Medical Image Classification Using Generalized Zero Shot Learning

Dwarikanath Mahapatra[1] Behzad Bozorgtabar[2,3] Zongyuan Ge[4,5]
[1]Inception Institute of Artificial Intelligence, UAE
[2]LTS5, EPFL, Switzerland [3]CIBM, Switzerland [4]Monash University, Australia
[5]Airdoc Research Melbourne, Australia
dwarikanath.mahapatra@inceptioniai.org, behzad.bozorgtabar@epfl.ch, zongyuan.ge@monash.edu

Abstract

In many real world medical image classification settings we do not have access to samples of all possible disease classes, while a robust system is expected to give high performance in recognizing novel test data. We propose a generalized zero shot learning (GZSL) method that uses self supervised learning (SSL) for: 1) selecting anchor vectors of different disease classes; and 2) training a feature generator. Our approach does not require class attribute vectors which are available for natural images but not for medical images. SSL ensures that the anchor vectors are representative of each class. SSL is also used to generate synthetic features of unseen classes. Using a simpler architecture, our method matches a state of the art SSL based GZSL method for natural images and outperforms all methods for medical images. Our method is adaptable enough to accommodate class attribute vectors when they are available for natural images.

1. Introduction

Medical image classification is an important step in computer aided diagnosis. In the present era, deep learning methods have achieved state of the art results for many medical image classification tasks such as diabetic retinopathy grading[15], digital pathology image classification [26] and chest xray images [18, 46], to name a few. Fully supervised learning (FSL) methods that achieve state of the art results have access to disease classes (labels) in the training and test sets. However in many real-world scenarios we may not have access to samples of all possible diseases. A common scenario is the diagnosis of radiological images, such as chest xrays. Unseen classes are generally classified into one of the seen classes, resulting in wrong diagnosis and treatment planning. For deployment in clinical settings it is essential that a machine learning model learns to recognize novel test cases.

Zero shot learning (ZSL) aims to learn plausible representations of unseen classes from labeled data of seen classes, and recognize unseen classes during test time. In a more generalized setting we expect to encounter both seen and unseen classes during the test phase, and a reliable model should accurately predict both classes. This is a case of generalized zero shot learning (GZSL) which is a challenging scenario since we do not want to predict unseen classes as one of the seen classes. We propose a GZSL method for medical image classification using self supervised learning (SSL), demonstrate its effectiveness across different datasets and also shows it's applicability to natural images.

GZSL has been a widely explored topic for natural images [14, 44, 51] 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 class 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 straightforward task to apply state of the art GZSL methods to medical image classification. While this makes GZSL for medical images a challenging task, it is nevertheless essential to tackle this problem due to the potentially immense benefit.

Initial approaches to tackle ZSL [11, 49] learnt cross-modal relationships between visual feature and semantic embeddings (class attribute vectors). Subsequently, recent generative approaches to GZSL [50, 14], used generative adversarial networks (GANs) to optimize the divergence between the data distribution of seen classes and generated features. Consequently, 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 [37, 31] utilize unlabeled data of

unseen classes in a transductive way. However they require two GANs for seen and unseen classes as they do not consider the relations between source and target domains.

However, methods leveraging transductive approaches are particularly relevant for medical classification tasks [47]. Absence of any supervised information from the unseen domain makes it very challenging to differentiate between disease labels, especially when many labels show similar appearance to the untrained eye. In our method we also leverage the unlabeled data of unseen classes as a guidance to train our GZSL method.

Another tricky issue facing GZSL applications in general and medical images in particular is the potentially large semantic gap between images of different classes. Consequently synthesizing such unseen class features from the seen classes can be challenging. Leveraging unlabeled unseen class data (e.g., using anchors) can be effective in bridging the semantic gap [47].In an attemp to address the above challenges our paper makes the following contributions:

1. We propose a GZSL approach using self supervised learning (SSL) for medical image classification. Our method outperforms state of the art methods for multiple medical image datasets, and matches their performance on natural images.

2. We use SSL for: 1) deriving anchor vectors through **improved** clustering; and 2) feature synthesis of seen and unseen classes.

3. We achieve GZSL of medical images without using class attribute vectors commonly used for natural images. This is important for real world clinical scenarios where defining class attribute vectors is a time consuming and expensive task.

2. Prior Work

(Generalized) Zero-Shot Learning: In Zero-Shot Learning [49], the goal is to recognize classes not encountered during training. External information about the novel classes may be provided in forms of semantic attributes [24], visual descriptions [1], or word embeddings [33]. Zero-shot learning has been addressed using Generative Adversarial Networks (GANs) [50], Variational Autoencoders (VAE) [41] or both of them [51].

In generalized zero-shot learning (GZSL), the purpose is to recognize images from known and unknown domains. Many works [14, 44, 41, 50, 51] obtain impressive 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. [17] describes a Generative Dual Adversarial Network (GDAN) which couples a Generator, a Regressor and a Discriminator. The interaction between the three components produces various visual features conditioned on class labels. Keshari et al. [21] use overcomplete distributions to generate features of the unseen classes, while Min et al. [34] 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 specifying the class label of the desired output feature. GZSL for medical image tasks have seen limited applications such as registration [23] and artefact reduction [13].

Self-Supervised Learning: These methods consist of two main approaches; 1) pre-text tasks and 2) down-stream tasks. Solving pre-text tasks learns a proper data representation, although the task itself may not be relevant, while 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 [16] and SimCLR [12] are popular and give state-of-the-art results for down-stream task-based methods. Self-supervised learning (SSL) also addresses labeled data shortage and has found wide use in medical image analysis by using innovative pre-text tasks for active learning [30], anomaly detection [6], data augmentation [28], semi-supervised histology classification [27],stain normalization [29] and registration [43]. Recent works also use self supervision for domain adaptation [40] and perhaps the first work to combine GZSL and SSL [47]. While our work is inspired from [47] 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 medical images do not have defined class attribute vectors we use a simpler architecture for GZSL. 2) [47] use 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.

3. Method

3.1. Method Overview

Figure 1 depicts our proposed workflow. In the first step we generate anchor vectors (cluster centroids) by using SSL within the SwAV clustering approach [10]. 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. The second step involves feature generation that

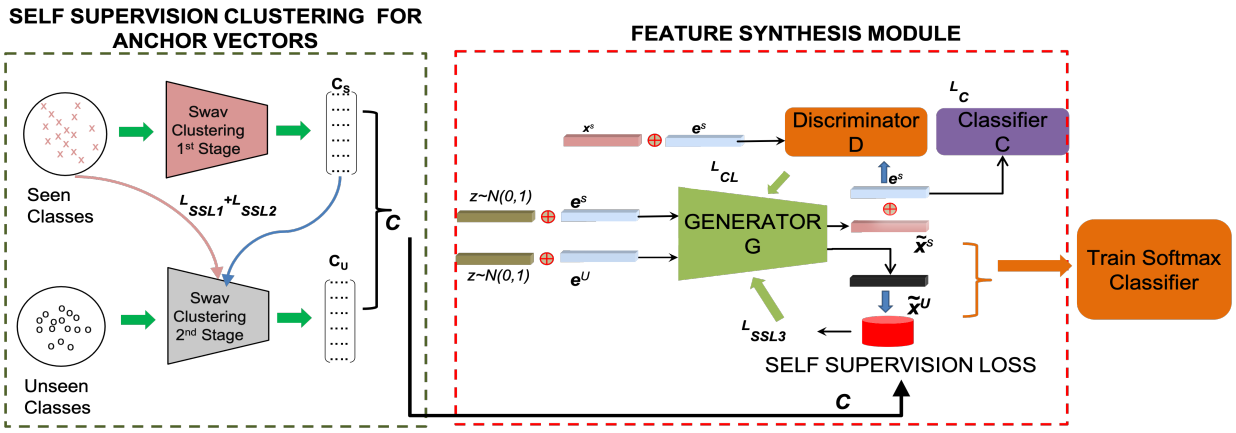


Figure 1. Architecture of proposed SC-GZSL method. In the first step we generate anchor vectors (cluster centroids) by using SSL within the SwAV clustering approach [10]. We have two clustering stages: one for Seen class samples and second for Unseen classes. Feature generation leverages one Generator and one Discriminator alongwith anchor vectors (from clustering) to derive SSL loss terms.

takes a noise vector and desired class label of output vector to synthesize features. Anchor vectors from the clustering stage are used to derive SSL based loss terms. Synthesized and real features of unseen and seen classes are used to train a softmax classifier for identifying different disease classes.

3.2. SSL Clustering To Obtain Anchor Vectors

Let the number of classes in the Seen set be n_S, and the number of classes in the Unseen set is n_U. We assume that the total number of classes is known.We learn anchor vectors of the different classes by using the SSL based online clustering approach SwAV (**Sw**apping **A**ssignments between multiple **V**iews) [10], and introduce additional SSL inspired loss terms. Typical offline clustering methods [4, 9] alternate between cluster assignment and centroid update. Since they require multiple passes over the dataset, such methods are slow for online clustering. To overcome the high training time and inspired by contrastive instance learning [48], [10] enforce 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 take the cluster centers to be class anchor vectors since they give a reliable representation of the corresponding class. We choose to compute the anchor vectors in an online fashion since the number of unseen classes may change in a dynamic way depending upon the specific use case. 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 computing 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:

$$\mathcal{L}(x_t, x_s) = \ell(x_t, q_s) + \ell(x_s, q_t) \tag{1}$$

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_θ transforms I_{nt} to a feature vector which is projected to the unit sphere, i.e., $x_{nt} = f_\theta(x_{nt})/\|f_\theta(x_{nt})\|_2$. The cluster assignment q_{nt} is computed by determining the distance of x_{nt} to the set of cluster centroids, $c_1, \cdots, c_K$. C denotes a matrix whose columns are $c_1, \cdots, c_k$.

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_i and all columns in C, i.e.,

$$\ell(x_t, q_s) = -\sum_k q_s^{(k)} \log p_t^{(k)}, \;\; p_t^{(k)} = \frac{\exp \frac{x_t^\top c_k}{\tau}}{\sum_{k'} \exp \frac{x_t^\top c_k}{\tau}} \tag{2}$$

where $\tau = 0.1$ is the temperature parameter [48]. Computing this loss over all images and augmentations results in the following loss function for swapped prediction:

$$\mathcal{L}(x_t, x_s) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{s,t\sim T}\Bigg[\frac{x_{nt}^\top C q_{ns}}{\tau} + \frac{x_{ns}^\top C q_{nt}}{\tau} - \log\sum_{k=1}^{K}\exp\left(\frac{x_{nt}^\top c_k}{\tau}\right) - \log\sum_{k=1}^{K}\exp\left(\frac{x_{ns}^\top c_k}{\tau}\right)\Bigg]. \tag{3}$$

This loss function is jointly minimized with respect to the centroids in C and parameters θ of f_θ.

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,

$$\max_{Q \in \mathcal{Q}} Tr(Q^\top C^\top X) + \epsilon H(Q), \tag{4}$$

where H is the entropy function, $H(Q) = -\sum_{ij} Q_{ij} \log Q_{ij}$ and $\epsilon = 0.05$ controls smoothness of mapping. A high ϵ could potentially results in a trivial solution where all samples collapse into an unique representation and are assigned uniformly to all prototypes.

Our Novel Contribution: We use the concept of anchor vectors to bridge the gap between seen and unseen classes, which is determined by the following steps: Assuming we have n_S seen classes we first cluster the Seen class images into n_S clusters and obtain their centroids as $C_S = c_1, \cdots, c_{n_S}$. 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 do not change since they have been computed from the seen classes.

2. A self supervised constraint is added where the centroids of the unseen classes are forced to be different from the seen class centroids. This is done to account for the situation that some of the Unseen classes may be semantically close to one or more Seen classes. This may happen when images of different disease labels have very similar appearance which can be a common occurrence for radiological images. This condition is implemented using:

$$\mathcal{L}_{SSL1} = \min \Big(CoSim(C_S^i, C_U^j), \sigma_1 \Big) \tag{5}$$

   Here $\sigma_1 = 0.15$ is a parameter that determines the semantic distance between the centroids, and CoSim denotes cosine similarity.

3. We add a second self supervised constraint that the similarity of seen class sample, x_s^i, with its corresponding class centroid C_S^i is higher than their similarity w.r.t all 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 constraint is implemented by

$$\mathcal{L}_{SSL2} = \max \Big(CoSim(x_S^i, C_S^i) - CoSim(x_S^i, C_U^j), \sigma_2 \Big) \; \forall j \tag{6}$$

   $\sigma_2 = 0.25$ controls the minimum degree of semantic difference between different classes.

The final loss term for clustering the Unseen class samples is $\mathcal{L}_{Unseen} = \mathcal{L}(x_s, x_t) + \lambda_1 L_{SSL1} - \lambda_2 L_{SSL2}$, where $\mathcal{L}(x_s, x_t)$ is defined in Eqn. 3. $\lambda_1 = 1.1, \lambda_2 = 0.7$ are the weights. The '$-\lambda_2 L_{SSL2}$' ensures that the loss term does not increase arbitrarily which is possible for '$+\lambda_2 L_{SSL2}$'.

3.3. Feature Generation Network

Given the training images of Seen classes and unlabeled images of the Unseen classes we learn a generator $G : \mathcal{E}, \mathcal{Z} \longrightarrow \mathcal{X}$, 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 $\tilde{x} \in \mathcal{X}$. The discriminator $D : \mathcal{X}, \mathcal{E} \rightarrow [0, 1]$ 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]:

$$\begin{aligned} \mathcal{L}_{WGAN} = \min_G \max_D \mathbb{E}[D(x, e^y)] - \mathbb{E}[D(\tilde{x}, e^y)] \\ -\lambda \mathbb{E}[(\|\nabla_{\tilde{x}} D(\tilde{x}, e^y)\|_2 - 1)^2] \end{aligned} \tag{7}$$

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.

3.3.1 Self Supervised Loss From Anchor Vectors

The discriminator D is a classifier that determines whether the generated feature vector $\tilde{x}$ 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 distance between the generated vector $\tilde{x}$ and the anchor vector corresponding to the desired class y, i.e.

$$\mathcal{L}_{SSL3} = 1 - CoSim(\tilde{x}, c_y) \tag{8}$$

If $\tilde{x}$ truly represents the desired class y then the cosine similarity between $\tilde{x}$ 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.

3.3.2 Classifier Loss

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

$$\mathcal{L}_{CL} = -\mathbb{E}_{(\tilde{x}^s, y^s) \sim P_{\tilde{x}^s}} [\log P(y^s | \tilde{x}^s, \theta_{CL})] \tag{9}$$

where $P(y^s | \tilde{x}^s, \theta_{CL})$ is the classification probability and θ_{CL} denotes fixed parameters of the pre-trained classifier.

3.4. Training and Implementation

The final loss function is defined as

$$\mathcal{L} = \mathcal{L}_{WGAN} + \lambda_{CL}\mathcal{L}_{CL} + \lambda_3 L_{SSL3} \tag{10}$$

where λ_{CL}, λ_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 $\tilde{x}^u$ and real and synthetic features of seen class $x^s, \tilde{x}^s$ to construct the training set. Then we train a softmax classifier by minimizing the negative log likelihood loss:

$$\min_{\theta} -\frac{1}{|\mathcal{X}|} \sum_{(x,y)\in(\mathcal{X},\mathcal{Y})} \log P(y|x,\theta), \tag{11}$$

where $P(y|x,\theta) = \frac{\exp(\theta_y^T x)}{\sum_{j=1}^{|\mathcal{Y}|}\exp(\theta_y^T x)}$ is the classification probability and θ denotes classifier parameters. The final class prediction is by $f(x) = \arg\max_y P(y|x,\theta)$

Implementation Details: We show results for natural and medical images and compare with existing GZSL methods. Extending our method to natural images is straightforward where in we replace the class label vector e^y with the corresponding class attribute vectors. For feature extraction, similar to [49], we use a pre-trained ResNet-101 to extract 2048 dimensional CNN features for natural images. 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 choose Adam [22] as our optimizer, and the momentum is 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.

3.5. 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 S_{Te}. 2) **U**: training is performed on synthesized samples of $S + U$ classes and test on U. We also report the harmonic mean defined as

$$H = \frac{2 \times Acc_U \times Acc_S}{Acc_U + Acc_S} \tag{12}$$

where Acc_S and Acc_U denote the accuracy of images from seen (setting S) and unseen (setting U) classes respectively:

4. Experimental Results

4.1. Dataset Description

We demonstrate our method's effectiveness on natural images and the following medical imaging datasets for classification tasks. Datasets with a minimum of 3 disease classes (excluding normal label) were chosen to highlight the performance of feature synthesis.

1. **CAMELYON17** dataset [7]: contains 1000 whole slide images (WSIs) with 5 slides per patient: 500 slides for training and 500 slides for test. 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 224×224 patches from the different slides and obtain $130,000$ tumor patches and $200,000$ normal patches. We take a pre-trained ResNet101 and finetune 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.

2. **NIH Chest Xray** Dataset: For lung disease classification we adopted the NIH ChestXray14 dataset [46] 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 finetuned using the CheXpert dataset [18] and the chosen baseline FSL was from [39].

3. **CheXpert** Dataset: We used the CheXpert dataset [18] 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 [46] and the baseline FSL method was of [38] which is ranked second for the dataset with shared code.

4. **Kaggle Diabetic Retinopathy** dataset: has approximately $35,000$ images in the provided training set [19]. 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 [42] which has

[1] https://grand-challenge-public.s3.amazonaws.com/evaluation-supplementary/80/46fc579c-51f0-40c4-bd1a-7c28e8033f33/Camelyon17˙.pdf

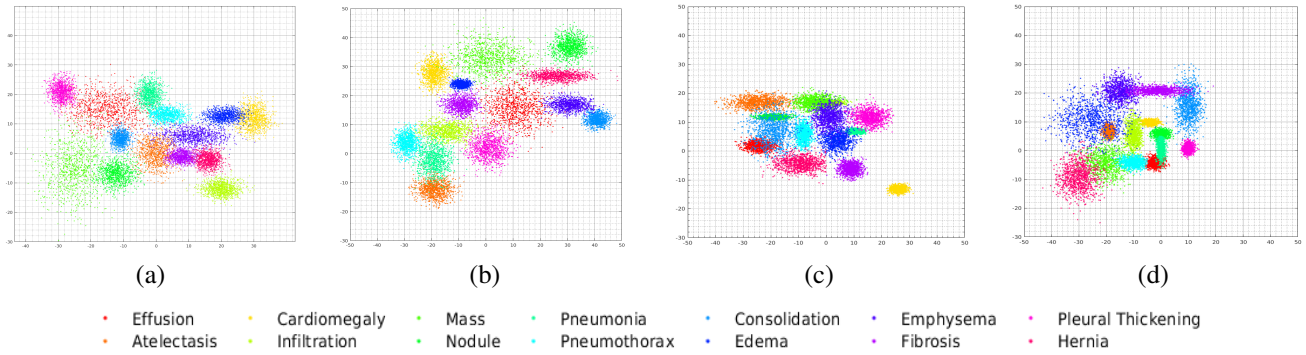


Figure 2. Feature visualizations for NIH ChestXray Dataset: (a) Seen+Unseen classes from actual dataset; distribution of synthetic samples generated by (b) SC-GZSL; (c) SC-GZSL$_{w\mathcal{L}_{SSL3}}$; (d) SDGN [47]. Different colours represent different classes. (b) is closer to (a), while (c) and (d) are quite different.

9939 color fundus images (2720×2720) 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 224×224.

5. **Gleason grading challenge** dataset [2] for prostate cancer (PCA) [20]. 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. Although a much larger dataset for PCA using WSIs is available[3], the data cannot be used for external submissions[4]. The baseline FSL was the classification outcome of the top ranked method[5]. The feature extractor was a pre-trained ResNet101 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 224×224 patches. The individual labels patches from normal images were all 'normal'. 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.

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 Xray, CheXpert and Kaggle DR datasets.

[2]https://gleason2019.grand-challenge.org/Home

[3]https://www.kaggle.com/c/prostate-cancer-grade-assessment/overview

[4]https://www.kaggle.com/c/prostate-cancer-grade-assessment/discussion/201117

[5]https://github.com/hubutui/Gleason

For natural images we use the following five datasets: 1) CUB [45], AwA1 [24], AwA2 [49], SUN [36], and FLO [35]. CUB includes 11K images and 200 species of birds labeled with 312-D attributes. AwA1 and AwA2 consist of 50 kinds of animals described by 85- D attributes, containing 30K and 37K images respectively. SUN is a large-scale scene attribute dataset, including 717 classes and 14K images with 102-D attributes. FLO con- sists of 8K images from 102 flower classes. Adapting our method to natural images is done by replacing the class vector (e) with the class attribute vector.

4.2. Baseline Methods

We compare our method's performance with the following GZSL methods employing different feature generation approaches such as CVAE or GANs: 1) CVAE based generation method of [17]; 2) over complete distribution (OCD) method of [21]; 3) self-supervised learning GZSL method of [47]; 4) FSL- Top performing FSL methods of corresponding datasets. Following GZSL protocol we report performance for Seen and Unseen classes. Our method is denoted as SC-GZSL (**S**elfsupervised **C**lustering based **GZSL**).

4.3. Visualization of Synthetic Image Features

Figure 2 (a) shows t-SNE plot of features from actual data from the NIH chest Xray 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. Figure 2 (c) shows the feature distribution for our method without using self-supervision. The resulting distribution is compact without overlap between classes, which is not representative of the real-world case. Clas-

sifiers trained on such distributions perform poorly on unseen classes. Figure 2 (d) shows the feature distributions using SDGN [47]. Although it also uses SSL the resulting feature representation is less accurate than our proposed method which contributes to the corresponding inferior performance.

4.4. Generalized Zero Shot Learning Results

Table 1 summarizes the results of our algorithm on natural images. The best performing method amongst all competing methods is SDGN [47]. However we are able to outperform it despite using a much simpler architecture. A McNemar's statistical test [32] shows that the results between SC-GZSL and SDGN is not very significant ($p = 0.062$), except for the SUN dataset (p=0.01). This dataset is particularly challenging as demonstrated by the fact that accuracy values are lower than other datasets.

The results for medical images shown in Table 2 shows our proposed method outperforms all competing GZSL methods including SDGN. 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. On the other hand medical images require a different approach.

4.5. Ablation Studies

Table 2 also shows results for the following ablation studies: 1) SCGZSL$_{w\mathcal{L}_{SSL1}}$- SCGZSL without the loss term $\mathcal{L}_{SSL1}$ (Eqn.5) for obtaining the anchor vectors, 2) SCGZSL$_{w\mathcal{L}_{SSL2}}$ - SCGZSL without the loss term $\mathcal{L}_{SSL2}$ (Eqn.6) to get anchor vectors, 3) SCGZSL$_{\mathcal{L}}$- Using only the baseline loss term $\mathcal{L}(z_s, z_t)$ (Eqn.3) for clustering all seen and unseen classes together, and no $\mathcal{L}_{SSL3}$ for feature synthesis; 4) SCGZSL$_{w\mathcal{L}_{SSL3}}$- SCGZSL without the loss term $\mathcal{L}_{SSL3}$ (Eqn.8) for training the feature synthesis network; 5) SCGZSL$-$only$\mathcal{L}_{SSL3}$- SCGZSL using only $\mathcal{L}_{SSL3}$ for feature synthesis without $\mathcal{L}_{SSL1}$, $\mathcal{L}_{SSL2}$.

The first three ablation studies investigate the effect of clustering on the final classification results. Their significant performance degradation compared to SCGZSL indicates the importance of our novel SSL based terms ($\mathcal{L}_{SSL1}, \mathcal{L}_{SSL2}$) in obtaining accurate anchor vectors. The baseline method, SCGZSL$_{\mathcal{L}}$, does not use any form of self supervision and has lowest H values. Compared to SCGZSL, we observe that excluding $\mathcal{L}_{SSL3}$ (SCGZSL$_{w\mathcal{L}_{SSL3}}$) leads to maximum reduction of H (more than 3.5%) across all datasets . This indicates that $\mathcal{L}_{SSL3}$ makes the most significant contribution to our method's performance. The use of anchor vectors makes it easier to synthesize features of unseen classes.

The influence of $\mathcal{L}_{SSL1}, \mathcal{L}_{SSL2}$ is quantitatively similar as shown by similar H values of SCGZSL$_{w\mathcal{L}_{SSL1}}$, SCGZSL$_{w\mathcal{L}_{SSL2}}$ 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 $\mathcal{L}_{SSL1}$ and $\mathcal{L}_{SSL2}$ sigificantly improves clustering accuracy. Excluding both $\mathcal{L}_{SSL1}, \mathcal{L}_{SSL2}$ and using the baseline SwAV ('only $w\mathcal{L}_{SSL3}$') gives significantly reduced H values for the different datasets despite using $\mathcal{L}_{SSL3}$ for feature synthesis. This clearly indicates the importance of having accurate anchor vectors for our method. SCGZSL$_{\mathcal{L}}$ can be considered as the most basic method without using any of our proposed novel loss terms, and unsurprisingly gives the worst results.

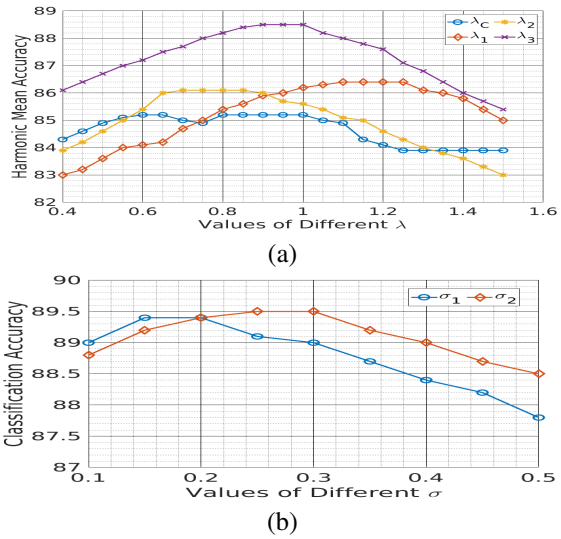


Figure 3. Hyperparameter Plots showing the value of H and classification accuracy for different values of;(a) λ; (b) σ.

4.6. Hyperparameter Selection

For all the competing methods in the case of medical images 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.

Figure 3 (a) shows the harmonic mean values for the NIH Chest Xray dataset for different values of hyperparameters $\lambda_1, \lambda_2, \lambda_3$, while Figure 3 (b) shows the corresponding plots for different values of σ_1, σ_2. 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 was monitored. We start with the base cost function of Eqn. 7, and first select the optimum value of λ_1. λ_1 values is fixed and we then deter-

Method	Natural Images														
	CUB			AwA1			AwA2			SUN			FLO		
	S	U	H	S	U	H	S	U	H	S	U	H	S	U	H
f-Vaegan [51]	65.1	61.4	63.2	-	-	-	88.6	84.8	86.7	41.9	60.6	49.6	87.2	78.7	82.7
GXE [25]	68.7	57.0	62.3	89.0	87.7	88.4	90.0	80.2	84.8	58.1	45.4	51.0	-	-	-
SDGN [47]	70.2	69.9	70.1	88.1	87.3	87.7	89.3	88.8	89.1	46.0	62.0	52.8	91.4	78.3	84.4
GDAN [17]	66.7	39.3	49.5	-	-	-	67.5	32.1	43.5	89.9	38.1	53.4	-	-	-
OCD[21]	59.9	44.8	51.3	-	-	-	73.4	59.5	65.7	42.9	44.8	43.8	-	-	-
SCGZSL	71.7	70.6	71.1	88.5	88.1	88.3	89.9	89.3	89.6	50.3	62.1	55.6	91.8	79.4	85.2

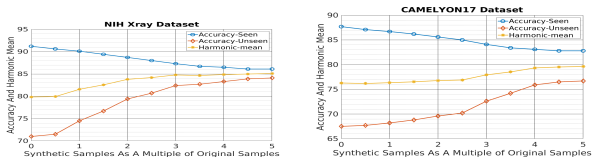
Table 1. **GZSL Results For Natural Images:**Average per-class classification accuracy (%) and harmonic mean (H) accuracy of generalized zero-shot learning when test samples are from Seen(S) or Unseen (U) classes. Numbers for competing methods are taken from [47]. S, U denote Acc_S, Acc_U.

Method	Multiple Medical Image Datasets														
	CAMELYON17			NIH Xray			CheXpert			Kaggle DR			Gleason		
	S	U	H	S	U	H	S	U	H	S	U	H	S	U	H
f-VAEGAN [51]	90.2	88.2	89.2	82.9	80.0	81.4	88.5	87.6	88.0	92.8	90.2	91.5	88.2	85.1	86.6
GDAN [17]	91.1	89.1	90.1	83.8	80.9	82.3	89.2	88.0	88.6	94.2	91.0	92.6	88.8	86	87.4
OCD[21]	91.5	89.3	90.4	84.7	81.3	83.0	89.9	88.1	89.0	94.8	91.3	93.0	89.2	86.9	88
SDGN [47]	92.1	89.5	90.8	84.4	81.1	82.7	90.2	88.2	89.2	95.0	91.9	93.4	90.0	87.8	88.9
SCGZSL	93.5	91.1	92.3	87.2	84.3	85.7	91.8	89.4	90.6	96.1	93.2	94.7	92.1	89.5	90.8
FSL	93.7	93.5	93.6	87.4	86.9	87.1	92.1	92.5	92.3	96.4	96.1	96.2	92.4	92.2	92.3
SCGZSL	**Ablation Studies**														
$w\mathcal{L}_{SSL1}$	91.2	88.7	89.9	84.5	82.1	83.3	89.1	86.9	88.0	92.2	89.6	90.9	90.3	86.9	88.6
$w\mathcal{L}_{SSL2}$	90.8	88.1	89.4	84.0	82.2	83.1	88.8	86.2	87.5	91.8	88.2	90.0	89.2	86	87.6
$w\mathcal{L}_{SSL3}$	90.0	87.0	88.5	83.2	81.0	82.1	87.6	85.1	86.3	90.1	86.7	88.4	88.4	85.5	86.9
only $\mathcal{L}_{SSL3}$	89.3	86.4	87.8	82.6	80.7	81.6	87.0	84.5	85.7	88.9	85.9	87.4	87.7	84.9	86.3
$\mathcal{L}$	87.2	84.1	85.6	80.7	79.1	79.7	84.6	82.7	83.6	86.5	83.7	85.1	86.1	82.8	84.4

Table 2. **GZSL Results For Medical Images:**Average per-class classification accuracy (%) and harmonic mean accuracy of generalized zero-shot learning when test samples are from Seen (Setting S) or unseen (Setting U) classes. Results of ablation studies are also shown.

mine λ_2, and then λ_3 by fixing λ_1, λ_2. 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 σ's were varied between $[0.1, 0.5]$ in steps of 0.05, and the resulting classification accuracy of the Xray images was determined. i.e., whether they were assigned to the correct cluster (class).

Figure 4 shows, for the NIH Chest Xray and CAMELYON17 dataset, the effect of adding synthetic samples on Acc_S, Acc_U as a function of dataset augmentation factor. Increasing synthesized examples increases Acc_U at a high rate while reducing Acc_S, although at a lower rate. TSynthetic samples improve discriminative power of classifiers and reduce bias towards Seen classes.



(a) (b)

Figure 4. Value of accuracy and H when adding synthetic samples to the dataset: (a) NIH dataset; (b) CAMELYON17 dataset.

5. 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 the feature generator. Using self supervision allows us to bridge the semantic gap between Seen and Unseen classes. The distribution of synthetic features generated by our method are close to the actual distribution, while removing the self-supervised term results in unrealistic distributions. Experimental results show our method outperforms other GZSL approaches in literature.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *In Proc. IEEE CVPR*, pages 2927–2936, 2015. 2

[2] Teresa Araujo, Guilherme Aresta, Luís Mendonca, Susana Penas, Carolina Maia, Angela Carneiro, Ana Maria Mendonca, and Aurelio Campilho. DR—GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 2020. 6

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *arXiv preprint arXiv:1701.07875*, 2017. 4

[4] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[5] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. van der Laak, , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 2017. 5, 6

[6] B. Bozorgtabar, D. Mahapatra, J.-P. Thiran, and L. Shao. SALAD: Self-supervised aggregation learning for anomaly detection on x-rays. In *In Proc. MICCAI*, pages 468–478, 2020. 2

[7] P. Bándi, , and et al. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Trans. Med. Imag.*, 38(2):550–560, 2019. 5

[8] Gabriele Campanella, Vitor M.K. Silva, and Thomas J. Fuchs. Terabyte-scale deep multiple instance learning for classification and localization in pathology. In *arXiv preprint arXiv:1805.06983*, 2018. 6

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018. 3

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand 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. 2, 3

[11] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5327–5336, 2016. 1

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *arXiv preprint arXiv:2002.05709*, 2020. 2

[13] 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. 2

[14] Rafael Felix, Vijay Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018. 1, 2

[15] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale 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. 1

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, pages 9729–9738, 2020. 2

[17] He Huang, Changhu Wang, Philip S. Yu, and Chang-Dong 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. 2, 6, 8

[18] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *arXiv preprint arXiv:1901.07031*, 2017. 1, 5

[19] Kaggle and EyePacs. Kaggle diabetic retinopathy detection. https://www.kaggle.com/c/diabetic-retinopathy-detection/data, jul 2015. 5

[20] D. Karimi, G. Nir, L. Fazli, P.C. Black, L. Goldenberg, and S.E. Salcudean. Deep learning-based gleason grading of prostate cancer from histopathology images-role of multi-scale decision aggregation and data augmentation. *IEEE J Biomed Health Inform.*, 24(5):1413–1426, 2020. 6

[21] Rohit Keshari, Richa Singh, and Mayank 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. 2, 6, 8

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980,*, 2014. 5

[23] Avinash Kori and Ganapathi Krishnamurthi. Zero shot learning for multi-modal real time image registration. In *arXiv preprint arXiv:1908.06213*, 2019. 2

[24] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Analysis Machine Intelligence*, 36(3):453–465, 2013. 2, 6

[25] Kai Li, Martin Renqiang Min, and Yun Fu. Rethinking zero-shot learning: A conditional visual classification perspective. In *CVPR*, pages 3583–3592, 2019. 8

[26] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venu-

gopalan, Aleksei Timofeev, Philip Q. Nelson, Greg S. Corrado, Jason D. Hipp, Lily Peng, and Martin C. Stumpe. Detecting cancer metastases on gigapixel pathology images. In *arXiv preprint arXiv:1703.02442*, 2017. 1

[27] Ming Y. Lu, Richard J. Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. In *arXiv:1910.10825*, 2019. 2

[28] D. Mahapatra, B. Bozorgtabar, and L. Shao. Pathological retinal region segmentation from oct images using geometric relation based augmentation. In *In Proc. IEEE CVPR*, pages 9611–9620, 2020. 2

[29] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and L. Shao. Structure preserving stain normalization of histopathology images using self supervised semantic guidance. In *In Proc. MICCAI*, pages 309–319, 2020. 2

[30] Dwarikanath Mahapatra, Alexander Poellinger, Ling Shao, and Mauricio Reyes. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. *IEEE TMI*, pages 1–15, 2021. 2

[31] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9977–9985, 2019. 1

[32] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. 7

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *In Proc. ICLR Workshops*, 2013. 2

[34] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong 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. 2

[35] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008. 6

[36] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 6

[37] Akanksha Paul, Narayanan C Krishnan, and Prateek Munjal. Semantically aligned bias reducing zero shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7056–7065, 2019. 1

[38] Hieu H. Pham, Tung T. Le, Dat Q. Tran, Dat 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. 5

[39] 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. 5

[40] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. In *In Proc. NeurIPS*, 2020. 2

[41] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *In Proc. IEEE CVPR*, pages 8247–8255, 2019. 2

[42] Hidenori Takahashi, Hironobu Tampo, Yusuke Arai, Yuji Inoue, and Hidetoshi Kawashima. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. *Plos One*, 12(6), 2017. 5

[43] J. Tong, D. Mahapatra, P. Bonnington, T. Drummond, and Z. Ge. Registration of histopathology images using self supervised fine grained feature maps. In *In Proc. MICCAI-DART Workshop*, pages 41–51, 2020. 2

[44] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4281–4289, 2018. 1, 2

[45] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[46] 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. 1, 5

[47] Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, and Feng 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. 2, 6, 7, 8

[48] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3

[49] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep 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. 1, 2, 5, 6

[50] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *In Proc. IEEE CVPR*, pages 5542–5551, 2018. 1, 2

[51] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep 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. 1, 2, 8