Multi-Label Generalized Zero Shot Chest Xray Classification By Combining Image-Text Information With Feature Disentanglement

Dwarikanath Mahapatra, Antonio Jimeno Yepes, Behzad Bozorgtabar, Sudipta Roy, Zongyuan Ge, Mauricio Reyes

***Abstract*— In fully supervised learning-based medical image classification, the robustness of a trained model is influenced by its exposure to the range of candidate disease classes. Generalized Zero Shot Learning (GZSL) aims to correctly predict seen and novel unseen classes. Current GZSL approaches have focused mostly on the single-label case. However, it is common for chest X-rays to be labelled with multiple disease classes. We propose a novel multi-modal multi-label GZSL approach that leverages feature disentanglement and multi-modal information to synthesize features of unseen classes. Disease labels are processed through a pre-trained BioBert model to obtain text embeddings that are used to create a dictionary encoding similarity among different labels. We then use disentangled features and graph aggregation to learn a second dictionary of inter-label similarities. A subsequent clustering step helps to identify representative vectors for each class. The multi-modal multi-label dictionaries and the class representative vectors are used to guide the feature synthesis step, which is the most important component of our pipeline, for generating realistic multi-label disease samples of seen and unseen classes. Our method is benchmarked against multiple competing methods and we outperform all of them based on experiments conducted on the publicly available NIH and CheXpert chest X-ray datasets.**

***Index Terms*— Multi-label, GZSL, Text Embeddings, Chest x-rays, Feature synthesis, Disentanglement**

I. Introduction

Fully supervised deep learning methods provide state-of-the-art (SOTA) performance for a variety of medical image analysis tasks, such as diabetic retinopathy grading [17] and chest X-ray diagnosis [21]. A key element to the success of fully supervised methods is having access to all classes during the training process. However, in a radiological workflow disease types not seen previously can be encountered, e.g., new strains of COVID-19 or tumour types in histopathological data. Hence, in conventional fully supervised approaches, new disease subtypes will be misclassified into one of the previously seen classes. Apart from these unfavorable misclassifications, the lack of adaptability of deep learning classification systems to new classes can also result in lengthy system re-certification loops for clinically deployed AI systems.

In contrast, self-supervised learning offers a paradigm that does not rely entirely on labeled data, potentially enhancing the model's ability to generalize to unseen classes. By learning rich feature representations from unlabeled data, self-supervised learning methods [4], [5] can complement supervised approaches by providing preliminary insights into novel classes without explicit prior knowledge [51]. However, while self-supervised learning can mitigate some of the challenges posed by unlabeled and unseen data, it typically requires subsequent fine-tuning with labeled data to achieve optimal performance, which may not be feasible in scenarios where new class data remains scarce or unavailable.

Zero-Shot Learning (ZSL) aims to learn plausible representations of unseen classes from the available features of seen classes. In a more generalized setting, we expect to encounter both seen and unseen classes during the test phase. This is the case of Generalized Zero-Shot Learning (GZSL), which is more challenging. Previous works on GZSL in medical images have mostly focused on the single label scenario where an image is assigned a single disease class [34], [36], [42]. However, chest X-ray (CXR) datasets have multiple labels assigned to the images and single-label methods do not work well in this setting. [18] proposed a multi-label GZSL method to predict multiple seen and unseen diseases in CXR images. The approach consists of mapping both visual and semantic modalities to a latent feature and learn a visual representation guided by the input's corresponding semantics extracted from a medical text corpus. However, their approach yields sub-optimal results on the external NIH chest xray dataset [53] in terms of AUROC values of seen (0.79) and unseen (0.66) classes. This is possibly due to the sub-optimal use of text and imaging data. We propose a multi-label GZSL approach that uses multi-modal dictionaries encoding text and imaging information to encode the semantic relationship between

D. Mahapatra is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. (email: dwarikanath.mahapatra@inceptioniai.org)

A. Jimeno Yepes is with RMIT University, Melbourne, Australia and Unstructured Technologies, USA

B. Bozorgtabar is with the Signal Processing Laboratory (LT55), at the École Polytechnique Fédérale de Lausanne (EPFL), and the Radiology Department, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland (e-mail: behzad.bozorgtabar@epfl.ch)

S. Roy is with Jio Institute, Navi Mumbai, India.

Z. Ge is with Monash University, Melbourne, Australia, and Airdoc, Melbourne

M. Reyes is with the ARTORG Center for Biomedical Engineering Research, University of Bern and the Department of Radiation Oncology, University Hospital Bern, University of Bern, Switzerland

multiple disease labels. This enables us to learn a highly accurate feature representation which plays an important role in synthetic feature generation.

In contrast with medical imaging datasets for GZSL, datasets for GZSL in natural images [15], [47] have the advantage of providing attribute vectors for all classes to enable a model to correlate between attribute vectors and corresponding feature representations of the seen classes. Defining unambiguous attribute vectors for medical images requires deep clinical expertise and extensive invested time to annotate radiological images. This complexity is further exacerbated for the multi-label scenario, where many disease conditions have similar appearances and textures. In our previous work [36] we proposed a method to perform GZSL without using attribute vectors. In this approach, we used a baseline clustering method, called SwAV [8], and incorporated additional constraints based on self-supervised learning to perform single-label GZSL. However, the primary challenge of multi-label GZSL is to generate features incorporating characteristics of multiple labels. This is a challenging task since it requires an appropriate disentanglement between class-specific and class-agnostic features. To address this non-trivial feature generation challenge, we build upon [36] and introduce the following contributions for multi-label GZSL:

1) We propose a novel feature disentanglement method where a given image is decomposed into class-specific and class-agnostic features. This step is necessary since for multi-label problems an accurate combination of class-specific features is needed.
2) We apply graph aggregation on class-specific features to learn an image feature based multi-label dictionary based on interactions between different labels at a global scale. This leads to more discriminative feature learning and contributes to better multi-label feature synthesis.
3) We learn the semantic relationships between text embeddings of different disease classes and use this knowledge to guide the generation of realistic feature vectors that preserve the semantic relationship among multiple disease labels.

II. Prior Work

A. Feature Disentanglement

[32] provide a comprehensive overview of feature disentanglement techniques in medical image analysis. Feature disentanglement has been used for various tasks like segmentation [52], classification [37] and superresolution [35]. [9] propose Spatial Decomposition Network (SDNet) to decompose 2D medical images into spatial anatomical factors and non-spatial modality factors. They use it for different cross modal segmentation tasks. [40] propose a disentanglement approach using margin loss, conditional convolution and a fusion function, with applications to three multi-modal neuroimaging datasets for brain tumor segmentation. [48] propose Attention-enhanced Disentangled Representation (ADR) learning model for unsupervised domain adaptation in cardiac segmentation.

B. (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 the form of semantic attributes [27], visual descriptions [1], or word embeddings [38]. ZSL has been addressed using GANs [13], Variational Autoencoders (VAE) [45] or both of them [58]. In GZSL, the purpose is to recognize images from known and unknown domains. Prior work on natural images show promising results by training GANs in the known domain and generating unseen visual features from semantic labels [14], [58].

The work by [19] describes a Generative Dual Adversarial Network (GDAN) that couples a generator, a regressor, and a discriminator. In [23], the authors used over-complete distributions to generate features of the unseen classes, while [39] used domain-aware visual bias elimination for synthetic feature generation. [15] proposes a non-generative model for synthesizing edge-pseudo and center-pseudo samples to introduce greater diversity. The work by [25] proposes an Intra-Class Compactness Enhancement method (ICCE) for GZSL, which promotes intra-class compactness with inter-class separability on both seen and unseen classes in the embedding space and visual feature space. [47] leverage visual and semantic modalities to distinguish seen and unseen categories by deploying two variational autoencoders to generate latent representations for visual and semantic modalities in a shared latent space.

C. Multi-Label Zero-Shot Learning

The work of [28] proposes a novel deep learning architecture for multi-label zero-shot learning (ML-ZSL), which is able to predict multiple unseen class labels for each input instance using an information propagation mechanism from the semantic label space. As an extension to ZSL, ML-ZSL further requires one to assign multiple unseen labels. The work of [61] considers the separability of relevant and irrelevant labels, proposing a model that learns principal directions for images in the embedding space. Differently, the work of [16] leverages co-occurrence statistics of seen and unseen labels and learns a graphical model that jointly models the label matrix and the co-occurrence matrix.

D. GZSL In Medical Images

GZSL in medical image analysis is a much less explored topic with limited applications primarily because conventional methods from the natural image domain cannot be directly applied due to lack of class attribute vectors for medical images. Some initial works explored registration [26] and artifact reduction [10]. In earlier work [**?**], [34], [36] we proposed a class-attribute-free method for GZSL on different medical images by using saliency maps and self-supervised learning. In [42], the authors 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. The primary challenge of multi-label GZSL lies in synthesizing features that capture the characteristics of multiple classes. A

robust method requires appropriate disentanglement between class-specific and class-agnostic features. The class specific features can then be appropriately combined to synthesize feature vectors representing multiple-labels.

Recently, language models pre-trained on large corpora have been considered for ZSL [6]. More specifically in the biomedical domain, [18] learn an image's visual representation guided by the input's corresponding semantics extracted from BioBERT [29], a BERT [12]-based language model. [49] propose a method that relies on image similarity and embeddings with self-supervised learning. Different from other works, our method, combining image and text, works with images from a single modality and shows state-of-the-art performance on multiple public CXR datasets. In the following, we provide an overview of the proposed method as well as a detailed description of its components, followed by experimental setup and results obtained comparing seven previously proposed GZSL on two publicly available chest X-ray datasets.

III. Method

A. Method Overview:

Figure 1 depicts the proposed workflow. Let us denote a given image as x_i and the corresponding latent representation is denoted as z_i. The corresponding encoder for class L is denoted as E_l and the decoder is denoted as G_l. Our method consists of the following stages: 1) Image feature disentanglement to get class-specific component, $z_i^{spec_l}$ for class l, and a class-agnostic component, $z_i^{agn_l}$ from the original latent vector z_i using $\mathscr{L}_{Disent}$ (Eqn. 1); 2) Creating two multi-label dictionaries, $Dict_{Spe}$ (from class specific image features) and $Dict_{Text}$ using text embeddings of disease labels. The class-specific features are used to learn more global relationship between label features, whereas the text embeddings for different labels are obtained from BioBert [29]; 3) Clustering of seen and unseen class samples using $\mathscr{L}_{ML-Seen}$ in Eqn 7 and $\mathscr{L}_{ML-All}$ in Eqn.8 to obtain class centroids that function as class representative vectors. $\mathscr{L}_{ML-Seen}$ is the difference between Dic_{Spe} and centroid of seen classes and defined in Eq. 7. $\mathscr{L}_{ML-All}$ is the difference between Dic_{Text} and centroids of all classes and is defined in Eq. 8. Both are part of the clustering step.; 4) Feature synthesis to generate multi-label features of different label combinations using Eqn.13. The centroid vectors are used as reference vectors for feature synthesis. The synthesized vectors are compared with the centroids using $\mathscr{L}_{ML-Syn}$ defined in Eq. 12 to determine whether they belong to the desired classes; 5) Training a classifier to identify the correct set of labels for each test image (Eqn.14). Synthesized and real features of unseen and seen classes are used to train a multi-label classifier. Different from [36] we propose novel loss functions introduced in the clustering stage (Eqns. 7,8). The feature synthesis stage (Step 4) is similar to [36], but we use a completely different loss function (Eqn. 12) for multi-label considerations.

B. Feature Disentanglement

Feature disentanglement for domain adaptation separates the features into domain-specific and domain-invariant components [41]. The task of domain adaptation becomes one of minimizing the distance among domain-invariant features. In the case of GZSL, the data is from the same domain with different labels. Hence we propose to decompose the feature space of the seen class samples into 'class-specific' and 'class-agnostic' features. The class-specific features of each class will encode information specific to the particular class, and the class-specific features of different classes will be dissimilar. On the other hand, the class-agnostic features (e.g., characterization of bone in X-ray scans) of each class will be highly similar to each other. In this setup, we aim to yield class-specific and class-agnostic features to be mutually complementary and hence have minimal overlap in semantic content. This feature disentanglement helps to obtain features specific to each class which in turn allows for more accurate synthesis of multi-label features, by combining incorporating characteristics of the desired classes.

Figure 2 shows the architecture of our feature disentanglement network (FDN). The FDN consists of L encoder-decoder architectures corresponding to the L classes in the training data. We train different autoencoders for each class in order to obtain class specific features. The encoders and decoders (generators) are denoted, respectively, as $E_l(\cdot)$ and $G_l(\cdot)$. Similar to a classic autoencoder, the encoder, E_l ($l \in (1, \cdots, L)$), produces a latent code z_i for image $x_i \sim p$. Each decoder, G_l, reconstructs the original image from z_i. Furthermore, to divide the latent code, z_i, into two components we have two heads for the final embedding output (insead of one) corresponding to: a class-specific component, $z_i^{spec_l}$ for class l, and a class-agnostic component, $z_i^{agn_l}$. Both components are vectors, and they are combined and fed to the decoder, which reconstructs the original input. The disentanglement network is trained using the following loss function:

$$\mathscr{L}_{Disent} = \mathscr{L}_{Rec} + \lambda_1 \mathscr{L}_{spec} + \lambda_2 \mathscr{L}_{agn} + \lambda_3 \mathscr{L}_{agn-spec} \quad (1)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weights for above loss terms. **Reconstruction Loss**: $\mathscr{L}_{Rec}$, is the commonly used image reconstruction loss and is defined as:

$$\mathscr{L}_{Rec} = \sum_{l=1}^{L} \mathbb{E}_{x_i \sim p_l} \left[\left\| x_i^l - G_l(E_l(x_i^l)) \right\| \right] \quad (2)$$

The above term is a sum of the reconstruction losses from the class specific autoencoders.

Class Specific Loss: For given class l the class specific component $z_i^{spec_l}$ will have high similarity, according to some metric (e.g. cosine similarity), with samples from the same class. Since this feature is class specific it will have low similarity with the $z_i^{spec_k}$ of other classes k ($k \neq l$). These two conditions are incorporated using the following terms

$$\mathscr{L}_{spec} = \sum_{i,j} \sum_{l} \left(1 - \langle z_i^{spec_l}, z_j^{spec_l} \rangle\right) + \sum_{k} \langle z_i^{spec_l}, z_j^{spec_k} \rangle \quad (3)$$

where $\langle . \rangle$ denotes cosine similarity. The first term encourages high similarity for class specific features of samples having the same training labels. The second term encourages different classes, l and k to have highly dissimilar class specific features. The sum is calculated for all classes indexed by $\sum_l$ and over all samples indexed by i, j.

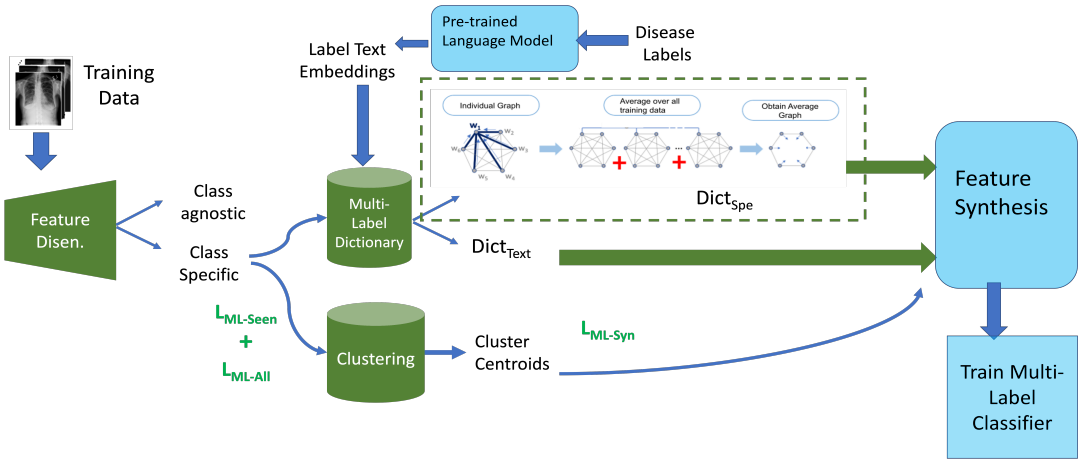

Fig. 1: Workflow of the proposed method. Training data goes through a feature disentanglement stage, followed by multi-modal and multi-label dictionary learning and clustering, feature synthesis and training of a multi-label classifier. Our novel contributions and loss functions are highlighted as green blocks and letters. Dic_{Spe} is the dictionary created from class specific features of seen classes and Dic_{Text} is the dictionary obtained from label texts. $\mathscr{L}_{ML-Seen}$ is the difference between Dic_{Spe} and centroid of seen classes and defined in Eq. 7. $\mathscr{L}_{ML-All}$ is the difference between Dic_{Text} and centroids of all classes and is defined in Eq. 8. Both are part of the clustering step. $\mathscr{L}_{ML-Syn}$ is defined in Eq. 12 and is part of the feature generation step.

Class Agnostic Loss: The class agnostic features of different classes have, by definition, similar semantic content and hence they will have high cosine similarity. L_{agn} is defined as,

$$\mathscr{L}_{agn} = \sum_{i,j} \sum_{l} \sum_{k} \left(1 - \langle z_i^{agn_l}, z_j^{agn_k} \rangle\right). \tag{4}$$

The above formulation ensures that the loss is minimized.

Finally, we want the class specific and class agnostic features of same-class samples to be mutually complementary and have minimal overlap in semantic content. This implies that their cosine similarity values should be minimal. Hence the final loss term is defined as

$$\mathscr{L}_{agn-spec} = \sum_{l} \langle z_i^{agn_l}, z_j^{spec_l} \rangle \tag{5}$$

Since the above loss terms are minimized it helps us achieve our stated objectives.

Figure 3 (a) shows the t-sne plots of image features (taken from the fully connected layer of a DenseNet-121 trained for image classification) while Figure 3 (b) shows the plot using the class-specific features. The plots of the original features shows different image class clusters that overlap and that makes it challenging to have good classification. On the other hand, the clusters obtained using the class-specific features are well separated and there is less overlap between different clusters. Figure 3 (c) shows the output of using class agnostic features where a significant overlap is observed among classes. This clearly demonstrates the efficacy of our feature disentanglement method, i.e., the class-specific and class-agnostic features fulfil their desired objectives. In the example in Figure 3, the features are taken from images belonging to 5 classes (Atelectasis,Consolidation, Effusion, Infiltration and Nodule) from the NIH dataset.

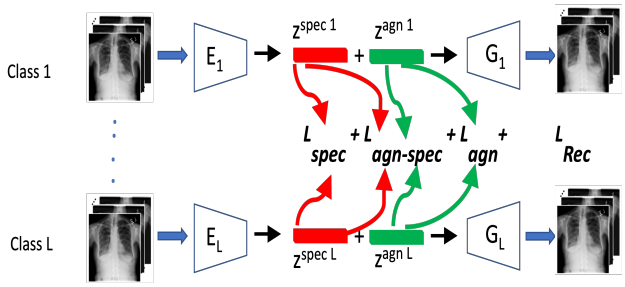

Fig. 2: Architecture of class specific feature disentanglement network. Given training images from different classes of the same domain, we disentangle features into class-specific and class-agnostic using autoencoders. The different feature components are used to define the different loss terms.

C. Embeddings

We generate embeddings of image class labels using BioBERT [29], a BERT [12]-like pre-trained model. BioBERT is pre-trained on biomedical literature, more specifically the model available from Huggingface[1], which is a base and cased model. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [29] is a pre-trained language representation model for the biomedical domain. It's input is the label (disease name of the sample) and the output is the corresponding label embedding vector. BioBERT is initialized with weights from BERT (Bidirectional Encoder Representations from Transformers) , which was pre-trained on general domain corpora (English Wikipedia and BooksCorpus). Then, BioBERT is pre-trained on biomedical domain corpora (PubMed abstracts and PMC full-text articles). BioBERT is fine-tuned and evaluated on three popular

[1] https://huggingface.co/dmis-lab/biobert-v1.1

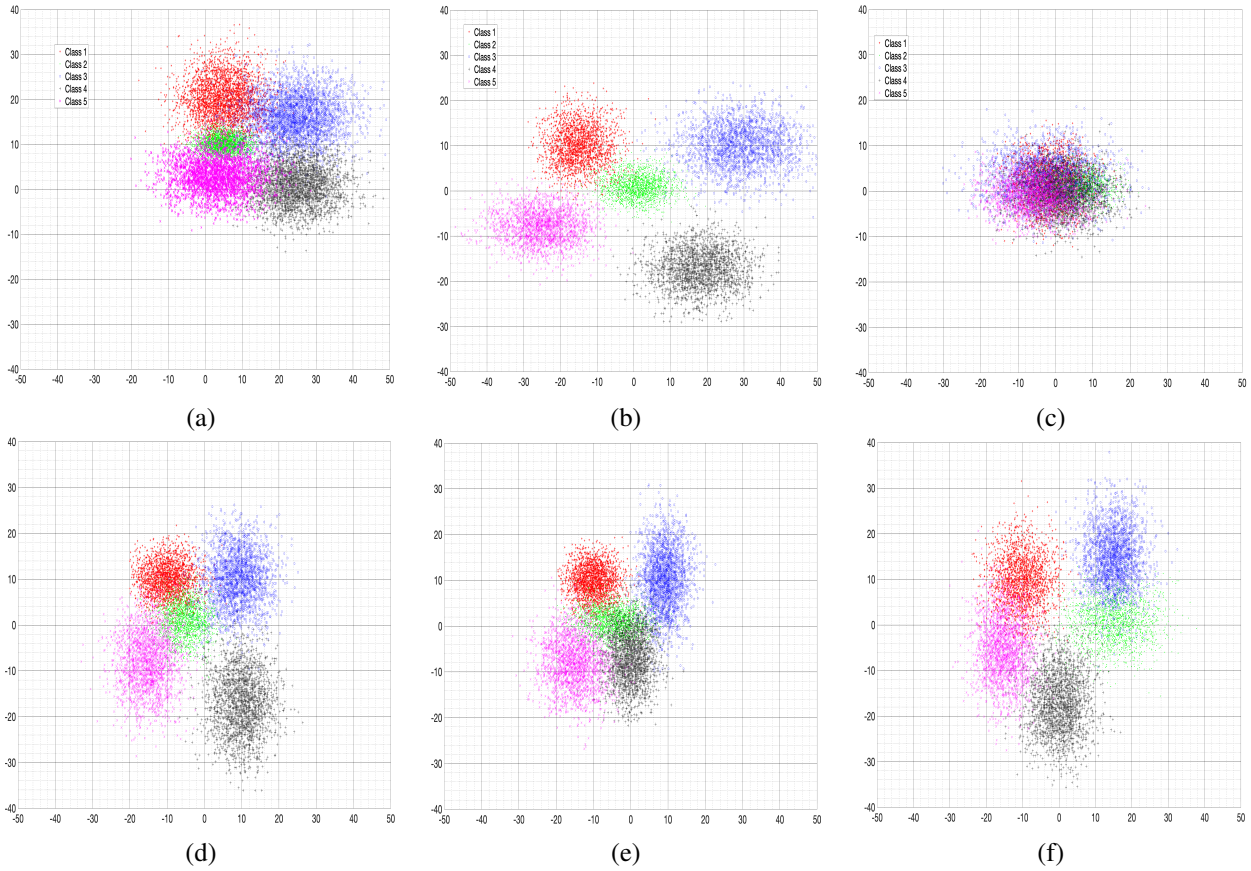

Fig. 3: T-sne results comparison between original image features and feature disentanglement output. (a) Original image features; (b) Class specific features; (c) Class agnostic features. Visualizations of synthetic features for: (d) ML-GZSL$_{w/o\ \mathscr{L}_{spec}}$; (e) ML-GZSL$_{w/o\ \mathscr{L}_{ML-Seen}}$; (f) ML-GZSL$_{w/o\ \mathscr{L}_{ML-All}}$

biomedical text mining tasks: named entity recognition (NER), relation extraction (RE) and question answering (QA). Various pre-training strategies with different combinations and sizes of general domain corpora and biomedical corpora are tested, and the results of the effect of each corpus on pre-training is analyzed. The pooled embeddings have 768 dimensions. [18] used a multi-layer perceptron and further adapted the embeddings to a lower dimension in a supervised fashion. In our work, we do not fine-tune the language model and propose reducing the dimensionality by projecting the embeddings to a lower dimensional space using t-sne (t distributed stochastic neighbor embedding) [50]. As per the implementation[2], to guarantee reproducibility, we set the random seed to a specific value. Table I shows the actual cosine similarity values between all the 15 classes - 14 diseased classes and 'No Finding'. The 15×15 matrix in Table I has all diagonal elements equal to 1 as it is the cosine similarity of a class's embedding with itself. During the feature generation stage, we enforce the constraint that the unseen class feature vectors should have cosine similarity values (with respect to other seen and unseen classes) close to the values shown in Table I. This matrix, which we refer as $Dict_{Text}$ - dictionary for text embeddings, is a realistic substitute for class attribute vectors. Notably, $Dict_{Text}$ avoids the labor-intensive process of defining class attribute vectors and provides a quantitative relationship between different disease labels.

Since our approach to obtain the BioBERT feature embeddings is based on t-sne, reproducibility is an important factor. We use different fixed values (total of 10 values) of the seed parameter and obtain the corresponding set of feature values. Thereafter we calculate the cosine similarity of each label pair similar to what is shown in Table I ((seed value$=$ 1367). We calculate the element-wise difference of each such table to the values shown in Table I. The diagonal elements are not considered since they are always equal to 1. The mean element wise difference is 0.012 (max$=$ 0.02, min=0.0004) which indicates a minor difference in terms of semantic similarity for each seed value. This shows that despite using different seeds the relative semantic similarity between feature embedding vectors does not change significantly.

D. Learning a Multi-Modal Multi-Label Dictionary

A multi-label dictionary is useful in quantifying the relationship between different labels and is used to guide the feature synthesis module. The dictionary is constructed from two sources (modalities): 1) class-specific features of seen class samples from images; 2) from text embeddings of the label vectors for all classes. We have already described in Section III-C the steps to generate $Dict_{Text}$ - the dictionary

[2]https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

	Atl.	Cardio.	Consol.	Edema	Eff.	Emphy.	Fibr.	Hernia	Infil.	Mass	None	Nodule	Pl.Th.	Pneu.	Pneumoth.
Atelectasis	1.00	0.84	0.93	0.92	0.66	0.99	0.77	0.99	0.93	0.93	0.49	0.70	0.79	0.99	0.89
Cardiomegaly	0.84	1.00	0.97	0.97	0.93	0.88	0.98	0.83	0.95	0.97	0.81	0.96	0.98	0.87	0.60
Consolidation	0.93	0.97	1.00	0.99	0.84	0.95	0.93	0.92	0.99	0.99	0.69	0.88	0.93	0.94	0.72
Edema	0.92	0.97	0.99	1.00	0.86	0.95	0.93	0.91	0.99	0.99	0.70	0.89	0.94	0.94	0.71
Effusion	0.66	0.93	0.84	0.86	1.00	0.71	0.96	0.65	0.84	0.85	0.91	0.98	0.95	0.70	0.40
Emphysema	0.99	0.88	0.95	0.95	0.71	1.00	0.82	0.99	0.95	0.95	0.54	0.75	0.83	0.99	0.86
Fibrosis	0.77	0.98	0.93	0.93	0.96	0.82	1.00	0.76	0.91	0.93	0.87	0.98	0.99	0.80	0.52
Hernia	0.99	0.83	0.92	0.91	0.65	0.99	0.76	1.00	0.92	0.91	0.48	0.70	0.78	0.99	0.91
Infiltration	0.93	0.95	0.99	0.99	0.84	0.95	0.91	0.92	1.00	0.99	0.68	0.87	0.92	0.95	0.73
Mass	0.93	0.97	0.99	0.99	0.85	0.95	0.93	0.91	0.99	1.00	0.70	0.88	0.94	0.95	0.72
No Finding	0.49	0.81	0.69	0.70	0.91	0.54	0.87	0.48	0.68	0.70	1.00	0.91	0.85	0.53	0.23
Nodule	0.70	0.96	0.88	0.89	0.98	0.75	0.98	0.70	0.87	0.88	0.91	1.00	0.97	0.74	0.45
Pleural_Thickening	0.79	0.98	0.93	0.94	0.95	0.83	0.99	0.78	0.92	0.94	0.85	0.97	1.00	0.82	0.54
Pneumonia	0.99	0.87	0.94	0.94	0.70	0.99	0.80	0.99	0.95	0.95	0.53	0.74	0.82	1.00	0.87
Pneumothorax	0.89	0.60	0.72	0.71	0.40	0.86	0.52	0.91	0.73	0.72	0.23	0.45	0.54	0.87	1.00

TABLE I: Table showing the Cosine similarity values of the labels' BioBERT embeddings (seed value=1367). This information is used to guide the clustering and feature generation stages.

from text embedding vectors of all the disease classes. In this section, we further describe our approach of using the class-specific features and graph aggregation to learn a dictionary for seen class samples.

Note that we only learn an image feature based dictionary of the seen classes. Since we do not know the actual image samples of unseen classes, it is difficult to identify features corresponding to them, and hence we cannot construct the corresponding dictionary. We construct a graph from the **seen** class samples in the following manner:

1) We represent each image sample as a separate graph.
2) Within a graph each of the seen class labels (representing a disease or condition) is a node, which is represented by the class-specific features.
3) Edge weights in the graph represent the similarity between corresponding nodes using cosine similarity of class specific features.

Assuming there are K nodes in each graph (i.e., K seen classes), each node has $K-1$ edge weights to all the other nodes. We define the edge weight w_{ij} between nodes i, j as

$$w_{ij} = cosine\_similarity(z_I^{spec_l}, z_I^{spec_k}) = S_c\left(z_I^{spec_l}, z_I^{spec_k}\right), \quad (6)$$

where $z_I^{spec_l}$ and $z_I^{spec_k}$ are the class specific features, respectively, of classes l and k for sample images I. Cosine similarity is a commonly used metric employed to compare latent representations. Since its range of values is bounded , the cosine similarity is also a good option for its inclusion in a loss. Note that each graph has a total of $\frac{K(K-1)}{2}$ edge links.

1) Informativeness Dictionary: Our objective is to create a dictionary that quantifies the multi-label relationships. We average the inter-node edge weights across all graphs, to get an 'average' graph. Each inter-node link value quantifies the average cosine similarity across all training samples from the seen class. An example inter-node similarity matrix is depicted in Table II with the mean and standard deviations(std), and we refer to this matrix as $Dict_{Spe}$ the multi-label dictionary from class specific features. Each row shows the average cosine similarity for the label with other corresponding labels. The diagonal elements are all one and the matrix is symmetric. Note that the std values are provided for completeness whereas the mean values are used for further calculations. Any synthetically generated sample will preserve this relationship between seen labels by using appropriate loss functions. The values of cosine similarity for the same labels is different than in Table I as the features are taken from different sources. However, we do observe that the values are similar for many label pairs in Table I.

	Atl.	Card.	Cons.	Edema	Eff.
Atelectasis	1	0.80(0.37)	0.89(0.32)	0.90(0.30)	0.68(0.39)
Cardiomegaly	0.80(0.37)	1	0.91(0.31)	0.91(0.32)	0.89(0.33)
Consolidation	0.89(0.32)	0.91(0.31)	1	0.94(0.30)	0.80(0.35)
Edema	0.90(0.30)	0.91(0.32)	0.94(0.30)	1	0.83(0.35)
Effusion	0.68(0.39)	0.89(0.33)	0.80(0.35)	0.83(0.35)	1

TABLE II: Example of the multi-label similarity dictionary from saliency maps for seen classes only. This is an example dictionary for K=5 seen classes.

E. SSL Based Clustering

Having created multi-label dictionaries from text reports and imaging, our next step is to synthesize multi-label features that will play an important role in training the classifier to recognize seen and unseen classes. Before determining the centroids of different class specific features, which function as reference vectors (or *class anchor vectors* for individual classes [30]), to determine whether synthesized features have characteristics of the desired classes. We use class-specific features, z^{spec_l}, and apply self-supervised learning (SSL) based online clustering approach, SwAV (**Sw**apping **A**ssignments between multiple **V**iews) [8] to determine the seen class centroids. Our experimental results in Figure 3 show clustering using class-specific features results in better cluster separability than image features obtained from pre-trained feature extractors.

Let the number of seen and unseen classes be, respectively, n_S and n_U. We first cluster seen class features into n_S clusters and obtain their centroids as $C_S = c_1, \cdots, c_{n_S}$. We enforce the constraint that the semantic relationship between the seen class centroids should be close to that obtained from $Dict_{Spe}$. This is achieved by constructing a matrix of inter-label similarities using the cosine distance between the cluster centroids at each iteration, denoted as $Cent_{Seen}(i, j)$. We

then calculate an element-wise difference between $Dict_{Spe}$ and $Cent_{Seen}(i,j)$:

$$\mathscr{L}_{ML-Seen} = \frac{1}{n_S^2}\sum_i\sum_j Dict_{Spe}(i,j) - Cent_{Seen}(i,j). \tag{7}$$

Since the matrix of cosine similarities is a square matrix having n_S rows and columns, it is divided by a factor of n_S^2 to get a normalized distance measure. $L_{ML-Seen}$ is the loss for multi-label **seen** classes.

In the next pass, we compute the clusters $C_U = c_{n_S+1}, \cdots, c_{n_S+n_U}$ of the n_U unseen classes using the following additional constraints:

1) The centroids in C_S are kept fixed. Since the centroids C_S have been computed from labeled samples, we assume that the computed centroids are reliable and are not changed in the second stage.
2) We add a constraint that the semantic relationship between the seen and unseen class centroids should follow the dictionary $Dict_{Text}$ created using the text embedding vectors, as described in Section III-C. This condition is implemented using:

$$\mathscr{L}_{ML-All} = \frac{1}{N^2}\sum_i\sum_j Dict_{Text}(i,j) - Cent_{All}(i,j), \tag{8}$$

where $Cent_{All}$ refers to the changing matrix of cluster centroid similarities for all seen and unseen classes. $N = n_S + n_U$ is the total number of classes - including seen and unseen classes.

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 [8]:

$$\mathscr{L}(x_t, x_s) = \ell(x_t, q_s) + \ell(x_s, q_t), \tag{9}$$

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 the same information, we can predict the cluster assignment from the other feature.

The final loss term for **clustering** all class samples is

$$\mathscr{L}_{Clust} = \mathscr{L}(x_s, x_t) + \lambda_4\mathscr{L}_{ML-Seen} + \lambda_5\mathscr{L}_{ML-All}. \tag{10}$$

where λ_4, λ_5 are the weights for above loss terms. Thus, we obtain a set of cluster centroids for both the seen and unseen classes which guide the feature generation step.

F. Feature Generation Network

In the feature generation step we synthesize the class-specific features of unseen and seen classes following the steps in [57]. Given the training images of seen classes and unlabeled images of the unseen classes, we learn a generator $G : \mathscr{E}, \mathscr{Z} \longrightarrow \mathscr{X}$, which takes a class label vector $e^y \in \mathscr{E}$ and a Gaussian noise vector $z \in \mathscr{Z}$ as inputs, and generates a feature vector $\tilde{x} \in \mathscr{X}$. The discriminator $D : \mathscr{X}, \mathscr{E} \rightarrow [0,1]$ takes a real feature x, or synthetic feature $\tilde{x}$, 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]:

$$\begin{aligned}\mathscr{L}_{WGAN} = \min_G \max_D \mathbb{E}[D(x,e^y)] - \mathbb{E}[D(\tilde{x},e^y)] \\ -\lambda\mathbb{E}[(\|\nabla_{\tilde{x}} D(\tilde{x},e^y)\|_2 - 1)^2],\end{aligned} \tag{11}$$

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

The discriminator D is a classifier that determines whether the generated feature vector $\tilde{x}$ belongs to one of the seen classes. As the anchor vectors (i.e., the cluster centers) are fixed, we calculate the cosine similarity between the generated vector $\tilde{x}$ and the anchor vector corresponding to the desired classes. Since we are synthesizing multi-label features, it is expected that the cosine similarities of the synthetic vector will be high with respect to the centroids of the desired classes. We integrate these conditions in the following formulation:

$$\mathscr{L}_{ML-Syn} = \sum_{l_y}(1 - \langle \tilde{x}, c_y \rangle) \tag{12}$$

$\mathscr{L}_{ML-Syn}$ is termed as the multi-label synthetic loss. If $\tilde{x}$ truly represents the set of desired classes y, then the cosine similarity between $\tilde{x}$ and the corresponding anchor vectors c_y should be high and the corresponding loss is low.

As part of our method we assume that the total number of classes are known which we divide into seen and unseen classes. During model training we have knowledge of the seen class samples and their labels. However, for the unseen classes we only know the number of unseen classes without any label information. The standard practice in GZSL [56] is to learn plausible representations of unseen classes from seen class features, and a reference for the unseen classes is assumed, e.g. class attribute vectors in natural images or embeddings of labels of unseen classes as in our proposed approach. The label domain for seen and unseen classes is the same.

G. Training, Inference and Implementation

The final loss function for **feature generation** is:

$$\mathscr{L} = \mathscr{L}_{WGAN} + \lambda_6\mathscr{L}_{ML-Syn} \tag{13}$$

where λ_6 is a weight balancing the contribution of the different terms. Once training is complete, we specify the label of desired classes and input a noise vector to G which synthesizes a new feature vector. We combine the synthesized target features of the unseen classes $\tilde{x}^u$ with real and synthetic features of seen class $x^s, \tilde{x}^s$ to construct the training set. We then train a multi-label sigmoid classifier by minimizing the negative log-likelihood loss:

$$\min_\theta -\frac{1}{|\mathscr{X}|}\sum_{(x,y)\in(\mathscr{X},\mathscr{Y})} \log P(y|x,\theta), \tag{14}$$

where $P(y|x,\theta) = \frac{\exp(\theta_y^T x)}{1+\exp(\theta_y^T x)}$ is the classification probability and θ denotes classifier parameters.

The steps in Eqns.11,12 are part of the training process. The core step is the feature generation or synthesis part

(Section III-F) where the objective is to generate features for unseen classes from the available seen class features. To achieve this objective we first obtain cluster centroids of seen and unseen classes. Thereafter when a feature vector is generated we compare it with the centroid of the desired class label (either seen or unseen class) to determine its authenticity. After generating multiple samples of unseen classes, they are combined with seen class samples to train a multi-label classifier. This multi-label classifier is then used at test time to classify each sample. We note that relying only on clustering unseen classes is not very informative due to which we incorporate additional information from multi-modal dictionaries.

Inference: Given initial seen and unseen class samples, the clustering stages yield class centroids. The subsequent feature synthesis module generates samples of different classes for classifier training, and applying to test features.

Implementation Details: We compare the results of our method for medical images with seven other 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, we use our feature disentanglement approach to obtain class-specific features. 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 [24] as our optimizer, and the momentum was set to $(0.9, 0.999)$. The values of loss term weights are $\lambda_{CL} = 0.6, \lambda_3 = 0.9$. 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.

H. 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,

$$H = \frac{2 \times Acc_U \times Acc_S}{Acc_U + Acc_S}, \tag{15}$$

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 the following chest X-ray datasets for multi-label classification tasks.

1) **NIH Chest X-ray** Dataset: For lung disease classification we adopted the NIH Chest X-ray14 dataset [53] having $112,120$ expert-annotated frontal-view X-rays from $30,805$ unique patients and has 14 disease labels. Original images were resized to 224×224. A pre-trained ResNet-101 was fine-tuned using the CheXpert dataset [21] and the chosen baseline FSL was from [44]. 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 $\lambda_1 = 1.1, \lambda_2 = 0.7, \lambda_3 = 0.9, \lambda_4 = 1, \lambda_5 = 1.1, \lambda_6 = 0.9$.
2) **CheXpert** Dataset: We used the CheXpert dataset [21] consisting of $224,316$ chest radiographs of $65,240$ patients labeled for the presence of 14 common chest conditions. Original images were resized to 224×224. A pre-trained ResNet-101 was finetuned using the NIH dataset [53] and the baseline FSL method was of [43], 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 $\lambda_1 = 1.2, \lambda_2 = 0.8, \lambda_3 = 1.1, \lambda_4 = 1.1, \lambda_5 = 1.0, \lambda_6 = 1.1$.
3) **PadChest** Dataset [7]: consisting of $160,868$ images from $67,625$ patients. Hyperparameter values are $\lambda_1 = 1.3, \lambda_2 = 0.9, \lambda_3 = 0.9, \lambda_4 = 1.3, \lambda_5 = 1.2, \lambda_6 = 1.1$.

A $70/10/20$ split at patient level was done to get training, validation and test sets for all datasets.

B. Comparative Study Methods

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

1) SDGN- Self-supervised learning GZSL method of [55].
2) FSL- Top performing fully supervised methods of corresponding datasets. For FSL baselines we implement the different methods referred in the description of individual datasets.
3) Method of [15] using feature disentanglement and controllable pseudo-sample synthesis.
4) [25] that promotes intra-class compactness with inter-class separability on both seen and unseen classes in the embedding space and visual feature space.
5) [47] leveraging visual and semantic modalities to distinguish seen and unseen categories.
6) [18]: the Multi-label GZSL method using BioBERT features.
7) [28]: a graph-based Multi-Label GZSL approach
8) MedCLIP: the medical vision language model trained on Xrays [54]

Following common practices for GZSL we report accuracy for seen and unseen classes. Our method is denoted as ML-GZSL (**M**ulti **L**abel **GZSL**) and is suffixed with the appropriate language model used for encoding text features. The GZSL methods dealing with natural images use class attribute vectors, and when applying them to medical images we replace the attribute vectors with class centroids.

C. Generalized Zero Shot Learning Results

Classification results for medical images shown in Table III 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.Our proposed ML-GZSL method does as good as the multi-label fully supervised learning (FSL) benchmark (using a DenseNet-121 classifier). Class specific features play an important role here since they focus on the features relevant to specific classes and provide more discriminatory information than an FSL approach.

The feature generation network is an important component of the entire pipeline. Its purpose is to generate plausible representations of the unseen class. Consequently the generated feature vectors are compared with representative vectors (or centroids) of the desired class. Without using the feature generation component, an alternative approach to unseen class feature generation is a weighted combination of seen class features (similar to MixUp [59]). However, it is observed that such an approach generates unrealistic feature vectors which leads to poor performance (as shown by the values in Table III).

We used the image encoder output, and the text encoder output of MedCLIP to create new multimodal dictionaries and the results are shown in Table III (ML-GZSL$_{MedCLIP-Dict}$). The output performance is slightly better than our proposed method since the text encoder is a BioClinicalBERT model[3], which is an enhanced version of the BioBERT model used in our method. The use of SwinTransformer as the vision encoder model also contributes to the better performance over Dict$_{Spe}$. Additionally, the MedCLIP model can also output a joint visual-text encoder embedding, which is a multimodal feature vector. When using this vector for classification we obtain much improved results (ML-GZSL$_{MedCLIP-Joint}$) than our approach and ML-GZSL$_{MedCLIP}$. This is due to the more representative vectors learned by using the joint learning layer in MedCLIP and enables the model to better capture the multimodal interactions. We also show results when using the clinical version of RoBERTa (**R**obustly **o**ptimized **BERT** **a**pproach) [33]. The results demonstrate the modular nature of our approach wherein different text encoders and vision language models can be used without affecting system performance.

D. Ablation Studies

Table IV shows results for ablation studies, which are grouped under two categories: 1) Feature disentanglement and 2) Clustering using multi-label dictionaries. For the ablation methods related to feature disentanglement we exclude each of the three loss terms - $\mathscr{L}_{agn}$,$\mathscr{L}_{spec}$ and $\mathscr{L}_{agn-spec}$- and report the results as ML-GZSL$_{MedCLIP-Joint}$ $-$ w/o $\mathscr{L}_{agn}$, ML-GZSL$_{MedCLIP-Joint}-w/o$ $\mathscr{L}_{spec}$, and ML-GZSL$_{MedCLIP-Joint}$ $-$ w/o $\mathscr{L}_{agn-spec}$. We also compare with the results of using image features obtained from a CNN-based feature extractor (ResNet50 trained on Imagenet), which we denote as 'pre-train'. We observe that the class-specific features has the greatest influence on the results and excluding it, ML-GZSL$_{MedCLIP-Joint}$ $-$ w/o $\mathscr{L}_{spec}$ results in maximum degradation of performance compared to ML-GZSL. ML-GZSL$_{MedCLIP-Joint}-w/o$ $\mathscr{L}_{agn-spec}$ shows the next worst performance, while ML-GZSL$_{MedCLIP-Joint}$ $-$ w/o $\mathscr{L}_{agn}$ shows the least difference among the three methods. These results highlight the importance of the class-specific features and at the same time illustrate that the class-agnostic features have a relatively smaller influence on the method's performance. This is desirable since our objective for feature disentanglement was to get complementary features.

The second category of ablation experiments are related to learning the multi-modal multi-label dictionary, clustering and feature synthesis. The primary goal of dictionary learning is to influence clustering and feature synthesis. We conduct two set of experiments where we exclude Dict$_{Spe}$ (ML-GZSL$_{MedCLIP-Joint}$ $-$ w/o $\mathscr{L}_{ML-Seen}$) and Dict$_{Text}$ (ML-GZSL$_{MedCLIP-Joint}$ $-$ w/o $\mathscr{L}_{ML-All}$). The results in Table IV show the two dictionaries have similar influences on the outcome, with Dict$_{Text}$ exerting a greater influence due to its ability to encode more information from all classes. Excluding $\mathscr{L}_{ML-Syn}$ uses only the Wasserstein loss for feature synthesis, without including the class centroids. This results in significant performance degradation since there is no mechanism to check the realism of synthetic features. This leads to a severe reduction in performance as the classifier is trained with lots of spurious samples, which affects the final performance.

In Figures 3 (d,e,f) we show the t-sne visualizations of different settings where specific terms are excluded in the cost function. We observe that the worst results are shown in Figure 3 (f) for ML-GZSL$_{w/o\ \mathscr{L}_{ML-All}}$ (since the clusters are very close to each other, thus hampering accurate classification), followed by ML-GZSL$_{w/o\ \mathscr{L}_{ML-Seen}}$ (Figure 3 (e)) and ML-GZSL$_{w/o\ \mathscr{L}_{spec}}$ (Figure 3 (d)). The visualizations support the observations reported in Table IV about the performance of the corresponding ablation methods.

E. Hyperparameter Selection

Figure 4 shows the harmonic mean values for the NIH Chest X-ray dataset for different values of hyperparameters $\lambda_1, \lambda_2, \lambda_3$. The λ'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 were monitored. We start with the base cost function of Eqn. 1, and first select the optimum value of λ_1 by keeping $\lambda_2 = \lambda_3 = 1$. λ_1 value is fixed and we then determine optimal λ_2, and subsequently λ_3. Similarly for values of λ_4, λ_5, we start with the cost function of Eqn. 10, fix $\lambda_5 = 1$ and search for the optimum value of λ_4. Then we fix λ_4 and search for the optimal value of λ_5. 'Finally we search for the optimal value of λ_6 in Eqn. 13. The plots for the loss function with different values of λ are shown in Figure 4.

[3] $https://huggingface.co/emilyalsentzer/Bio\_ClinicalBERT$

Method	NIH X-ray			CheXpert			PadChest		
	S	U	H	S	U	H	S	U	H
	Single Label GZSL Methods								
f-VAEGAN [58]	82.9(3.6)	80.0(3.7)	81.4(3.7)	88.5(3.3)	87.6(3.7)	88.0(3.7)	81.0(3.5)	78.4(3.7)	79.7(3.6)
SDGN [55]	84.4(3.3)	81.1(3.7)	82.7(3.5)	89.8(3.3)	88.3(3.4)	89.0(3.4)	82.3(3.2)	80.0(3.5)	81.1(3.5)
Feng [15]	84.7(3.4)	81.4(3.9)	83.0(3.7)	90.2(3.2)	88.6(3.4)	89.4(3.3)	82.5(3.3)	80.2(3.4)	81.3(3.4)
Kong [25]	84.8(3.6)	81.2(3.7)	82.9(3.7)	90.0(3.3)	88.7(3.4)	89.3(3.4)	82.7(3.4)	80.5(3.5)	81.6(3.5)
Su [47]	84.5(3.5)	81.4(3.5)	82.9(3.5)	90.3(3.3)	88.6(3.4)	89.4(3.4)	82.3(3.6)	79.8(3.8)	81.03(3.7)
	Multi Label GZSL Methods								
Hayat [18]	79.1(3.8)	69.2(4.3)	73.8(4.1)	81.2(3.7)	79.8(3.9)	80.5(3.8)	77.3(4.2)	68.1(4.3)	72.4(4.3)
Lee [28]	85.1(3.5)	81.3(3.7)	83.1(3.6)	87.4(3.2)	85.7(3.1)	86.5(3.2)	82.9(3.5)	78.4(3.6)	80.6(3.6)
Huynh [20]	84.7(3.4)	80.8(3.5)	82.7(3.5)	86.9(3.1)	85.1(3.3)	86.0(3.2)	82.5(3.3)	77.3(3.6)	79.8(3.5)
Mixup [59]	79.3(3.8)	77.1(3.8)	78.2(3.8)	81.6(3.4)	80.2(3.5)	80.9(3.5)	78.4(3.5)	75.8(3.7)	77.1(3.6)
	Multi Label Benchmarks								
FSL(Multi Label)	86.0(3.2)	85.1(3.3)	85.5(3.3)	90.8(3.1)	90.5(3.1)	90.6(3.1)	88.4(3.3)	86.5(3.4)	87.4(3.4)
Mahapatra [36]	84.3(3.3)	83.2(3.6)	83.7(3.5)	88.9(3.1)	88.5(3.2)	88.7(3.2)	86.2(3.3)	84.1(3.5)	85.1(3.5)
	Proposed Method And Variants								
ML-GZSL$_{BioBERT}$	86.2(3.4)	85.0(3.6)	85.6(3.5)	90.8(3.1)	90.2(2.9)	90.5(3.0)	88.2(3.4)	86.1(3.6)	87.1(3.5)
ML-GZSL$_{MedCLIP-Dict}$	87.7(2.2)	86.8(2.3)	87.2(2.3)	91.9(2.3)	91.6(2.4)	91.7(2.4)	89.5(2.6)	87.7(2.7)	88.6(2.7)
ML-GZSL$_{MedCLIP-Joint}$	**88.3(2.3)**	**87.2(2.2)**	**87.7(2.2)**	**92.2(2.0)**	**92.0(2.4)**	**92.1(2.2)**	**90.1(2.4)**	**88.3(2.5)**	**89.2(2.4)**
ML-GZSL$_{RoBERTa}$	87.0(3.5)	85.8(3.6)	86.4(3.6)	91.8(3.1)	91.2(3.3)	91.4(3.3)	89.4(3.3)	87.2(3.5)	88.1(3.4)

TABLE III: **GZSL Results For chest xray Images in Multi-Label setting:** Average per-class classification accuracy (%) and harmonic mean accuracy (H) of generalized zero-shot learning when test samples are from seen or unseen classes. Results demonstrate the superior performance of our proposed method. The FSL performance is the upper bound for a specific classifier. The best results are shown in bold.

Method	NIH X-ray			CheXpert			PadChest		
	S	U	H	S	U	H	S	U	H
ML-GZSL$_{MedCLIP-Joint}$	**88.3(2.3)**	**87.2(2.2)**	**87.7(2.2)**	**92.2(2.0)**	**92.0(2.4)**	**92.1(2.2)**	**90.1(2.4)**	**88.3(2.5)**	**89.2(2.4)**
	Feature Disentanglement Effects								
$w/o\ \mathscr{L}_{agn-spec}$ (Eqn. 5)	85.6(2.5)	84.1(2.2)	84.8(2.3)	90.1(2.5)	89.8(2.4)	89.9(2.4)	88.3(2.1)	85.7(2.3)	87.0(2.1)
pre-train	85.3(2.7)	84.1(2.6)	84.7(2.6)	89.7(2.6)	89.1(2.5)	89.4(2.5)	87.6(2.6)	85.1(2.8)	86.3(2.7)
$w/o\ \mathscr{L}_{agn}$ (Eqn. 4)	86.4(2.8)	84.3(2.9)	85.3(2.8)	90.6(2.4)	88.0(2.7)	89.3(2.6)	88.2(2.9)	85.4(3.1)	86.8(3.0)
$w/o\ \mathscr{L}_{spec}$ (Eqn. 3)	84.4(3.1)	82.7(3.4)	83.5(3.3)	88.9(3.0)	86.9(2.9)	87.9(3.0)	87.0(3.1)	86.1(3.2)	86.5(3.2)
	Effect of Dictionary/Clustering								
$w/o\ \mathscr{L}_{ML-Seen}$ (Eqn. 7)	85.1(3.2)	82.8(3.4)	83.9(3.4)	88.8(2.9)	86.6(3.2)	87.7(3.1)	87.4(3.1)	84.4(3.4)	85.9(3.3)
$w/o\ \mathscr{L}_{ML-All}$ (Eqn. 8)	83.9(3.2)	82.1(3.6)	83.0(3.3)	88.2(2.9)	86.2(3.4)	87.2(3.2)	86.1(3.2)	83.8(3.5)	84.9(3.3)
$w/o\ \mathscr{L}_{ML-Syn}$ (Eqn. 12)	82.5(3.6)	81.1(3.7)	81.8(3.6)	87.0(3.1)	84.1(3.4)	85.5(3.2)	85.0(3.4)	83.4(3.7)	84.2(3.5)

TABLE IV: Ablation Results using ML-GZSL$_{MedCLIP-Joint}$: Average per-class classification accuracy (%) and harmonic mean accuracy (H) of generalized zero-shot learning when test samples are from seen (Setting S) or unseen (Setting U) classes. The best results are shown in bold.

F. Realism of Synthetic Features

We reconstruct the x-ray images from the synthetic feature vectors using the feature disentanglement autoencoders' decoder part. We select 1000 such synthetic images from 14 classes of the NIH dataset and ask two trained radiologists, having 12 and 14 years experience in examining chest xray images for abnormalities, to identify whether the images are realistic or not in terms of images with the correct type of the disease. Each radiologist was blinded to the other's answers.

Results for ML-GZSL show one radiologist (RAD 1) identified $912/1000$ (91.2%) images as realistic while RAD 2 identified 919 (91.9%) generated images as realistic. Both of them had a high agreement with 890 common images (89.0% -"$Both\ Experts$" in Table V) identified as realistic. Considering both RAD 1 and RAD 2 feedback, a total of 941 (94.1%) unique images were identified as realistic ("$Atleast\ 1\ Expert$"). Subsequently, $59/1000$ (5.9%) of the images were not identified as realistic by any of the experts ("$No\ Expert$"). Agreement statistics for other methods are summarized in Table V.

Agreement Statistics	Both Experts	Atleast 1 Expert	No Expert
ML-GZSL$_{BioBERT}$	89.0 (890)	94.1 (941)	5.9 (59)
ML-GZSL$_{MedCLIP-Joint}$	**92.2** (922)	**95.2** (952)	**4.8** (48)
[28]	85.1 (851)	88.0 (880)	12.0 (120)
[20]	83.4 (834)	85.1 (851)	14.9 (149)
[55]	81.9 (819)	83.9 (839)	16.1 (161)

TABLE V: Agreement statistics on NIH dataset for different GZSL methods amongst 2 radiologists. Numbers outside the bracket indicate agreement percentage while numbers within brackets indicate actual numbers out of 1000 samples. The best results are shown in bold.

G. Results on Additional Datasets

We also show in Table VI results on the *multi-class* MedMNIST dataset [60] due to its balanced and standardized datasets spanning across various modalities. The im-

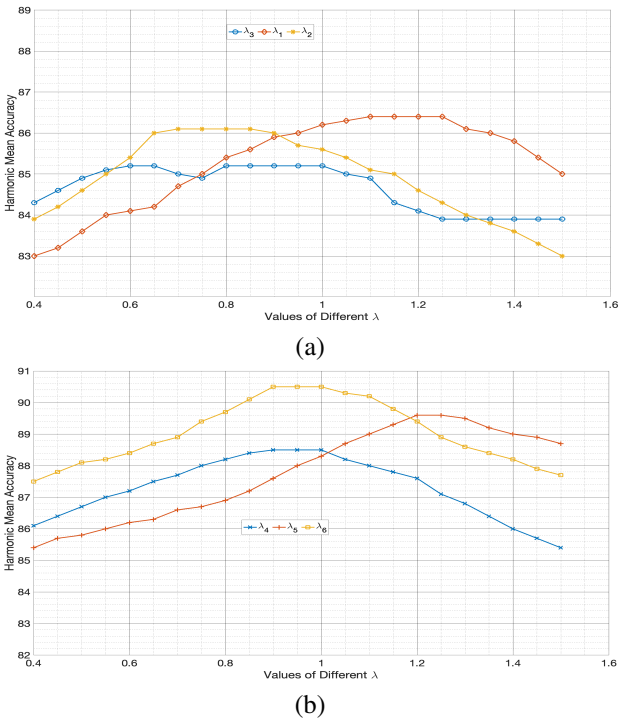

Fig. 4: Hyperparameter Plots showing the value of H and classification accuracy for different values of λ. The observed trends justify our final choice of the values.

ages in the dataset have one label out of multiple possible labels. We select subsets of the collection appropriate for multi-class disease classification, namely, BreastMNIST [2] having $546/78/156$ breast ultrasound images in the training/validation/test split for malignancy detection, RetinaMNIST [31] having $1080/120/400$ training/validation/test fundus images for diabetic retinopathy severity grading, and TissueMNIST having $165,466/23,640/47,280$ training/validation/test Kidney Cortex Microscope images for multiple disease classification. The results show clearly that our approach outperforms other competing methods for the multi-class setting where images can have only one label out of multiple possible labels.

V. Discussion

Our approach is particularly useful in scenarios where the number of disease classes are known but labeled samples of all classes cannot be accessed due to the infrequent occurrence of such cases or lack of expert clinicians to annotate complex cases. While fully supervised settings still provide the best performance, they are dependent upon sufficient labeled samples. Our system can identify the new diseases and clinical experts will be able to diagnose and label them. There is also a need for systems that can adapt to new classes introduced and incorporated into the nomenclature used in radiology (e.g. Fleischner society pulmonary nodule where recommendations are provided regarding the follow-up and management of indeterminate pulmonary nodules detected incidentally on CT).

Importance of Multi-Modal Multi-Label Dictionary: We create two multi-label dictionaries based on class-specific and text features. These help us have two references for determining the class centroids of seen and unseen classes, which subsequently improves the accuracy of synthesized features. The cosine similarities between text embeddings of different disease classes quantifies their mutual semantic relationships, which is helpful in generating features involving unseen disease classes. We also quantify the semantic relationship between seen classes using class-specific features, which plays an important role during clustering and feature generation steps. Our experiments show the importance of the two multi-label dictionaries in improving overall GZSL performance. A limitation of our method is the use of multiple encoders (equal to the number of classes) for feature disentanglement. This adds to computational complexity with increasing number of labels. However a mitigating factor is the computation burden is encountered during the training stage and during test-time we only use the class-specific encoders. Nevertheless we aim to address this factor in future work by use of foundation models trained on multimodal data.

Importance of Multi-Label GZSL: Multi-label GZSL is important in the context of CXRs in particular and medical image analysis in general. Since new images may have multiple co-occurring disease labels it is essential to have a mechanism that can generate synthetic samples with different co-occurring disease labels. Since SOTA methods focus on synthesizing samples of single-label cases, there is a need for methods that can effectively learn discriminative features for multi-label samples. This makes multi-label feature synthesis an important step in the whole setup. Our experiments show the importance of creating text and class-specific feature-based dictionaries for better multi-label feature synthesis.

Importance of Feature Disentanglement: We propose a novel method to disentangle image features into class-specific and class-agnostic features. This is motivated by previous work [22], which show that the current approach of using pre-trained networks to learn image features is sub-optimal for multi-label classification. The original features for all 14 classes when viewed through t-sne does not show a common region for all the classes although there are varying degree of overlap across different pairs of classes. The obtained class-specific features are highly discriminative for the specific class and the class-agnostic features are more common for all classes. In the feature synthesis step we focus on synthesizing features that are a good match across different disease labels. As a result, we are able to achieve high agreement with the template vectors of different desired classes.

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. This shows that our method generates realistic features and does not suffer from unconstrained feature generation wherein the features come from arbitrary distributions.

Performance In Extreme Low-Data Scenarios:

In the GZSL scenario there are many unseen classes. However there is no general assumption on the number of

Method	Breast MNIST			Retina MNIST			Tissue MNIST		
	S	U	H	S	U	H	S	U	H
	Single Label GZSL Methods								
f-VAEGAN [58]	90.2(0.39)	88.2(0.41)	89.2(0.4)	92.8(0.37)	90.2(0.34)	91.5(0.36)	88.2(0.43)	85.1(0.44)	86.6(0.44)
SDGN [55]	92.1(0.35)	89.5(0.39)	90.8(0.37)	95.0(0.31)	91.9(0.34)	93.4(0.33)	90.0(0.36)	87.8(0.39)	88.9(0.38)
Feng [15]	93.7(0.35)	92.1(0.34)	92.9(0.34)	96.1(0.31)	94.2(0.33)	95.1(0.32)	92.1(0.35)	91.1(0.38)	91.6(0.36)
Kong [25]	91.6(0.38)	89.1(0.39)	90.3(0.38)	95.1(0.34)	91.7(0.35)	93.3(0.35)	89.8(0.36)	87.9(0.33)	88.8(0.34)
Su [47]	90.6(0.33)	88.7(0.36)	89.6(0.34)	92.9(0.34)	90.6(0.32)	91.7(0.33)	88.4(0.36)	85.7(0.38)	87.0(0.38)
	Multi Label Benchmarks								
FSL(Multi Class)	96.0(0.31)	95.1(0.32)	95.6(0.32)	97.8(0.29)	96.5(0.3)	97.1(0.32)	95.4(0.33)	93.2(0.35)	94.3(0.35)
Mahapatra [36]	94.0(0.3)	93.1(0.32)	93.5(0.31)	96.4(0.28)	94.5(0.32)	95.4(0.31)	93.4(0.33)	91.2(0.35)	92.3(0.35)
	Proposed Method And Variations								
ML-GZSL$_{BioBERT}$	95.5(0.27)	94.7(0.3)	95.1(0.29)	96.9(0.26)	95.8(0.29)	96.3(0.28)	94.8(0.29)	92.9(0.33)	93.8(0.31)
ML-GZSL$_{MedCLIP-Dict}$	96.7(0.26)	96.1(0.29)	96.4(0.27)	97.1(0.23)	96.3(0.26)	96.7(0.25)	95.7(0.30)	93.7(0.33)	94.6(0.32)
ML-GZSL$_{MedCLIP-Joint}$	**97.3(0.26)**	**96.9(0.29)**	**97.1(0.28)**	**97.6(0.22)**	**96.9(0.24)**	**97.3(0.23)**	**96.2(0.24)**	**94.1(0.25)**	**95.2(0.25)**
ML-GZSL$_{RoBERTa}$	96.1(0.25)	95.2(0.27)	95.6(0.26)	97.5(0.24)	96.3(0.27)	96.8(0.25)	95.4(0.25)	93.5(0.29)	94.4(0.28)

TABLE VI: **GZSL Results for different subsets of MedMNIST dataset in Single-Label setting:** Average per-class classification accuracy (%) and harmonic mean accuracy (H) of generalized zero-shot learning when test samples are from seen or unseen classes. Results demonstrate the superior performance of our proposed method. The FSL performance is the upper bound for a specific classifier. The best results are shown in bold.

labeled samples available for the seen classes. In all datasets and previous works it is assumed that a sizable number of samples are available for the seen classes. However, in many scenarios we face the situation of having very few samples of seen (and unseen) classes. Our preliminary experiments in such low data scenarios (less than 5% labeled samples) suggest that our current method demonstrates very poor performance (AUC< 0.70). This is due to the fact that the current frameworks require many samples to be trained or finetuned. In the absence of networks pre-trained on suitable images the task is very challenging and we hypothesize that use of pre-trained networks like MedCLIP might make it easier to learn suitable representations. In future work we aim to address this problem with potential directions being few-shot classification of seen and unseen classes [11], [46].

VI. Conclusion

We propose a multi-label GZSL approach for chest xray images. Our novel method can accurately synthesize feature vectors of unseen classes by learning multi-modal multi-label dictionary using graph aggregation and class-specific features, along with text embedding relationships. Experimental results show our method outperforms other recent GZSL approaches in literature, and is consistently better across multiple public CXR 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 the infrequent occurrence of such cases or lack of expert clinicians to annotate complex cases. While fully supervised settings still provide the best performance, they are dependent upon sufficient labeled samples.

VII. Acknowledgements

This work was supported by the Swiss National Foundation grant number 212939, and Innosuisse grant number 31274.1.

References

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *In Proc. IEEE CVPR*, pages 2927–2936, 2015.

[2] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. In *Data Brief 28, https://doi.org/10.1016/j.dib.2019.104863*, 2020.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. In *arXiv preprint arXiv:1701.07875*, 2017.

[4] Behzad Bozorgtabar, Dwarikanath Mahapatra, and Jean-Philippe Thiran. Amae: Adaptation of pre-trained masked autoencoder for dual-distribution anomaly detection in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 195–205. Springer, 2023.

[5] Behzad Bozorgtabar, Dwarikanath Mahapatra, Guillaume Vray, and Jean-Philippe Thiran. Salad: Self-supervised aggregation learning for anomaly detection on x-rays. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 468–478. Springer, 2020.

[6] S. Bujwid and J. Sullivan. Large-scale zero-shot image classification from rich and diverse textual descriptions. *arXiv preprint arXiv:2103.09669*, 2021.

[7] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

[8] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[9] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58:101535, 2019.

[10] Y. Chen, Y. Chang, S. Wen, Y. Shi, X. Xu, T. Ho, Q. Jia, M. Huang, and J. Zhuang. Zero-shot medical image artifact reduction. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 862–866, 2020.

[11] Zitian Chen, Subhransu Maji, and Erik Learned-Miller. Shot in the dark: Few-shot learning with no base-class labels, 2021.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] M. Elhoseiny and M. Elfeki. Creativity inspired zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5784–5793, October 2019.

[14] R. Felix, V. Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018.

[15] Y. Feng, X. Huang, P. Yang, J. Yu, and J. Sang. Non-generative generalized zero-shot learning via task-correlated disentanglement and controllable samples synthesis. In *2022 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 9336–9345, 2022.

[16] A. Gaure, A. Gupta, V. K. Verma, and P. Rai. A probabilistic framework for zero-shot multi-label learning. In *The Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

[17] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, 12 2016.

[18] N. Hayat, H. Lashen, and F.E Shamout. Multi-label generalized zero shot learning for the classification of disease in chest radiographs. In *Proceeding of the Machine Learning for Healthcare Conference)*, pages 461–477, 2021.

[19] H. Huang, C. Wang, P. S. Yu, and C.-D. Wang. Generative dual adversarial network for generalized zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 801–810, June 2019.

[20] D. Huynh and E. Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8773–8783, 2020.

[21] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, , and et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *arXiv preprint arXiv:1901.07031*, 2017.

[22] J. Jia, F. He, N. Gao, X. Chen, and K. Huang. Learning disentangled label representations for multi-label classification, 2022.

[23] R. Keshari, R. Singh, and M. Vatsa. Generalized zero-shot learning via over-complete distribution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13300–13308, June 2020.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980,*, 2014.

[25] X. Kong, Z. Gao, X. Li, M. Hong, J. Liu, C. Wang, Y. Xie, and Y. Qu. En-compactness: Self-distillation embedding and contrastive generation for generalized zero-shot learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9296–9305, 2022.

[26] A. Kori and G. Krishnamurthi. Zero shot learning for multi-modal real time image registration. In *arXiv preprint arXiv:1908.06213*, 2019.

[27] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Analysis Machine Intelligence*, 36(3):453–465, 2013.

[28] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. Wang. Multi-label zero-shot learning with structured knowledge graphs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1576–1585, 2018.

[29] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[30] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang. Leveraging the invariant side of generative zero-shot learning. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019.

[31] Ruhan Liu, Xiangning Wang, and et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. volume 3, page 100512, 2022.

[32] X. Liu, P. Sanchez, S. Thermos, A.Q. O'Neil, and S.A. Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516, 2022.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[34] Dwarikanath Mahapatra, Behzad Bozorgtabar, and Zongyuan Ge. Medical image classification using generalized zero shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3344–3353, October 2021.

[35] Dwarikanath Mahapatra and Zongyuan Ge. MR image super resolution by combining feature disentanglement cnns and vision transformers. In *International Conference on Medical Imaging with Deep Learning, MIDL 2022*, volume 172 of *Proceedings of Machine Learning Research*, pages 858–878. PMLR, 2022.

[36] D. Mahapatra, Z. Ge, and M. Reyes. Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. *IEEE Transactions on Medical Imaging*, 41(9):2443–2456, 2022.

[37] Dwarikanath Mahapatra, Steven Korevaar, Behzad Bozorgtabar, and Ruwan B. Tennakoon. Unsupervised domain adaptation using feature disentanglement and gcns for medical image classification. In *ECCV 2022 Workshops*, volume 13807 of *Lecture Notes in Computer Science*, pages 735–748. Springer, 2022.

[38] T. Mikolov, K. Chen, Gr. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *In Proc. ICLR Workshops*, 2013.

[39] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12664–12673, June 2020.

[40] J. Ouyang, E. Adeli, K. M. Pohl, Q. Zhao, and G. Zaharchuk. Representation disentanglement for multi-modal brain mri analysis. In *Information Processing in Medical Imaging*, pages 321–333, 2021.

[41] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. A. Efros, and R. Zhang. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems*, 2020.

[42] A. Paul, T. C. Shen, S. Lee, N. Balachandar, Y. Peng, Z. Lu, and R. M. Summers. Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training. *IEEE Transactions on Medical Imaging*, pages 1–1, 2021.

[43] H.H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and Ha Q. Nguyen. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. In *arXiv preprint arXiv:1911.06475,*, 2020.

[44] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P Lungren, and A.Y Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. In *arXiv preprint arXiv:1711.05225,*, 2017.

[45] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *In Proc. IEEE CVPR*, pages 8247–8255, 2019.

[46] Jin-Woo Seo, Hong-Gyu Jung, and Seong-Whan Lee. Self-augmentation: Generalizing deep networks to unseen classes for few-shot learning. *Neural Networks*, 138:140–149, 2021.

[47] H. Su, J. Li, Z. Chen, L. Zhu, and K. Lu. Distinguishing unseen from seen for generalized zero-shot learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7875–7884, 2022.

[48] X. Sun, Z. Liu, S. Zheng, C. Lin, Z. Zhu, and Y. Zhao. Attention-enhanced disentangled representation learning for unsupervised domain adaptation in cardiac segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 745–754, 2022.

[49] E. Tiu, E. Talius, P. Patel, C.P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 2022.

[50] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[51] Guillaume Vray, Devavrat Tomar, Behzad Bozorgtabar, and Jean-Philippe Thiran. Distill-soda: Distilling self-supervised vision transformer for source-free open-set domain adaptation in computational pathology. *IEEE Transactions on Medical Imaging*, 2024.

[52] J. Wang, C. Zhong, C. Feng, Y. Zhang, J. Sun, and Y. Yokota. Disentangled representation for cross-domain medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2023.

[53] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R.M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *In Proc. CVPR*, 2017.

[54] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. pages 3876–3887, 2022. Funding Information: This work was supported by NSF award SCH-2205289, SCH-2014438, IIS-1838042, NIH award R01 1R01NS107291-01. Publisher Copyright: © 2022 Association for Computational Linguistics.; 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 ; Conference date: 07-12-2022 Through 11-12-2022.

[55] J. Wu, T. Zhang, Z.Zha, J. Luo, Y. Zhang, and F. Wu. Self-supervised domain-aware generative network for generalized zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12767–12776, June 2020.

[56] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Analysis Machine Intelligence*, 41(9):2251–2265, 2018.

[57] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *In Proc. IEEE CVPR*, pages 5542–5551, 2018.

[58] Y. Xian, S. Sharma, B. Schiele, and Z. Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10275–10284,

June 2019.

[59] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Qi Tian, and W. Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, pages 6502–6509, 2020.

[60] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *ISBI*, 2021.

[61] Y. Zhang, B. Gong, and M. Shah. Fast zero-shot image tagging. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.