Self-Supervised Generalized Zero Shot Learning For Medical Image Classification Using Novel Interpretable Saliency Maps

Dwarikanath Mahapatra, Zongyuan Ge, Mauricio Reyes

Abstract—In many real world medical image classification settings, access to samples of all disease classes is not feasible, affecting the robustness of a system expected to have high performance in analyzing novel test data. This is a case of generalized zero shot learning (GZSL) aiming to recognize seen and unseen classes. We propose a GZSL method that uses self-supervised learning (SSL) for: 1) selecting representative vectors of disease classes; and 2) synthesizing features of unseen classes. We also propose a novel approach to generate GradCAM saliency maps that highlight diseased regions with greater accuracy. We exploit information from the novel saliency maps to improve the clustering process by: 1) Enforcing the saliency maps of different classes to be different; and 2) Ensuring that clusters in the space of image and saliency features should yield class centroids having similar semantic information. This ensures the anchor vectors are representative of each class. Different from previous approaches, our proposed approach does not require class attribute vectors which are essential part of GZSL methods for natural images but are not available for medical images. Using a simple architecture the proposed method outperforms state of the art SSL based GZSL performance for natural images as well as multiple types of medical images. We also conduct many ablation studies to investigate the influence of different loss terms in our method.

Index Terms—Generalized zero shot learning, self-supervised learning, saliency, classification, X-ray, pathology

I. INTRODUCTION

In the present era, deep learning methods have achieved state of the art performance for many medical image classification tasks such as diabetic retinopathy grading [22], digital pathology image classification [36] and chest X-ray diagnosis [26], [62], to name a few. State of the art (SOTA) fully supervised methods have access to both the ‘seen’ and ‘unseen’ class labels, and trained models learn the characteristics of all classes. However many real-world scenarios do not provide access to samples of all possible diseases. As a result, unseen classes are generally classified into one of the seen classes, resulting in wrong diagnosis. For deployment in clinical settings we expect to encounter both seen and unseen classes during the test phase, where a reliable model should accurately predict both classes. This is a case of generalized zero shot learning (GZSL) and is challenging since predicting unseen classes as one of the seen classes can lead to incorrect diagnosis. In this work we propose a GZSL method for medical image classification using self supervised learning (SSL) and knowledge derived from saliency maps, and demonstrate its effectiveness across multiple medical image datasets.

GZSL is a widely explored topic for natural images [61], [67] where seen and unseen classes are characterized by class attribute vectors. A model learns to correlate between class attribute vectors and corresponding feature representations. This gives a strong reference point in synthesizing features of both seen and unseen classes, since by inputting the attribute vector of the desired class the corresponding feature representation can be generated. However medical images do not have such well defined class attributes since it requires high clinical expertise and time to define unambiguous attribute vectors for different disease classes. Hence it is not a trivial task to apply state of the art GZSL methods from natural image applications to medical image classification. For example, in the case of lung X-ray diagnosis many conditions co-occur frequently such as Atelectasis, Effusion, and Infiltration. An effective class attribute vector should be able to precisely identify the attribute categories and the corresponding entries, which is very challenging. Solving the GZSL problem for medical images without using attribute vectors is a challenging task but essential nevertheless due to the potentially immense benefits of reducing annotation effort of clinicians. It also helps to alleviate the critical issue of data shortage for many classes. The main contribution of our work is to perform GZSL for medical images without class attribute vectors.

Initial approaches to tackle ZSL [12] learnt cross-modal relationships between visual feature and semantic embeddings
(class attribute vectors) of natural images. Subsequently, recent generative approaches to GZSL [19], used generative adversarial networks (GANs) to optimize the divergence between the data distribution of seen classes and generated features. However, generators trained on seen class features cannot accurately represent unseen classes. The sub-optimal synthetic data does not lead to high performance of such models. As an attempt to circumvent this problem, some methods [46] utilize unlabeled data of unseen classes in a transductive way. However it leads to increased system complexity since two GANs are needed to model seen and unseen classes as they do not consider the relations between source and target domains.

Methods leveraging transductive approaches are particularly relevant for medical image classification tasks [63] since a large amount of medical images are acquired but not annotated due to high expertise required for such tasks. Consequently many image analysis tasks make use of semi-supervised learning to leverage the information from unlabeled classes. Absence of any supervised information from the unseen domain makes it very challenging to differentiate among disease labels, especially when many labels show similar appearance to the untrained eye.

Another tricky issue facing GZSL applications in general, and medical images in particular, is the different complexity of feature representations between images of seen and unseen classes. For natural image classification the difference in feature representations is addressed using attribute vectors. Since it is not feasible in practice to get manually defined attribute vectors of medical images, synthesizing unseen class features from seen classes is challenging. Leveraging unlabeled unseen class data (e.g., using anchors) can be effective in bridging the semantic gap [63]. In an attempt to address the above challenges our paper makes the following contributions:

1) Our main contribution is to achieve GZSL of medical images without using class attribute vectors, commonly required for natural image applications. This is important for real world clinical scenarios where defining class attribute vectors is a time consuming and expensive task.

2) We build on a contrastive learning baseline clustering method and propose novel additional SSL loss terms for: 1) deriving anchor vectors through clustering; and 2) feature synthesis of seen and unseen classes.

3) We propose a new approach to calculate interpretable GradCAM saliency maps that highlights disease regions with greater accuracy. This is particularly helpful when different diseases are localized in nearby locations.

4) We exploit attention focused information from the enhanced saliency maps to improve the clustering process by: 1) enforcing the saliency maps of different classes to be different; and 2) ensuring that clustering in the space of image and saliency features yield class centroids having similar semantic information.

In [38] we proposed a preliminary version of our method and, in comparison, our current work has the following novelties: 1) we propose a novel approach to calculate GradCAM saliency maps and use saliency information for clustering; 2) we perform extensive ablation and validation studies under different settings to determine the contribution of different components, and realism of synthetic features.

II. PRIOR WORK

A. (Generalized) Zero-Shot Learning

In ZSL the goal is to recognize classes not encountered during training. External information about the novel classes may be provided in form of semantic attributes [33], visual descriptions [1], or word embeddings [42]. ZSL has been addressed using GANs [19], Variational Autoencoders (VAE) [55] or both of them [67].

In GZSL the purpose is to recognize images from known and unknown domains. Many works [20], [61], [67] have shown promising results by training GANs in the known domain and generate unseen visual features from the semantic labels. This allows them to train a fully supervised classifier for two domains, which is robust to the biased recognition problem. The work by Huang et. al. [25] describes a Generative Dual Adversarial Network (GDAN) which couples a generator, a regressor and a discriminator. The interaction among the three components produces various visual features conditioned on class labels. Keshari et al. [30] use over-complete distributions to generate features of the unseen classes, while Min et. al. [43] use domain aware visual bias elimination for synthetic feature generation. Different from the above works we achieve GZSL without the need for descriptive class attribute vectors, but by generating anchor vectors that define specific classes and specifying the class label of the desired output feature.

B. Self-Supervised Learning

SSL methods consist of two main approaches; 1) pretext tasks and 2) contrastive learning based approaches. Common pretext tasks include estimating relative position of patches [17], local context [45], colour [68] and exemplar learning [18]. Down-stream tasks are used to evaluate the quality of features learned by self-supervised learning and are independent of pre-text tasks. Contrastive learning approaches such as MoCo [24] and SimCLR [14] are popular and give state-of-the-art results for down-stream task-based methods.

Recent works also use self-supervision for domain adaptation [54] and can be considered as the first work to combine GZSL and SSL [63]. SSL has found wide use in medical image analysis by using innovative pretext tasks such as patients’ MR scan recognition to detect vertebra [27], context restoration for classification, segmentation, and disease localization [13], image registration [60], and anomaly detection [6]. Other works include surgical video re-colourisation as a pretext task for surgical instrument segmentation [53], rotation prediction for lung lobe segmentation and nodule detection [58], and data augmentation [40]. For histopathology based domain specific pretext tasks SSL has been used for semi-supervised histology classification [37], active learning [41], stain normalization [39], and cancer sub-typing using visual dictionaries [44].

While our work is inspired from [63] in using SSL for GZSL, and using GANs for feature synthesis, there are significant differences such as: 1) we do not use class attribute vectors for training. Since definition of class attribute vectors

```
for medical images is infeasible we use a simpler yet effective architecture for GZSL. 2) [63] used a single generator but two discriminators to differentiate between seen and unseen classes. However we make use of a single generator and one discriminator to differentiate between all classes by leveraging anchor vectors; 3) we use a SSL based clustering approach to derive the anchor vectors of each class, including unseen classes. We use high level knowledge of the number of classes as a supervisory signal.

C. Few/Zero Shot Learning In Medical Images

Few-shot methods have been relatively less explored in medical image analysis applications. Chen et. al. in [15] propose a generative network based approach for one-shot MRI segmentation. Paul et. al. in [48] proposed an ensemble learning based approach for chest X-ray diagnosis. Other applications for FSL have been proposed for brain imaging modality recognition [50], volumetric medical image segmentation [21], and differential diagnosis of brain MRI [52]. Zero shot medical image analysis is a much less explored topic with limited applications such as registration [32] and artefact reduction [16]. Paul et al. [47] proposed a GZSL method for chest X-ray diagnosis by learning the relationship between multiple semantic spaces (from X-ray, CT images and reports). However not all datasets have multiple image modalities and text reports. Hayat et al. [23] learn an image’s visual and text representation guided by the input’s corresponding semantics extracted from a rich medical text corpus such as BioBert [34]. Our proposed method works with images from a single modality and shows state of the art performance on multiple public datasets.

III. METHOD

Method Overview: Figure 1 depicts our proposed workflow. In the first step we generate anchor vectors (cluster centroids) by modifying the SwAV clustering approach [11]. We have two clustering stages: one for seen class samples and second for unseen classes. Anchor vectors of the seen class samples are used to get SSL based loss terms for the second clustering stage, and we also use saliency map features to improve the clustering output. The second step involves feature generation, which takes a noise vector and desired class label of output vector to synthesize features. Synthesized and real features of unseen and seen classes are used to train a softmax classifier for identifying different disease classes.

A. SSL And Saliency Based Clustering

Let the number of classes in the seen set be \( n_S \), and the number of classes in the unseen set be \( n_U \). We assume that the total number of classes is known. We learn anchor vectors of different classes by using the SSL based online clustering approach SwAV (Swapping Assignments between multiple Views) [11], and introduce additional SSL and saliency inspired loss terms. Typical offline clustering methods [4], [9] alternate between cluster assignment and centroid update resulting in high training time. To overcome this and inspired by contrastive instance learning [64], [11] enforces that different augmentations of the same image are mapped to the same cluster. Multiple image views are contrasted by comparing their cluster assignments instead of features.

We use cluster centers as class anchor vectors since they provide a reliable representation of the corresponding class [35]. The anchor vectors are computed in an online fashion since the number of unseen classes may change in a dynamic way when the system adds new classes. We describe the baseline SwAV approach using notation from the original paper [11]. Given image features \( x_t \) and \( x_s \) from two different transformations of the same image, we compute their cluster assignments \( q_t \) and \( q_s \) by assessing the distance of the features to a set of \( K \) cluster centers \( c_1, \cdots, c_K \). A “swapped” prediction problem is solved with the following loss function:

\[
\mathcal{L}(x_t, x_s) = \ell(x_t, q_s) + \ell(x_s, q_t),
\]

where \( \ell(x, q) \) measures the fit between features \( x \) and assignment \( q \). Thus we compare features \( x_t \) and \( x_s \) using their intermediate cluster assignments \( q_t \) and \( q_s \). If the two \( x \)’s capture same information, we can predict the cluster assignment from the other feature.

Online clustering: Given image \( I_n \), it is transformed to \( I_{nt} \) using transformation \( t \) from a set \( T \) of image transformations. A non-linear mapping \( f_0 \) transforms \( I_{nt} \) to a feature vector which is projected to the unit sphere, i.e., \( x_{nt} = f_0(I_{nt})/\|f_0(I_{nt})\|_2 \). The cluster assignment \( q_{nt} \) is computed by determining the distance between \( x_{nt} \) and the set of cluster centroids, \( c_1, \cdots, c_K \). \( C \) denotes a matrix whose columns are \( c_1, \cdots, c_K \).

a) Swapped prediction problem: Each term in Eq.1 represents the cross entropy loss between \( q \) and the probability obtained by taking a softmax of the dot products of \( x_t \) and all columns in \( C \), i.e.,

\[
\ell(x_t, q_s) = - \sum_k q_s(k) \log p_t(k), \quad p_t(k) = \frac{\exp \frac{x_t^\top c_k}{\tau}}{\sum_k \exp \frac{x_t^\top c_k}{\tau}},
\]

where \( \tau = 0.1 \) is a temperature parameter [64]. Computing this loss over all images and augmentations results in the following loss function for swapped prediction:

\[
\mathcal{L}(x_t, x_s) = - \sum_{n=1}^{N} \sum_{s,t \sim T} \left[ \frac{x_{nt}^\top C q_{ns}}{\tau} + \frac{x_{ns}^\top C q_{nt}}{\tau} \right] - \log \sum_{k=1}^{K} \exp \left( \frac{x_{nt}^\top c_k}{\tau} \right) - \log \sum_{k=1}^{K} \exp \left( \frac{x_{ns}^\top c_k}{\tau} \right).
\]

This loss function is jointly minimized with respect to the centroids in \( C \) and parameters \( \theta \) of \( f_0 \).

Computing the cluster assignments: The clustering assignments \( q \) are computed in an online fashion using image features within a batch. Since the centroids in \( C \) are used across different batches, SwAV clusters multiple instances to their appropriate clusters. Given feature vectors \( X = [x_1, \cdots, x_B] \), we map them to centroids \( C = [c_1, \cdots, c_K] \) using \( Q = [q_1, \cdots, q_B] \), and we optimize \( Q \) to maximize the similarity between \( X \) and \( C \),

\[
\max_{Q \in \mathbb{R}^B} \text{Tr}(Q^\top C^\top X) + cH(Q),
\]
where $H$ is the entropy function, $H(Q) = -\sum_{ij} Q_{ij} \log Q_{ij}$ and $\epsilon = 0.05$ controls the mapping smoothness. A high $\epsilon$ could potentially result in a trivial solution where all samples collapse into an unique representation and are assigned uniformly to all prototypes.

1) **Our Novel Contribution:** We make two novel contributions in the clustering stage - the first involves the use of SSL based loss terms, and the second is the use of additional information from saliency maps to identify accurate cluster centers. To obtain cluster centers we use the concept of anchor vectors to bridge the gap between seen and unseen classes. The cluster centers also function as the anchor vectors and are obtained using the SwA V clustering approach [11]. We have two clustering stages: one for seen class samples and second for unseen classes. Feature generation leverages one Generator and one Discriminator along with anchor vectors to derive SSL loss terms.

$$L_{SSL1} = \min \left( CoSim(C_S^1, C_U^1), \sigma_1 \right)$$

(5)

Here $\sigma_1 = 0.15$ is a parameter that determines the semantic distance between the centroids, and CoSim denotes cosine similarity with values ranging from 0 (i.e. no similarity) and 1 (i.e. maximum similarity). The above equation imposes the constraint that the cosine similarity between centroids of seen and unseen classes should not exceed $\sigma_1$, in order to impose uniqueness.

3) We add a second self supervised constraint such that the similarity of seen class sample, $x^i_S$, with its corresponding class centroid $C_S^i$ is higher than its similarity with all the unseen class centroids $C_U^j$. This is achieved by randomly selecting samples from the seen class training set during minibatch training and computing the different cosine similarities. This is implemented by

$$L_{SSL2} = \max \left( CoSim(x_S^i, C_S^i) - CoSim(x_S^i, C_U^j), \sigma_2 \right) \forall j$$

(6)

$\sigma_2 = 0.25$ controls the minimum degree of semantic difference between different classes and $i,j$ index seen and unseen classes. This ensures that there is sufficient difference between seen and unseen class centroids in order not to have overlapping samples.

Note that while clustering the seen classes we do not add any label supervision. However in the second stage of clustering unseen classes we enforce that seen and unseen classes are different.

2) **Additional Constraints From Saliency Maps:** In medical images there are multiple structures of interest in a neighborhood. We propose to use an additional source of information in the form of interpretable saliency maps obtained using GradCAM [56], although other saliency methods can be used.

Interpretable saliency maps highlight regions identified as informative by the trained classifier. Consequently, for disease classification saliency maps highlight regions with disease activity. Saliency maps are an effective way to incorporate an attention mechanism. In our approach we seek to produce distinctive saliency maps for different class labels. Given image $I$ and a classification neural network $M$ the saliency
maps are obtained for the second to last layer as \( \{S_i,n\}_{n=1}^N \) for each class \( n \). We then calculate their corresponding latent representations \( \{z_{S_i,n}\}_{n=1}^N \) using an autoencoder trained to reconstruct the saliency maps. Deep latent representation has been effectively used as an image perception similarity metric [69], and for image retrieval [57]. In order to enhance differentiability of saliency maps for different classes, we calculate the following class distinctiveness loss term:

\[
L_{CD} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{CosSim}(z_{S_i}, z_{S_j}),
\]

Equation (7) therefore enforces distinctiveness of the different \( N \) saliency maps for each label class generated by the model \( M \).

An additional constraint from saliency information is also incorporated into the loss function. Saliency maps are expected to have semantically relevant information derived from the original images. Hence a clustering of the saliency features should map the clusters in such a way that they are semantically similar to the clusters in the image feature space. This is achieved by enforcing that the cosine similarity between image-based feature cluster centroids and saliency-based feature cluster centroids of the same class are as high as possible, and is implemented as the cosine loss of the two centroids of the corresponding classes.

\[
L_{Cent} = \sum_i \left( 1 - \text{CosSim}(C_i^{Image}, C_i^{Sal}) \right),
\]

where \( C_i^{Image} \) denotes the centroid of class \( i \) derived from image features, while \( C_i^{Sal} \) is the centroid of the same class from the saliency maps, and \( i \) denotes all classes. The final loss term for clustering the unseen class samples is

\[
L_{Unseen} = L(x_i, x_t) + \lambda_1 L_{SSL1} - \lambda_2 L_{SSL2} + \lambda_3 L_{CD} + \lambda_{Cent} L_{Cent}
\]

where \( L(x_i, x_t) \) is defined in Eqn. 3.

### B. Feature Generation Network

For feature generation we follow the steps in [66]. Given the training images of seen classes and unlabeled images of the unseen classes we learn a generator \( G : \mathcal{E}, \mathcal{X} \rightarrow \mathcal{X}^* \), which takes a class label vector \( e^y \in \mathcal{E} \) and a Gaussian noise vector \( z \in \mathcal{Z} \) as inputs, and generates a feature vector \( \hat{x} \in \mathcal{X}^* \). The discriminator \( D : \mathcal{X}, \mathcal{E} \rightarrow [0, 1] \) takes a real feature vector \( x \) or synthetic feature \( \hat{x} \) and corresponding class label vector \( e^y \) as input and determines whether the feature vector matches the class label vector. The generator \( G \) aims to fool \( D \) by producing features highly correlated with \( e^y \) using a Wasserstein adversarial loss [3]:

\[
L_{WGAN} = \min_G \max_D \mathbb{E}_x[D(x, e^y)] - \mathbb{E}_{\hat{x}}[D(\hat{x}, e^y)] - \lambda \mathbb{E}[\|\nabla_{\hat{x}} D(\hat{x}, e^y)\|_2 - 1]^2,
\]

where the third term is a gradient penalty term, and \( \hat{x} = \alpha x + (1 - \alpha) \hat{x} \). \( \alpha \sim U(0,1) \) is sampled from a uniform distribution.

#### 1) Our Contribution: Self Supervised Loss From Anchor Vectors

The discriminator \( D \) is a classifier that determines whether the generated feature vector \( \hat{x} \) belongs to one of the seen classes. Since the unseen classes are not labeled we do not have a data distribution for them and hence we use self-supervision to determine whether the generated feature vector matches an unseen class. As the anchor vectors (i.e., the cluster centers) are fixed, we calculate the cosine similarity between the generated vector \( \hat{x} \) and the anchor vector corresponding to the desired class \( y \), i.e.

\[
L_{SSL3} = 1 - \text{CosSim}(\hat{x}, c_y)
\]

If \( \hat{x} \) truly represents the desired class \( y \) then the cosine similarity between \( \hat{x} \) and the corresponding anchor vector \( c_y \) should be highest amongst all \( K (= n_s + n_u) \) anchor vectors, and the corresponding loss is lowest.

#### 2) Classifier Loss: We expect that \( \hat{x}^s \) (synthesized feature vector for seen classes) are predicted correctly by a pre-trained classifier \( CL \) with a loss defined as below

\[
L_{CL} = -\mathbb{E}_{(\hat{x}, y^s, \theta^{CL})} \left[ \log P(y^s|\hat{x}^s, \theta^{CL}) \right]
\]

where \( \theta^{CL} \) is the classification probability and \( \theta^{CL} \) denotes fixed parameters of the pre-trained classifier.

### C. Training, Inference and Implementation

The final loss function for feature generation is defined as

\[
L = L_{WGAN} + \lambda_{CL} L_{CL} + \lambda_3 L_{SSL3}
\]

where \( \lambda_{CL}, \lambda_3 \) are weights that balance the contribution of the different terms. Once training is complete we specify the label of desired class and input a noise vector to \( G \) which synthesizes a new feature vector. We combine the synthesized target features of the unseen class \( \hat{x}^u \) and real and synthetic features of seen class \( x^s, \hat{x} \) to construct the training set. Then we train a softmax classifier by minimizing the negative log likelihood loss:

\[
\min \theta \frac{1}{|\mathcal{Y}|} \sum_{(x,y) \in \mathcal{Y}} \log P(y|x, \theta)
\]

where \( P(y|x, \theta) = \frac{\exp(\theta^T y)}{\sum_y \exp(\theta^T y)} \) is the classification probability and \( \theta \) denotes classifier parameters. The final class prediction is by \( f(x) = \arg \max_y P(y|x, \theta) \).

**Inference:** Given a set of initial seen and unseen class samples, the clustering stages yields class centroids. The subsequent feature synthesis module generates samples of different classes which are used to train a softmax classifier. Given a test feature we use the softmax classifier to obtain its class label.

**Implementation Details:** We compare the results of our method for medical images with existing GZSL methods. For methods developed for natural images we replace the class label vector \( e^y \) with the corresponding class attribute vectors. For feature extraction, similar to [65], we use a pre-trained ResNet-101 to extract 2048 dimensional CNN features for natural images. The feature extractors for individual medical image datasets are described separately (ref Section IV-A).
The generator (G) and discriminator (D) are all multilayer perceptrons. G has two hidden layers of 2000 and 1000 units respectively while the discriminator D is implemented with one hidden layer of 1000 hidden units. We chose Adam [31] as our optimizer, and the momentum was set to (0.9, 0.999). The values of loss term weights are $\lambda_{C_{L}} = 0.6$, $\lambda_{3} = 0.9$. Training the Swav Clustering algorithm takes 12 hours and the feature synthesis network for 50 epochs takes 17 hours, all on a single NVIDIA V100 GPU (32 GB RAM). PyTorch was used for all implementations.

### D. Evaluation Protocol

The seen class $S$ can have samples from 2 or more disease classes, and the unseen class $U$ contains samples from the remaining classes. We use all possible combinations of labels in $S$ and $U$. Following standard practice for GZSL, average class accuracies are calculated for two settings: 1) S: training is performed on synthesized samples of $S + U$ classes and test on the seen test set $S_{Te}$. 2) U: training is performed on synthesized samples of $S + U$ classes and test on unseen test set $U_{Te}$. We also report the harmonic mean defined as,

$$H = \frac{2 \times Acc_{S} \times Acc_{U}}{Acc_{S} + Acc_{U}}$$  \hspace{1cm} (15)$$

where $Acc_{S}$ and $Acc_{U}$ denote the accuracy of images from seen (setting $S$) and unseen (setting $U$) classes respectively.

### IV. Experimental Results

#### A. Dataset Description

We demonstrate our method’s effectiveness on natural images and the following medical imaging datasets for classification tasks.

1) **CAMELYON17** dataset [7]: contains 1000 whole slide images (WSIs) with 5 slides per patient: 500 slides for training and 500 slides for testing. Training set has annotations of 3 categories of lymph node metastasis: Macro (Metastases greater than 2.0 mm), Micro (metastasis greater than 0.2 mm or more than 200 cells, but smaller than 2.0 mm), and ITC (single tumor cells or a cluster of tumor cells smaller than 0.2mm or less than 200 cells). We extract $224 \times 224$ patches from the different slides and obtain 130,000 tumor patches and 200,000 normal patches. We take a pre-trained ResNet-101 and fine tune the last FC layer using the CAMELYON16 dataset [5], which is closely related but different from CAMELYON17. A baseline fully supervised learning (FSL) method is implemented\(^1\) which is the top ranked in the leaderboard. We assume different combinations of 2 seen classes and 1 unseen classes, and the reported results are an average of 10 runs across different combinations. Hyperparameter values are $\lambda_{1} = 1.3$, $\lambda_{2} = 0.8$, $\lambda_{CD} = 1.0$, $\lambda_{C_{ent}} = 0.8$, $\lambda_{CL} = 0.7$, $\lambda_{3} = 1.0$.

2) **NIH Chest X-ray** Dataset: For lung disease classification we adopted the NIH Chest X-ray14 dataset [62] having 112,120 expert-annotated frontal-view X-rays from 30,805 unique patients and has 14 disease labels. Original images were resized to $224 \times 224$. A pre-trained ResNet-101 was finetuned using the CheXpert dataset [26] and the chosen baseline FSL was from [51]. We assume different combinations of 7 seen classes and 7 unseen classes, and the reported results are an average of 25 runs across different combinations. Hyperparameter values are $\lambda_{1} = 1.1$, $\lambda_{2} = 0.7$, $\lambda_{CD} = 1.2$, $\lambda_{C_{ent}} = 0.9$, $\lambda_{CL} = 0.6$, $\lambda_{3} = 0.9$.

3) **CheXpert** Dataset: We used the CheXpert dataset [26] consisting of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 common chest conditions. Original images were resized to $224 \times 224$. A pre-trained ResNet-101 was finetuned using the NIH dataset [62] and the baseline FSL method was of [49], which is ranked second for the dataset with shared code. We assume different combinations 7 seen classes and 7 unseen classes, and the reported results are an average of 25 runs across different combinations. Hyperparameter values are $\lambda_{1} = 1.2$, $\lambda_{2} = 0.8$, $\lambda_{CD} = 1.1$, $\lambda_{C_{ent}} = 1.3$, $\lambda_{CL} = 0.7$, $\lambda_{3} = 1.1$.

4) **Kaggle Diabetic Retinopathy** dataset: has approximately 35,000 images in the provided training set [28]. Images are labeled by a single clinician with the respective DR grade, out of 4 severity levels: 1-mild (2443 images), 2-moderate (5291 images), 3-severe (873 images), and 4-proliferative DR (708 images). The normal class 0 has 25810 images. A pre-trained ResNet-101 was finetuned using [59] which has 9939 color fundus images ($2720 \times 2720$) from 2740 diabetic patients. Although the number of classes are different from Kaggle the features are accurate since the end task is DR detection. The chosen baseline method was of [2]. Original images were resized to $224 \times 224$. We assume different combinations 2 seen classes and 2 unseen classes, and the reported results are an average of 15 runs across different combinations. Hyperparameter values are $\lambda_{1} = 1.4$, $\lambda_{2} = 0.9$, $\lambda_{CD} = 1.1$, $\lambda_{C_{ent}} = 0.8$, $\lambda_{CL} = 0.8$, $\lambda_{3} = 1.1$.

5) **Gleason grading challenge** dataset \(^2\) for prostate cancer (PCA) [29]. It has 333 Tissue Microarrays (TMAs) from 231 patients and has 5 Gleason grades. Six pathologists with 27,15,1,24,17, and 5 years of experience annotated the data and majority voting was used to construct the ”ground truth label”. The training set had 200 TMAs while the validation set had 44 TMAs. A separate test set consisting of 87 TMAs from 60 other patients. The baseline FSL was the classification outcome of the top ranked method\(^3\). The feature extractor was a pre-trained ResNet-101 finetuned using the CAMELYON16 dataset [5]. Since both are histopathology image datasets, the feature extractor is quite accurate. The high dimensional images were divided into $224 \times 224$ patches. The individual labels patches from normal images were all “normal”.

---

\(^1\)https://grand-challenge-public.s3.amazonaws.com/evaluation-supplementary/80/46fc579c-51f0-40c4-bd1a-7c28e033f133/Camelyon17.pdf

\(^2\)https://gleason2019.grand-challenge.org/Home

\(^3\)https://github.com/hubutui/Gleason
Fig. 2: Feature visualizations for NIH Chest X-ray Dataset: (a) All classes from actual dataset; distribution of synthetic samples generated by (b) SC-GZSL; (c) SC-GZSL \( w_\text{SSL3} \); (d) SDGN [63]. Different colours represent different classes. (b) is similar to (a) in terms of the clusters being separate with minimal overlap. (c) and (d) are quite different due to overlapping and compact clusters.

For the diseased images (all Gleason grades except 1), the labels of individual patches were obtained using the multiple instance learning method of [8]. Thus we obtained more than 5,000 patches of each label. Although a much larger dataset for PCA using WSIs is available\(^6\), the data cannot be used for external submissions\(^5\). We assume different combinations of 3 seen classes and 3 unseen classes, and the reported results are an average of 20 runs across different combinations. Hyperparameter values are \( \lambda_1 = 1.2, \lambda_2 = 0.8, \lambda_{CD} = 1.4, \lambda_{Cent} = 1.2, \lambda_{CL} = 0.9, \lambda_3 = 1.1 \).

Since we did not have labels of the organizer designated test sets of all datasets, a 70/10/20 split at patient level was done to get training, validation and test sets for NIH Chest X-ray, CheXpert and Kaggle DR datasets.

B. Baseline Methods

We compare our method’s performance with the following GZSL methods employing different feature generation approaches such as CVAE or GANs:

1) GDAN - CVAE based generation method of [25].
2) OCD - The over complete distribution (OCD) method of [30]
3) SDGN- Self-supervised learning GZSL method of [63].
4) SUP- Top performing fully supervised methods of corresponding datasets. For FSL baselines we implement the different methods referred in the description of individual datasets. For each case we do a 70/10/20 split for training/validation/test sets at the patient level to ensure images from the same patient are in one fold only. Note that the FSL baselines are different from the feature extractors, which are used for subsequent feature synthesis.

\(^6\)https://www.kaggle.com/c/prostate-cancer-grade-assessment
\(^5\)https://www.kaggle.com/c/prostate-cancer-grade-assessment/discussion/201117

Following common practices for GZSL we report accuracy for seen and unseen classes. Our method is denoted as SC-GZSL (Self-supervised Clustering based GZSL). The GZSL methods dealing with natural images use class attribute vectors, and when applying them for medical images we replace the attribute vectors with class centroids. SCGZSL\(_{GC}\) denotes the use of modified GradCAM saliency maps.

C. Visualization of Synthetic Image Features

Figure 2 (a) shows t-SNE plot of features of actual data from the NIH chest X-ray dataset where the different classes are spread over a wide area, with slight overlap between some classes. Figure 2 (b) shows the distribution of synthetic features generated by our method. Although the corresponding clusters for the different classes have separate locations in the two figures they are similar to that of Figure 2 (a) in the sense that the different classes are similarly separated and there is minimal overlap. Figure 2 (c) shows the feature distribution for our method without using self-supervision in the feature generation stage. The resulting distribution is compact without overlap between classes, which is not representative of the real-world case. Classifiers trained on such distributions perform poorly on unseen classes. Figure 2 (d) shows the feature distributions using SDGN [63] for feature synthesis. Although it also uses SSL the resulting feature representation is less accurate than our proposed method which contributes to the corresponding lower performance.

D. Generalized Zero Shot Learning Results

Classification results for medical images shown in Table I and the corresponding 95% confidence intervals for harmonic mean, \( H \), in Table II show our proposed method significantly outperforms all competing GZSL methods including SDGN. Note that we use the anchor vectors in place of attribute vectors for these feature synthesis methods. This significant difference in performance can be explained by the fact that the complex architectures that worked for natural images will...
not be equally effective for medical images which have less information. Absence of attribute vectors for medical images is another contributing factor. The class attributes provide a rich source of information about natural images which can be leveraged using existing architectures. Since those are not available for medical images these methods do not perform equally well.

To determine confidence values we calculate average Harmonic Mean across 5 runs and use the following standard formula

\[ z = \sqrt{(H \times (1 - H))/n}, \]

where \( H \) is the Harmonic mean, \( n \) is the sample size (of the test set), and \( z \) is the number of standard deviations from the Gaussian distribution. Commonly used number of standard deviations from the Gaussian distribution and their corresponding significance level (\( z \)) are: 1.64 (90%), 1.96 (95%), 2.33 (98%), 2.58 (99%). We also show results when using different clustering methods such as DeepCluster and k-means instead of SwAV, while using our feature generation method. The results are inferior to our proposed method thus demonstrating the fact that SwAV gives better representations of the cluster centroids.

### E. Ablation Studies For Medical Images

Table III shows results for the following ablation studies:

1. SCGZSL\(_{w\text{SSL1}}\) - SCGZSL without the loss term \( \mathcal{L}_{\text{SSL1}} \) (Eqn.5) for obtaining anchor vectors.
2. SCGZSL\(_{w\text{SSL1},\text{SSL2}}\) - SCGZSL without the loss term \( \mathcal{L}_{\text{SSL2}} \) (Eqn.6) to get anchor vectors.
3. SCGZSL\(_{w\text{CD}}\) - SCGZSL without the loss term \( \mathcal{L}_{\text{CD}} \) (Eqn.7) to get anchor vectors.
4. SCGZSL\(_{w\text{Cent}}\) - SCGZSL without the loss term \( \mathcal{L}_{\text{Cent}} \) (Eqn.8) to get anchor vectors.
5. SCGZSL\(_{w\text{SSL1},\text{SSL2}}\) - Using only the baseline loss term \( \mathcal{L}(z_1, z_2) \) (Eqn.3) for clustering all seen and unseen classes together, and no \( \mathcal{L}_{\text{SSL3}} \) for feature synthesis.
6. SCGZSL\(_{w\text{SSL3}}\) - SCGZSL without the loss term \( \mathcal{L}_{\text{SSL3}} \) (Eqn.11) for training the feature synthesis network.
7. SCGZSL\(_{w\text{SSL1}}\) - SCGZSL using only \( \mathcal{L}_{\text{SSL3}} \) for feature synthesis.

The baseline method, SCGZSL\(_{w\text{SSL1},\text{SSL2}}\) (last row of Table III), does not use any form of self supervision and has lowest \( H \) values. These values are very low compared to \( \text{SUP} \) in Table I. As we add more information through the different loss terms we observe a progressive improvement in performance.

The first five ablation studies investigate the effect of our modified clustering approach on the final classification results. Their significant performance degradation compared to SCGZSL indicates the importance of our novel SSL (\( \mathcal{L}_{\text{SSL1}}, \mathcal{L}_{\text{SSL2}} \)) and saliency based terms (\( \mathcal{L}_{\text{CD}}, \mathcal{L}_{\text{Cent}} \)) in obtaining accurate anchor vectors. Recall that our novel loss terms in the clustering stage can be categorized as SSL loss terms (\( \mathcal{L}_{\text{SSL1}}, \mathcal{L}_{\text{SSL2}} \)) and saliency based attention loss terms (\( \mathcal{L}_{\text{CD}}, \mathcal{L}_{\text{Cent}} \)). The numbers in Table III indicate that saliency and SSL terms have a similar contribution to the final model performance. Their combined contribution makes SCGZSL perform much better than compared methods although individually they perform slightly worse than baseline models.

Compared to SCGZSL, we observe that excluding \( \mathcal{L}_{\text{SSL3}} \) (SCGZSL\(_{w\text{SSL1},\text{SSL2}}\)) leads to significant reduction of \( H \) (more than 3.5%) across all datasets. This indicates that \( \mathcal{L}_{\text{SSL3}} \) makes important contributions to our method’s performance. This is reasonable since the feature synthesis is the most important step in GZSL, although it is supported by accurate representation of anchor vectors. The use of anchor vectors makes it easier to synthesize features of unseen classes.

The influence of \( \mathcal{L}_{\text{SSL1}}, \mathcal{L}_{\text{SSL2}} \) is quantitatively similar as shown by similar \( H \) values of SCGZSL\(_{w\text{SSL1},\text{SSL2}}\). SCGZSL\(_{w\text{SSL1}}\) across all datasets. However their difference in \( H \) values compared to SCGZSL is nearly 2.4% which is significant (\( p = 0.01 \)). Thus the use of self supervision is an important factor in obtaining accurate anchor vectors (cluster centroids). Although the baseline clustering mechanism, SwAV, uses self supervision in the form of contrastive loss, including \( \mathcal{L}_{\text{SSL1}} \) and \( \mathcal{L}_{\text{SSL2}} \) significantly improves clustering accuracy. Excluding both \( \mathcal{L}_{\text{SSL1}}, \mathcal{L}_{\text{SSL2}} \) and using the baseline SwAV (‘only \( w\mathcal{L}_{\text{SSL3}} \)’) gives significantly reduced \( H \) values for the different datasets despite using \( \mathcal{L}_{\text{SSL3}} \) for feature synthesis. This clearly indicates the importance of having accurate anchor vectors for our method. SCGZSL\(_{w\text{SSL1},\text{SSL2}}\) can be considered as the most basic method without using any of our proposed novel loss terms, and unsurprisingly gives the worst results.

### F. Hyperparameter Selection

To find the optimal set of parameters we tried different approaches such as exhaustive grid search, and random search. While they converge to the optimal values, we make use of information from prior work and adopt the following strategy. For all the competing synthesis methods we start with the original values provided by the authors and vary them in range \( x \pm 0.5x \) in steps of \( x/10 \), where \( x \) is the initial value. The best results are usually obtained using author provided values for each method. Note that this approach is efficient due to initial parameter values provided in the original works. To tune the values from scratch we recommend using a grid search approach. Figure 3 (a) shows the harmonic mean values for the NIH Chest X-ray dataset for different values of hyperparameters \( \lambda_{\text{CD}}, \lambda_1, \lambda_2 \), Figure 3 (b) shows plots for \( \lambda_{\text{CD}}, \lambda_{\text{Cent}}, \lambda_3 \), while Figure 3 (c) shows the corresponding plots for different values of \( \sigma_1, \sigma_2 \). The \( \lambda \)'s were varied between \([0.4 – 1.5]\) in steps of 0.05 and the performance on a separate test set of 10,000 images was monitored.

We start with the base cost function of Eqn. 10, and first select the optimum value of \( \lambda_1 \). \( \lambda_1 \) value is fixed and we then determine optimal \( \lambda_2 \), and subsequently \( \lambda_{\text{CD}} \) and \( \lambda_{\text{Cent}} \) by fixing the previous parameters. Then the optimal \( \lambda_{\text{CL}} \) is determined with fixed previous parameters and finally optimal \( \lambda_3 \) is determined. The order in which the parameters were set is important and we find the above order as giving the best results. Similarly the value of \( \sigma \)'s were varied between \([0.1, 0.5]\) in steps of 0.05, and the resulting classification accuracy of the X-ray images was determined, i.e., whether they were assigned to the correct cluster (class).

We show in Figure 4 plots for hyperparameters when they are optimized in a different order. The final AUC values are
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3163232, IEEE Transactions on Medical Imaging

MAHAPATRA et al.: SELF-SUPERVISED GENERALIZED ZERO SHOT LEARNING

<table>
<thead>
<tr>
<th>Method</th>
<th>CAMELYON17</th>
<th>NIH X-ray</th>
<th>CheXpert</th>
<th>Kaggle DR</th>
<th>Gleason</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>U</td>
<td>H</td>
<td>S</td>
<td>U</td>
</tr>
<tr>
<td>I-VAEGAN [67]</td>
<td>90.2</td>
<td>82.9</td>
<td>80.0</td>
<td>81.4</td>
<td>88.5</td>
</tr>
<tr>
<td>GDAN [25]</td>
<td>91.1</td>
<td>83.8</td>
<td>80.9</td>
<td>82.3</td>
<td>89.2</td>
</tr>
<tr>
<td>OCD [30]</td>
<td>91.5</td>
<td>84.7</td>
<td>81.3</td>
<td>83.0</td>
<td>89.9</td>
</tr>
<tr>
<td>SDGN [63]</td>
<td>92.1</td>
<td>84.4</td>
<td>81.1</td>
<td>82.7</td>
<td>90.2</td>
</tr>
<tr>
<td>SCGZSL $^{\text{GC}}$</td>
<td>93.7</td>
<td>87.2</td>
<td>85.1</td>
<td>86.1</td>
<td>91.8</td>
</tr>
<tr>
<td>SUP</td>
<td>93.7</td>
<td>87.4</td>
<td>86.9</td>
<td>87.1</td>
<td>92.1</td>
</tr>
<tr>
<td>Deep-Cluster</td>
<td>91.6</td>
<td>83.9</td>
<td>80.7</td>
<td>82.2</td>
<td>90.7</td>
</tr>
<tr>
<td>K-Means</td>
<td>90.6</td>
<td>83.4</td>
<td>80.7</td>
<td>82.0</td>
<td>88.9</td>
</tr>
</tbody>
</table>

![Table IV](image)

**TABLE IV.** Although the two sets of results in Tables IV, I are not directly comparable because of the different ground truth labels in single and multilabel settings, it is worth noting that multilabel classification tasks usually perform better than single label classification as the joint learning of multiple disease characteristics improves overall performance. However, the results are inferior compared to the original approach thus suggesting that the clustering based approach may not be optimal for multi-label classification and requires further investigation. Note that our approach is not directly comparable with the multi-label GZSL approach of [23] which uses semantic embeddings generated by BioBert, and requires a different approach to handle the multi-label setting.

**H. Enhanced Interpretability Saliency Maps**

Figure 5 (a) shows the expert delineated regions for pleural effusion (red outline) and atelectasis (blue outline). Saliency maps are shown for GradCAM (Fig. 5 (b)), our proposed modified method (Fig. 5 (c)), without $\mathcal{L}_{\text{Centr}}$ (Fig. 5 (d)), and without $\mathcal{L}_{\text{CD}}$ (Fig. 5 (e)). The GradCAM saliency maps in this particular example are very similar for the two disease classes and widely dispersed without much distinction between the two classes. On the other hand the maps generated by our method are much closer to the manual annotations and shows distinct regions for the two classes. Excluding $\mathcal{L}_{\text{Centr}}$ leads to inaccurate localization of the diseased region while excluding $\mathcal{L}_{\text{CD}}$ results in more dispersed salient regions, alluding to potential shortcut learning.

**I. Effect of Number of Synthetic Samples:**

Figure 6 shows, for the NIH chest X-ray and CAMELYON17 dataset, the effect of adding synthetic samples on $\text{Acc}_{\text{CS}}, \text{Acc}_{\text{CU}}$ as a function of dataset augmentation factor. Increasing synthesized examples increases $\text{Acc}_{\text{CU}}$ at a high rate while reducing $\text{Acc}_{\text{CS}}$, although at a lower rate. The overall H value keep increasing. Adding synthetic samples improves discriminative power of classifiers and reduces bias towards seen classes.

**J. Relationship Between Saliency and Image Features**

Recall that in Eqn. 8 we enforce the constraint that clustering of the saliency features should map the clusters in such a way that they are semantically similar to the clusters in the image feature space. The cosine similarity between image-based feature cluster centroids and saliency-based feature
cluster centroids of the same class are made as high as possible.

Figure 7 shows clustering results under different conditions for the Gleason 2019 challenge dataset that has 5 classes. Figure 7 (a) shows clustering output when using the labels of all classes while Figure 7 (b) shows the clustering output using our proposed modified SwAV approach with 3 seen and 2 unseen classes, which is very similar to the output of Figure 7 (a). Figures 7 (c)-(f) show clustering outputs when excluding, respectively, $L_{SSL1}$, $L_{SSL2}$, $L_{Cent}$, $L_{CD}$ (Equations 5,6,8,7). We observe that the clustering outputs degrade significantly when excluding the different terms. This establishes the important contributions of each of the clustering loss terms.

K. Realism of Synthetic Features

In order to evaluate the realism of synthetic features we let trained ophthalmologists analyze images corresponding to the features. First we train an auto encoder to reconstruct a given input fundus image. Thereafter, given a feature vector the decoder part of the autoencoder can reconstruct the original image. Using these reconstructed images we let two trained ophthalmologists examine them and rate whether they consider them realistic or not.

Two trained ophthalmologists having 12 and 14 years experience in examining retinal fundus images for abnormalities assessed realism of generated images. We present them with a common set of 500 synthetic images obtained from the features generated by our method and ask them to classify each as realistic or not. The evaluation sessions were conducted separately with each ophthalmologist blinded to other’s answers.

### Table III: Ablation Results For Medical Images

<table>
<thead>
<tr>
<th>Method</th>
<th>CAMELIYON17</th>
<th>NIH X-ray</th>
<th>CheXpert</th>
<th>Kaggle DR</th>
<th>Gleason</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>U</td>
<td>H</td>
<td>S</td>
<td>U</td>
</tr>
<tr>
<td>SCGZSL</td>
<td>93.5</td>
<td>91.7</td>
<td>92.6</td>
<td>87.2</td>
<td>85.1</td>
</tr>
<tr>
<td>$w_{SSL1}$</td>
<td>90.2</td>
<td>88.1</td>
<td>89.1</td>
<td>83.8</td>
<td>81.9</td>
</tr>
<tr>
<td>$w_{SSL2}$</td>
<td>90.1</td>
<td>87.3</td>
<td>88.7</td>
<td>83.4</td>
<td>82.0</td>
</tr>
<tr>
<td>$w_{CD}$</td>
<td>91.2</td>
<td>88.7</td>
<td>89.9</td>
<td>84.5</td>
<td>82.1</td>
</tr>
<tr>
<td>$w_{Cent}$</td>
<td>90.8</td>
<td>88.1</td>
<td>89.4</td>
<td>84.0</td>
<td>82.2</td>
</tr>
<tr>
<td>only $L_{SSL3}$</td>
<td>89.3</td>
<td>86.4</td>
<td>87.8</td>
<td>82.6</td>
<td>80.7</td>
</tr>
<tr>
<td>$\phi$</td>
<td>87.2</td>
<td>84.1</td>
<td>85.6</td>
<td>80.7</td>
<td>79.1</td>
</tr>
</tbody>
</table>

### Table IV: GZSL Results for multi-label scenario

Results are shown for two chest X-ray datasets, and the numbers indicate slight performance degradation over results reported in Table I.

### Table V: Agreement statistics for different image generation methods amongst 2 ophthalmologists. Numbers in bold indicate agreement percentage while numbers within brackets indicate actual numbers out of 500 samples.

Results for SCGZSL show one ophthalmologist (OPT 1) identified 451/500 (90.2%) images as realistic while OPT 2 identified 452 (90.4%) generated images as realistic. Both of them had a high agreement with 440 common images (88.0% “Both Experts” in Table V) identified as realistic. Considering both OPT 1 and PAT 2 feedback, a total of 463 (92.6%) unique images were identified as realistic (“Atleast 1 Expert” in Table V). Subsequently, 37/500 (7.4%) of the images were not identified as realistic by any of the experts (“No Expert” in Table V). Agreement statistics for other methods are summarized in Table V. The highest agreement between two ophthalmologists is obtained for images generated by our method. For all the other methods their difference from SCZSL is significant.

### Table VI: Additional Results For Diabetic Retinopathy

Table VI shows detailed results for diabetic retinopathy when the unseen class is a combination of different severity grades. The results indicate that it is easier to classify the least severe and most severe samples. The other labels are more challenging because of their being close on the severity scale which results in shared characteristics. This requires further investigation since 1) a reliable method should not be biased towards specific classes, and 2) the intermediate severity grades occur more frequently than the high severity cases and necessitate accurate detection. In future work we aim to explore methods to overcome this limitation.
**M. Visualization of Synthetic Images**

In this section we visualize some of the synthetic images obtained using our method. Note that we do not generate synthetic images but the corresponding features. A variational auto encoder is trained on the training images to reconstruct the original images. After generating the synthetic features we use the decoder of the VAE to reconstruct the image.

Figure 8 (a) shows examples of reconstructed X-ray images and the corresponding disease region (blue contour) identified by an expert. Figures 8 (b,c) shows, respectively, saliency maps obtained by GradCam and our proposed method on the reconstructed image. Results obtained by our method are closer to the ground truth than GradCAM thus highlighting its effectiveness.

**V. DISCUSSION**

**Importance of Self Supervised Learning:** The baseline clustering method, i.e. SwAV, is a self supervised approach using contrastive loss. However it needs to be modified since we need to separate out the seen and unseen classes. The cluster centroids act as representative vectors of the particular class and hence need to be accurate. Hence we include two additional self supervised loss terms in the clustering stage that enforce consistency amongst the seen class centroids and distinctiveness compared to unseen class centroids. As experimental results prove, the additional constraints improve clustering performance, which in turn improves GZSL classification.

**Not Relying on Class Attribute Vectors:** As mentioned previously, natural images have class attribute vectors of seen and unseen classes which acts as the auxiliary information for generating features of unseen classes. In the absence of such information for medical images we use the cluster centroids as anchors and representatives of each class. Our method’s performance on natural image datasets is impressive as it outperforms the state of the art methods on all datasets. Through our results we demonstrate that GZSL for medical images is possible without having to rely on class attribute vectors that are obtained after painstaking effort and uses valuable time of clinicians.
Fig. 5: Comparison with Radiologist’s Saliency Maps. (a) Original image with expert-annotated outlines of diagnosed conditions. Saliency Maps for different methods: (b) Original GradCAM; (c) Our proposed method; (d) Our method without $L_{\text{Cent}}$; (e) Our method without $L_{\text{CD}}$. Top row: Pleural effusion (red contour); Bottom row: Atelectasis (blue contour).

Fig. 6: Value of accuracy and H when adding synthetic samples to the dataset: (a) NIH dataset; (b) CAMELYON17 dataset.

A common scenario for disease cases is the evolution of disease characteristics (e.g., discovery of newer variants) which requires updating of symptoms and characteristics. Thus we need a system that can adapt with the evolving disease characteristics. Our proposed approach is better suited in this scenario as it does not require painstaking definition of attribute vectors.

Importance of Saliency: We use saliency maps as a way to integrate attention input into the clustering step. Saliency maps improve clustering performance in addition to the self-supervised losses. This is explained by the fact that saliency maps highlight a focused region relevant to the task in hand and hence provide better quality information than image features alone.

Realism of Synthetic Images: We engaged experienced clinicians to examine generated images (obtained from the synthetic features), and they determined that a high percentage of generated images are realistic (refer Supplementary material). This shows that our method generates realistic features and does not suffer from unconstrained feature generation wherein the features come from arbitrary distributions. Thus our approach of including self-supervised information in the feature generation step actually improves the quality of generated features.

Limitations: While our method performs well on multiple medical imaging datasets it also has some limitations. 1) Our approach cannot be adapted to an infinite number of unseen classes since we assume that the number of seen and unseen classes are known. This assumption is valid since it is expected that clinicians have a good idea of what they expect to encounter. 2) Relying solely on image features can affect robustness when the features are not easily separated. In such a scenario it is helpful to have an additional information (e.g. semantic embeddings). Recent work [23] have used BioBert [34] to generate semantic embeddings (similar to class attribute vectors) from images and this approach needs to be further investigated for generalizability. 3) As demonstrated by the results with retinal images (Section IV-L) our method shows a slight performance degradation for the intermediate severity grades which is a matter of concern that requires further exploration to learn better image features and have nearly equal performance across all disease labels.

Multi-Label Classification: While we adapt our method to multi-label classification we observe performance degradation.
This could be attributed to the fact that the clustering does not explicitly account for multi-label scenarios and hence leads to sub-par performance. However explicitly modeling the multi-label influence through cross-label interactions could further improve the performance.

**Workflow Design:** Since we do not have access to class attribute vectors of the medical images we employ multiple stages of clustering and self supervised learning to train the feature synthesis module. Wu et al. [63] also use self supervised learning for feature synthesis. However our workflow is simpler because we use a single generator while they used 2 generators for seen and unseen classes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3163232, IEEE Transactions on Medical Imaging

VI. CONCLUSION

We propose a GZSL approach for medical images without relying on class attribute vectors. Our novel method can accurately synthesize feature vectors of unseen classes by employing self supervised learning at different stages such as anchor vector selection, and training a feature generator. We also propose a novel approach to generate enhanced GradCAM saliency maps and integrate attention focused information from them for clustering. Using self supervision helps us learn feature of unseen classes from the seen classes. The distribution of synthetic features generated by our method are close to the actual distribution, while removing the self-supervised and saliency based terms results in unrealistic distributions. Experimental results show our method outperforms other GZSL approaches in literature, and is consistently better across multiple public datasets.

Our approach is useful in scenarios where the number of disease classes are known but labeled samples of all classes cannot be accessed due to infrequent occurrence of such cases or lack of expert clinicians to annotate complex cases. A specific example is Gleason grading where the number of Gleason grades is well known. While fully supervised settings still provide the best performance they are dependent upon sufficient labeled samples. Hence GZSL can be relevant to address the low data scenarios.

In future work we aim to explore the use of BioBert in generating semantic embeddings for different disease classes and its robustness for different medical imaging datasets for the purpose of GZSL. We expect that the generated semantic embeddings will reduce our method’s complexity. We also aim to make our method much more suitable for multi-label classification settings.

REFERENCES


