × 10^{-4}) updated the parameters. We adopted a batch size of 64 and set the learning rate as 1 × 10^{-3}. To evaluate the classification performance, we adopted the commonly used metrics in this study, including F1-score, Accuracy, Precision, Sensitivity, and Specificity.

- F1-score (F1):
  \[
  F1-score = \frac{2 \times TP}{2 \times TP + FP + FN}
  \]

- Accuracy (Acc):
  \[
  Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
  \]

- Precision (Pre):
  \[
  Precision = \frac{TP}{TP + FP}
  \]

2) Backbone Selection: In order to show the superiority of the proposed method, we compare the classical coarse-grained categorization networks including VGG [10], Inception v3 [11], ResNet [12], DenseNet [13], and EfficientNet [14] since there is little previous work focusing on the large-scale fine-grained OUS image categorization. As shown in Table III, a large-scale benchmark of our ZJUOUSD is reported on classical deep learning-based coarse-category classification networks. We select the EfficientNet-b3 as our backbone, as it achieves the state-of-the-art performance among the classical coarse-grained categorization networks. Our model specializes in the fine-grained OUS visual categorization task compared with classic coarse-grained models.

3) Comparison With Other Methods: To further investigate the OUS categorization task, state-of-the-art lesion categorization and localization networks, including MAG-SD [15], PCAN [56], and Probabilistic-CAM [57], were re-implemented for experiments. Meanwhile, MAG-SD based on EfficientNet is also reported for a fair comparison. The original MAG-SD achieves a 91.8 ± 1.44% F1 score on our ZJUOUSD and a 92.52 ± 0.46% F1 score based on EfficientNet-b3. Also, the PCAN achieves a 92.27 ± 1.04% F1 score, and Probabilistic-CAM achieves a 93.09 ± 0.15% F1 score on our ZJUOUSD. Furthermore, our HCD-Loss is replaced by SupCon Loss [32] (i.e., CDNet (SupCon) in Table III) for investigating the functionality of the feature disentanglement. Significant performance advancement is observed in our CDNet armed with HCD-Loss, which illustrates that only input decoupling is suboptimal for the OUS task. Specifically, our CDNet obtains a 94.13 ± 0.03% F1 score and 94.40 ± 0.05% accuracy, which is the highest in the benchmarks of the ZJUOUSD. As shown in Table III, our CDNet

| Dataset       | Class | Normal | BP (C19) | VP | Total |
|---------------|------|--------|---------|----|-------|
| CXRD          | Train| 1341   | 2530    | 1345| 5216  |
|               | Validation| 121 | 125    | 74    | 320   |
|               | Test  | 121    | 125    | 74    | 320   |
|               | Total | 1333   | 2780   | 1493 | 5585  |
| COVID-RD      | Train| 940    | 940    | 940  | 2820  |
|               | Validation| 200 | 200    | 200   | 600   |
|               | Test  | 205    | 205    | 205   | 615   |
|               | Total | 1345   | 1345   | 1345 | 4035  |

- Sensitivity (Sen):
  \[
  Sensitivity = \frac{TP}{TP + FN}
  \]

- Specificity (Spe):
  \[
  Specificity = \frac{TN}{TN + FP}
  \]
Typically, state-of-the-art coarse-grained CNN models were trained and evaluated on the same protocol. Components were trained and evaluated on the same protocol. Components validations were reported in Table VI.

C. Ablation Study of Framework Design

Our method components could be summarized as the WSLL module, CMZ strategy, and HCD-Loss. Each component was studied by evaluating its improvement in categorization performance, which is quantified by the metrics above. The following procedure obtained the performance gain: the proposed model was first trained on a specific dataset with metrics. Then, the single component was substituted or removed and validated the same dataset for a fair comparison. All the illustrated models were trained and evaluated on the same protocol. Components validations were reported in Table VI.

1) Efficacy of Weakly-Supervised Lesion Localization (WSLL): Typically, state-of-the-art coarse-grained CNN models suffer from similar and entangled features when dealing with FGIC tasks. Existing works toward the feature disentanglement highly rely on the lesion location, which could be effectively located by our WSLL module. Therefore, we first investigate the functionality of our WSLL, which provides our CDNet with lesions localization under a weakly-supervised protocol for further feature disentanglement. With the assistance of WSLL, the EfficientNet-b3 obtains superior performance on all categorization metrics than the baseline on the three datasets, as depicted in Table VI. As for the EfficientNet-b3 equipped with WSLL, F1 reached 91.74% in the ZJUOUSD dataset. Moreover, our CDNet armed with WSLL is capable of capturing global inspection and sensitive to tiny ocular lesions, as demonstrated in Fig. 8. The abnormal regions are subtle, while our CDNet is capable of detecting the lesions, for instance, the IOT row, where a mere tiny tumor is present.

Notably, the WSLL module is plug-and-play and independent of network architecture and input modality. It is capable of locating the lesions under various circumstances, e.g., x-rays. With the guidance of the WSLL, the proposed CDNet is capable of categorizing OUS images while locating lesion regions based on the saliency maps. Moreover, the quantitative and qualitative results demonstrate the efficacy of our WSLL while depicting the diagnosis consistency with experienced ophthalmologists.

2) Efficacy of Contrastive Multi-Zoom (CMZ): The attention mechanism emphasized local features that can guide examinations further and provide more interpretable decisions. The vanilla Zoom-in [15], [28] model that zooms into the center is expected. In contrast, our method comprehensively decouples the feature at both the input and the output aspects.

Consider the radiological characteristics of the datasets above, we compared our method with the current deep learning-based image-level classification methods [53], [54], which reveal the preliminary benchmarking results and corresponding state-of-the-art performance. With the assistance of the feature disentanglement, our CDNet surpasses the SEME-DenseNet169, the MAG-SD, and the 3D Non-Local DenseNet on all categorization metrics, as depicted in Tables IV and V.

Table IV: Comparison of the State-of-the-Art Models on the CXRD. The results of Base and SEME-Based Models are Borrowed from the Literature [53]. The Best Results are Shown in Bold

Table V: Comparison of the State-of-the-Art Models on the COVID-19. The Results of DenseNet-Based Models are Borrowed from the Literature [53]. The Best Results are Shown in Bold

Table VI: Ablation Study of Each Key Component

Table VII: The Comparison Results Between the Vanilla Zoom-In and Our CMZ Strategy on ZJUOUSD. Notably, Our HCD-Loss is Not Utilized Here

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
We proposed the HCD-Loss for the output enactment of our HCD-Loss. All settings are frozen except the λ to investigate the disentangled efficacy of the COVID-RD for λ, which is set to 0.2, 0.05, 0.1, and 0.4. The best results are shown in bold. Considering that λ is distributed from 0.05 to 0.2, the λ is suggested to be fine-tuned between 0.05 and 0.2 for a new dataset.

### TABLE VIII

| Dataset | λ | FI (%) | Acc (%) | Pre (%) | Sen (%) | Spe (%) | AVG (%) |
|---------|---|--------|---------|---------|---------|---------|---------|
| ZJUOUSD | 0.05 | 94.07±0.04 | 93.87±0.05 | 93.30±0.06 | 93.90±0.03 | 93.59±0.02 | 93.30±0.05 |
|         | 0.4 | 94.13±0.03 | 94.08±0.05 | 94.43±0.06 | 94.35±0.05 | 94.59±0.03 | 94.14±0.02 |
|         | 0.2 | 94.32±0.06 | 94.33±0.09 | 94.30±0.12 | 94.17±0.13 | 94.39±0.19 | 94.23±0.19 |
|         | 0.4 | 94.49±0.01 | 92.07±0.15 | 92.48±0.29 | 91.67±0.20 | 91.61±0.19 | 91.95±0.11 |
| CXRD    | 0.02 | 93.84±0.02 | 93.85±0.01 | 93.41±0.03 | 93.10±0.01 | 93.41±0.09 | 93.32±0.10 |
|         | 0.09 | 94.13±0.00 | 94.89±0.05 | 94.83±0.06 | 94.61±0.00 | 94.16±0.46 | 94.55±0.59 |
|         | 0.1 | 94.66±0.05 | 94.61±0.18 | 94.89±0.02 | 94.70±0.41 | 94.74±0.13 | 94.99±0.08 |
|         | 0.2 | 95.16±0.47 | 95.39±0.54 | 96.06±0.43 | 96.75±1.24 | 95.42±0.40 | 95.07±0.50 |
|         | 0.4 | 94.45±0.91 | 94.71±0.00 | 94.70±0.80 | 94.58±0.64 | 94.64±0.42 | 95.01±0.29 |
| COVID-RD| 0.04 | 91.64±0.10 | 91.83±0.10 | 91.62±0.11 | 91.43±0.10 | 91.24±0.05 | 91.34±0.09 |
|         | 0.05 | 94.38±0.33 | 94.31±0.35 | 94.39±0.32 | 94.37±0.33 | 94.39±0.27 | 94.75±0.17 |
|         | 0.1 | 95.00±0.49 | 95.00±0.50 | 95.00±0.46 | 95.00±0.50 | 95.00±0.25 | 95.00±0.44 |
|         | 0.2 | 95.00±0.47 | 95.00±0.47 | 95.01±0.46 | 95.00±0.47 | 95.00±0.24 | 95.00±0.16 |
|         | 0.4 | 97.99±0.32 | 97.99±0.32 | 97.92±0.31 | 97.99±0.32 | 97.90±0.26 | 97.93±0.19 |

3) Efficacy of Hyperspherical Contrastive Disentangled Loss (HCD-Loss): We proposed the HCD-Loss for the output disentanglement to obtain robust and disentangled representations for recognition. Three relationships among views, including pulling, pushing, and decoupling, are explored in our study. The term armed with HCD-Loss achieves promising performance in all three datasets, as depicted in Table VI. Our experiments indicate that only inputs feature decomposition is apt to be disturbed by classical CE-Loss, which is thus substituted by our novel HCD-Loss. Furthermore, we replace HCD-Loss with SupCon loss [32] to investigate the disentangled efficacy of the proposed HCD-Loss. Replacing HCD-Loss with the original SupCon loss deteriorates the main metrics, indicating that a straightforward substitution of SupCon loss without considering local views relationship and feature disentanglement struggles in FGIC tasks, as shown in Table III. Specifically, the CDNet with SupCon obtains an average 92.37% FI score, far inferior to that of the HCD-Loss. In contrast, our HCD-Loss complements our CMZ strategy regarding the outputs aspect, further demonstrating the importance of the simultaneous feature disentanglement in both the inputs and outputs aspects.

Besides, the global and local branches share the same name and train with different scales of the augmented images, from which the network views different lesions and makes various errors, HCD-Loss serves as an implicit constraint among various views due to the difficulties of self-correcting, promoting the network to extract robust and disentangled representations. Though the global attention did not locate the detached retina well, the local attention precisely located the partial retina for further diagnosis, as illustrated in the RD column in Fig. 8. Utilizing the common matrix operation promotes the feature decomposition via forcing different-dimension features of different views to be orthogonal and the same-dimension features of different views to be parallel. The HCD-Loss collaborates with the CMZ strategy for feature disentanglement in both the output and the input aspects.

4) Impact of the Hyperparameter λ: We conduct a series of hyperparameter tuning for better feature disentanglement and representation capacity, validating the trade-off functionality of the λ of our HCD-Loss. All settings are frozen except the λ of the HCD-Loss for condition control. All metrics above are depicted in Table VIII, from which we can observe that both too large and too small degrade the metrics of our CDNet, for instance, when

---

**Fig. 6.** Visualization of the vanilla zoom in strategy and our proposed CMZ strategy on the ZJUOUSD. The (G), (Z), (C1), and (C2) represent the global image with retinal detachment, vanilla zoom-in image, and two images generated by our CMZ strategy, respectively. The ocular abnormalities are marked with yellow arrows by medical experts. Moreover, the zoom-in image and the two images from the CMZ strategy are marked by green, blue, and orange bounding boxes, which are consistent with the boxes in the global image. Compared to the vanilla zoom-in image in (G), the two generated images in (C1) and (C2) are with more varied backgrounds and more tiny lesions to be detected, which effectively mimics the fine-grained challenge.

**Fig. 7.** Visualization of the feature distribution of ZJUOUSD via UMAP. The three main variants of our models and the baseline are illustrated. Each variant’s Davies-Bouldin score (DB score) is also depicted. The minimum DB score, in theory, is zero, and lower values indicate better clustering performance.

**Fig. 8.** Visualization of the attention heatmaps of the experimental datasets, where the left is the ZJUOUSD, the middle is the CXRD, and the right is the COVID-RD dataset. The abnormalities are marked with yellow arrows in rows (A) and (G) by medical experts. The brighter the heatmap is, the more critical it illustrates. Row (A) and (B) represent the global views and corresponding attention maps. Row (C) and (D) represent the local views (at the attention center) and corresponding attention maps.
Fig. 9. Three charts of confusion matrices generated by the proposed CDNet. The confusion matrices depict the distribution of predictions in three datasets. The color of confusion matrices relies on the normalized values of predictions for a better visualization, which is placed at the right of each matrix. The number of predictions is placed above the normalized values.

λ is set to 0.02 or 0.4. Therefore, we suggest that the λ can be fine-tuned between 0.05 and 0.2 for a new dataset.

5) Analysis of Feature Distribution: We visualize the features before the FC head of each model using UMAP [59], a technique to visualize high-dimensional features by identifying critical structures in the high-dimensional space and reserving them in the lower-dimensional embedding (e.g., 2-dimension). Compared with coarse-grained (i.e., Fig. 7(a)), variants armed with WSLL (i.e., Fig. 7(b)–(d)) have closer feature distribution among normal and ocular abnormalities.

Moreover, we further introduce the Davies-Bouldin Score (DB Score) [60] to measure the clustering performance, demonstrating the feature-aligned capacity of the variants of our model. All the lower-dimensional embeddings are normalized before calculating DB Score. Formally, the DB score for a cluster $c_i$ is expressed as (21).

$$DBscore = \max_{c_i \neq c_j} \frac{Intra(c_i) + Intra(c_j)}{Inter(c_i, c_j)}$$ (21)

where $Intra(c_i)$ represents the intra-category variance of the cluster $c_i$, and $Inter(c_i, c_j)$ represents the distance between the cluster centers of $c_i$ and $c_j$. The $Intra(c_j)$ is calculated by averaging squared deviations from the category center. Intuitively, closer and less incorrect clustering feature distribution declares finer recognition. As shown in Fig. 7, the CDNet armed with the HCD-Loss obtains the minimum DB Score (0.3615) among all variants, depicting a closer cluster for categorization than others. Specifically, the feature distribution with the maximum DB score (i.e., 0.8772) from the baseline (i.e., Fig. 7(a)) is more dispersed than the others. An improvement in the DB score can be observed in Fig. 7 comparing the (c) and the (d), where closer RD cases are visible and less incorrect VH cases are entangled with normal and PSS cases.

6) Analysis of Prediction Distribution: The distribution of prediction results is essential for assessing the efficacy of models, which is illustrated via the confusion matrix in Fig. 9. The CDNet is selected to obtain the charts to display the categorization result of deep learning models. It can be implied that the CDNet can extract robust and representative features from various samples in varied modalities as most samples are on the diagonal of matrices. Moreover, these results are consistent with the feature distributions in Fig. 7, indicating that the visible differences between IOT and PSS are incredibly subtle.

7) Analysis of Lesion Localization Capacity: To have a quantitative result for the localization capacity of our method, our ophthalmologists further spent several months labeling the bounding boxes of our ZJUOUSD test set, consisting of 1042 images. Consequently, we firstly apply a threshold to binarize the attention maps from our CDNet. Then bounding boxes covering the region of interest (e.g., lesion regions) are generated for the localization task. The localization capacity of each model is based on evaluating the Intersection over Union (i.e., IoU) metric, as depicted in Table IX. The IoU could be defined by (22), where the $A_t$ and $A_l$ represent the intersection area and the union area between the localization prediction and the ground truth, respectively. Results in Table IX demonstrate the lesion localization efficacy of our CDNet and achieves a significant potential for accurate lesion localization. Thus our CDNet will be very helpful for ophthalmologists in clinical practice.

$$IoU = \frac{A_t}{A_t + A_l}$$ (22)

D. Limitations of Our Work

In this work, we demonstrated the efficacy of our fine-grained categorization network, the CDNet, which significantly surpasses the previous methods in three datasets. Nonetheless, there are some limitations to our study. 1) Our CDNet requires triple forward propagations (typically, one global input and two zoom-in inputs) for training, which increases the training time. 2) 3D data, for instance, CT and MRI volumes, are not considered in our current work. 3) The CDNet is a weakly-supervised learning method that must be trained with category-labeled data. New disease occurrences or rare diseases without large labeled images may not be classified well, while the self-supervised mechanism may be a solution to alleviate the limitations. Therefore, we will attempt to reduce the forward times and extend the CDNet to other visual tasks in future works. Moreover, the few-shot and zero-shot settings will be explored to investigate the feature disentanglement capacity.
In this work, the OUS recognition tasks are combined as a five-category classification task, for which we propose a novel CDNet in a feature disentanglement paradigm. The WSSL enables our CDNet to locate the lesion without relying on strong annotations while providing an explicable diagnosis decision. The CMZ strategy and the HCD-Loss collaborate for feature disentanglement. We point out that disentangling both input images and output representations can facilitate the performance of ocular abnormalities diagnosis. We demonstrate the CDNet, which significantly surpasses the previous methods on three datasets. More challenging settings will be investigated for our CDNet in the future.

V. CONCLUSION

REFERENCES

[1] C. Stannard, W. Sauerwein, G. Maree, and K. Lecuona, “Radiotherapy for ocular tumours,” Ecre, vol. 27, no. 2, pp. 119–127, 2013.
[2] K. Ohno-Matsui and J. B. Jonas, “Posterior staphyloma in pathologic myopia,” Prog. Retinal Eye Res., vol. 70, pp. 99–109, 2019.
[3] G. Blindness, “Vision impairment, C; vision loss expert group of the global burden of disease, S., trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the global burden of disease study,” Lancet. Glob. Health, vol. 9, no. 2, pp. e130–e143, 2021.
[4] A. P. Marques et al., “Global economic productivity losses from vision impairment and blindness,” EclinicalMedicine, vol. 35, 2021, Art. no. 100852.
[5] J. E. W. Koh et al., “A novel hybrid approach for automated detection of retinal detachment using ultrasound images,” Comput. Biol. Med., vol. 120, 2020, Art. no. 103704.
[6] R. Neupane, R. Gaudana, and S. H. Boddu, “Imaging techniques in the diagnosis and management of ocular tumors: Prospects and challenges,” AAPS J., vol. 20, no. 6, pp. 1–12, 2018.
[7] M. De La Hoz Polo, A. Torramilans Lluís, O. Pozuelo Segura, A. Anguera Bosque, C. Esmerado Appiani, and J. M. Caminal Mitjana, “Ocular ultrasonography focused on the posterior eye segment: What radiologists should know,” Insights Imag., vol. 7, no. 3, pp. 351–364, 2016.
[8] M. Bláivas, D. Theodorou, and P. R. Sierzenski, “A study of bedside ocular ultrasonography in the emergency department,” Academic Emerg. Med., vol. 9, no. 8, pp. 791–799, 2002.
[9] R. Yoonessi, A. Hussain, and T. B. Jang, “Bedside ocular ultrasound for the diagnostic evaluation of acute near vision impairment over 30 years: An analysis for the global burden of disease study,” Lancet. Glob. Health, vol. 9, no. 2, pp. e130–e143, 2021.
[10] J. Xu, H. Le, M. Huang, S. Athar, and D. Samaras, “Variational feature disentanglement for fine-grained image categorization,” in Proc. IEEE/CVF Conf. Vis. Pat. Recognit., 2021, pp. 6626–6636.
[11] S.-X. Zhao, Y. Chen, K.-F. Yang, Y. Luo, B.-Y. Ma, and Y.-J. Li, “A local and global feature disentangled network: Towards classification of benign-malignant thyroid nodules from ultrasound images,” IEEE Trans. Med. Imag., vol. 41, no. 6, pp. 1497–1509, Jun. 2022.
[12] S. Feng, B. Liu, Y. Zhang, X. Zhang, and Y. Li, “Two-stream compare and contrast network for vertebral compression fracture diagnosis,” IEEE Trans. Med. Imag., vol. 40, no. 9, pp. 2496–2506, Sep. 2021.
[13] Y.-J. Huang et al., “Rectifying supporting regions with mixed and active supervision for rib fracture recognition,” IEEE Trans. Med. Imag., vol. 39, no. 12, pp. 3843–3854, Dec. 2020.
[14] R. Gupta, V. Gupta, B. Kumar, P. K. Singh, and A. K. Singh, “A novel method for automatic retinal detachment detection and estimation using ocular ultrasound image,” Multimedia Tools Appl., vol. 79, no. 15, pp. 11143–11161, 2020.
[15] Y. Wang, C. Tang, J. Wang, Y. Sang, and J. Lv, “Cataract detection based on ocular B-ultrasound images by collaborative monitoring deep learning,” Knowl.-Based Syst., vol. 231, 2021, Art. no. 107442.
[16] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, “Zoom-in-net: Deep mining lesions for diabetic retinopathy detection,” in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv., Springer, 2017, pp. 267–275.
[17] N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, W. Dai, and E. Xing, “Reinforced auto-zoom net: Towards accurate and fast breast cancer segmentation in whole-slide images,” in Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support, Springer, 2018, pp. 317–325.
[18] J. Zhang, Q. He, Y. Xiao, H. Zheng, C. Wang, and J. Luo, “Ultrasound image reconstruction from plane wave radio-frequency data by self-supervised deep neural network,” Med. Image Anal., vol. 70, 2021, Art. no. 102018.
[19] H.-Y. Zhou, C. Lu, S. Yang, X. Han, and Y. Yu, “Preservation learning improves self-supervised medical image models by reconstructing diverse contexts,” in Proc. IEEE/IV Conf. Int. Conf. Vis. Pat. Recognit., 2021, pp. 3399–3409.
[20] P. Khosla et al., “Supervised contrastive learning,” Adv. Neural Inf. Process. Syst., vol. 33, pp. 18661–18673, 2020.
[21] X. Hu, D. Zeng, X. Xu, and Y. Shi, “Semi-supervised contrastive learning for label-efficient medical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv., Springer, 2021, pp. 481–490.
[22] A. Chartsias et al., “Contrastive learning for view classification of echocardiograms,” in Proc. Int. Workshop Adv. Simplifying Med. Ultrasound, Springer, 2021, pp. 149–158.
[23] E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie, “When does contrastive visual representation learning work?,” in Proc. IEEE/CVF Conf. Vis. Pat. Recognit., 2022, pp. 14755–14764.
[24] M. Pezeshki, O. Kaba, Y. Bengio, A. C. Courville, D. Precup, and G. Lajoie, “Gradient starvation: A learning procrivity in neural networks,” Adv. Neural Inf. Process. Syst., vol. 34, pp. 1256–1272, 2021.
[25] J. Xu, H. Le, M. Huang, S. Athar, and D. Samaras, “Variational feature disentangling for fine-grained few-shot classification,” in Proc. IEEE/IV Conf. Int. Conf. Vis. Pat. Recognit., 2021, pp. 8812–8821.
[26] C. Pei, F. Wu, L. Huang, and X. Zhuang, “Disentangle domain features for cross-modality cardiac image segmentation,” Med. Image Anal., vol. 71, 2022, Art. no. 102078.
[27] J. Cheng et al., “Multimodal disentangled variational autoencoder with game theoretic interpretability for glioma grading,” IEEE J. Biomed. Health Inform., vol. 26, no. 2, pp. 673–684, Feb. 2022.
[28] D. Bau et al., “Seeing what a GAN cannot generate,” in Proc. IEEE/IV Conf. Int. Conf. Vis. Pat. Recognit., 2019, pp. 4502–4511.
[29] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conf. Comput. Vis. Pat. Recognit., 2015, pp. 815–823.
[42] Y. Wang et al., “Deep attentional features for prostate segmentation in ultrasound,” in Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv., Springer, 2018, pp. 523–530.

[43] Z. Han, Z. Fu, and J. Yang, “Learning the redundancy-free features for generalized zero-shot object recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 12865–12874.

[44] S. Zhang, R. Du, D. Chang, Z. Ma, and J. Guo, “Knowledge transfer based fine-grained visual classification,” in Proc. IEEE Int. Conf. Multimedia Expo, 2021, pp. 1–6.

[45] X. Peng, K. Wang, Z. Zhu, and Y. You, “Crafting better contrastive views for siamese representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 16031–16040.

[46] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 603–612.

[47] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, “Cross attention network for few-shot classification,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 4003–4014.

[48] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 12310–12320.

[49] C.-F. R. Chen, Q. Fan, and R. Panda, “CrossViT: Cross-attention multi-scale vision transformer for image classification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 357–366.

[50] D. S. Kermany et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” Cell, vol. 172, no. 5, pp. 1122–1131, 2018.

[51] T. Rahman et al., “Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images,” Comput. Biol. Med., vol. 132, 2021, Art. no. 104319.

[52] M. E. Chowdhury et al., “Can ai help in screening viral and COVID-19 pneumonia?,” IEEE Access, vol. 8, pp. 132665–132676, 2020.

[53] D. Lv et al., “A cascade-seme network for COVID-19 detection in chest X-ray images,” Med. Phys., vol. 48, no. 5, pp. 2337–2353, 2021.

[54] F. Rundo, A. Genovese, R. Leotta, F. Scotti, V. Piuri, and S. Battisti, “Advanced 3D deep non-local embedded system for self-augmented X-ray-based COVID-19 assessment,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 423–432.

[55] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and flexible image augmentations,” Information, vol. 11, no. 2, 2020, Art. no. 125. [Online]. Available: https://www.mdpi.com/2078-2499/11/2/125

[56] X. Zhu et al., “PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization,” Computerized Med. Imag. Graph., vol. 102, 2022, Art. no. 102137.

[57] W. Ye, J. Yao, H. Xue, and Y. Li, “Weakly supervised lesion localization with probabilistic-CAM pooling,” 2020, arXiv:2005.14480.

[58] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size,” 2016, arXiv:1602.07360.

[59] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” 2020, arXiv:1802.03426.

[60] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.