1-28-2021

Generative multi-label zero-shot learning

Akshita Gupta
Inception Institute of Artificial Intelligence

Sanath Narayan
Inception Institute of Artificial Intelligence

Salman Khan
Mohamed bin Zayed University of Artificial Intelligence

Fahad Shahbaz Khan
Mohamed bin Zayed University of Artificial Intelligence

Ling Shao
Inception Institute of Artificial Intelligence

See next page for additional authors

Recommended Citation

A. Gupta, S. Narayan, S. Khan, F.S. Khan, L. Shao, and J. van de Weijer, "Generative multi-label zero-shot learning", 2021, arXiv.2101.11606v2

Authors

Akshita Gupta, Sanath Narayan, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Joost van de Weijer

Generative Multi-Label Zero-Shot Learning

Akshita Gupta*[1]   Sanath Narayan*[1]   Salman Khan[2]   Fahad Shahbaz Khan[2]

Ling Shao[1]   Joost van de Weijer[3]

[1]Inception Institute of Artificial Intelligence, UAE   [2]Mohamed Bin Zayed University of AI, UAE

[3] Universitat Autònoma de Barcelona

Abstract

Multi-label zero-shot learning strives to classify images into multiple unseen categories for which no data is available during training. The test samples can additionally contain seen categories in the generalized variant. Existing approaches rely on learning either shared or label-specific attention from the seen classes. Nevertheless, computing reliable attention maps for unseen classes during inference in a multi-label setting is still a challenge. In contrast, state-of-the-art single-label generative adversarial network (GAN) based approaches learn to directly synthesize the class-specific visual features from the corresponding class attribute embeddings. However, synthesizing multi-label features from GANs is still unexplored in the context of zero-shot setting. In this work, we introduce different fusion approaches at the attribute-level, feature-level and cross-level (across attribute and feature-levels) for synthesizing multi-label features from their corresponding multi-label class embeddings. To the best of our knowledge, our work is the first to tackle the problem of multi-label feature synthesis in the (generalized) zero-shot setting. Comprehensive experiments are performed on three zero-shot image classification benchmarks: NUS-WIDE, Open Images and MS COCO. Our cross-level fusion-based generative approach outperforms the state-of-the-art on all three datasets. Furthermore, we show the generalization capabilities of our fusion approach in the zero-shot detection task on MS COCO, achieving favorable performance against existing methods. The source code is available at https://github.com/akshitac8/Generative_MLZSL.

1. Introduction

Multi-label classification is a challenging problem where the task is to recognize all labels in an image. Typical examples of multi-label classification include, MS COCO [22] and NUS-WIDE [5] datasets, where an image may contain several different categories (labels). Most recent multi-label classification approaches address the problem by utilizing attention mechanisms [38, 45, 47], recurrent neural networks [37, 44, 25], graph CNNs [18, 4] and label correlations [40, 6]. However, these approaches do not tackle the problem of multi-label zero-shot classification, where the task is to classify images into multiple new "unseen" categories at test time, without being given any corresponding visual example during the training. Different from zero-shot learning (ZSL), the test samples can belong to the seen or unseen classes in generalized zero-shot learning (GZSL). Here, we tackle the challenging problem of large-scale multi-label ZSL and GZSL.

Existing multi-label (G)ZSL approaches address the problem by utilizing global image representations [24, 48], structured knowledge graphs [20] and attention-based mechanisms [14]. In contrast to the multi-label setting, single-label (generalized) zero-shot learning, where an image contains at most one category label, has received significant attention [15, 9, 8, 33, 32, 46, 1, 41, 42, 43]. State-of-the-art single-label (G)ZSL approaches [42, 7, 21, 13, 23, 43, 26] are generative in nature. These approaches exploit the power of generative models, such as generative adversarial networks (GANs) [11] and variational autoencoder (VAE) [17] to synthesize unseen class features. Typically, a feature synthesizing generator is utilized to construct single-label features. The generative approaches currently dominate single-label ZSL due to their ability to synthesize unseen class (fake) features by learning the underlying data distribution of seen classes (real) features. Nevertheless, the generator only synthesizes single-label features in the existing ZSL frameworks. To the best of our knowledge, the problem of designing a feature synthesizing generator for *multi-label* ZSL paradigm is yet to be explored.

In this work, we address the problem of multi-label (generalized) zero-shot learning by introducing an approach based on the generative paradigm. When designing a generative multi-label zero-shot approach, the main objective is to synthesize semantically consistent *multi-label* visual features from their corresponding class attributes (embeddings). Multi-label visual features can be synthesized in two ways. (i) One approach is to integrate class-specific attribute embeddings at the input of the generator to produce a global image-level embedding vector. We call this approach
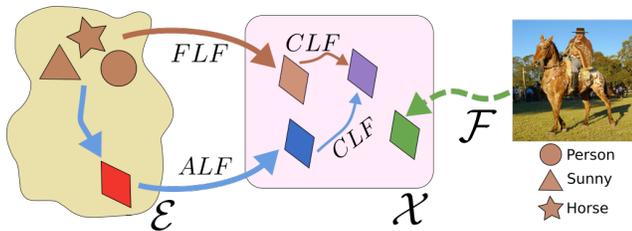
Figure 1. A conceptual illustration of our three fusion approaches: attribute-level (ALF), feature-level (FLF) and cross-level feature (CLF), on an example image with three classes. Given an image, a feature extractor $\mathcal{F}$ extracts real visual features (green diamond) in $\mathcal{X}$. ALF integrates the individual class-specific embeddings to generate a global image-level embedding (red diamond) in $\mathcal{E}$. The image-level embedding is then used to synthesize the multi-label features (blue diamond). Different from ALF, the FLF synthesizes class-specific features and integrates them in the feature space $\mathcal{X}$. These integrated features are shown as the yellow diamond. Our CLF takes as input synthesized features from ALF and FLF and enriches each respective feature by taking guidance from the other. The resulting enriched features are integrated to obtain the final image-level feature representation (purple diamond). The final image representation combines the advantage of label dependency of ALF and class-specific discriminability of FLF.

attribute-level fusion (ALF). Here, the image-level embedding represents the holistic distribution of the positive labels in the image (see Fig. 1). Since the generator in ALF performs global image-level feature generation, it is able to better capture label dependencies (correlations among the labels) in an image. However, such a feature generation has lower class-specific discriminability since the discriminative information with respect to the individual classes is not explicitly encoded. (ii) A second approach is to synthesize the features from the class-specific embeddings individually and then integrate them in the visual feature space. We call this approach feature-level fusion (FLF), as shown in Fig. 1. Although FLF better preserves the class-specific discriminative information in the synthesized features, it does not explicitly encode the label dependencies in an image due to synthesizing features independently of each other. This motivates us to investigate an alternative fusion approach for multi-label feature synthesis.

In this work, we introduce an alternative fusion approach that combines the advantage of label dependency of ALF and the class-specific discriminability of FLF, during multi-label feature synthesis. We call this approach as cross-level feature fusion (CLF), as in Fig. 1. The CLF approach utilizes each individual-level feature and attends to the bi-level context (from ALF and FLF). As a result, individual-level features adapt themselves to produce enriched synthesized features, which are then pooled to obtain the CLF output. In addition to multi-label zero-shot classification, we investigate the proposed multi-label feature generation CLF approach for (generalized) zero-shot object detection.

Contributions: We propose a generative approach for multi-label (generalized) zero-shot learning. To the best of our knowledge, we are the first to explore the problem of multi-label feature synthesis in the zero-shot setting. We investigate three different fusion approaches (ALF, FLF and CLF) to synthesize multi-label features. Our CLF approach combines the advantage of label dependency of ALF and the class-specific discriminability of FLF. Further, we integrate our fusion approaches into two representative generative architectures: f-CLSWGAN [42] and f-VAEGAN [43].

We evaluate our (generalized) zero-shot classification approach on three datasets: NUS-WIDE [5], Open Images [19] and MS COCO [22]. Our CLF approach achieves consistent improvement in performance over both ALF and FLF. Furthermore, CLF outperforms state-of-the-art methods on *all* datasets. On the large-scale Open Images dataset, our CLF achieves absolute gains of 18.7% and 17.6% over the state-of-the-art in terms of GZSL F1 score at *top-K* predictions, $K \in \{10, 20\}$, respectively. In addition to classification, we evaluate CLF for (generalized) zero-shot object detection, achieving favorable results against existing methods on MS COCO.

2. Generative Single-Label Zero-Shot Learning

As discussed earlier, state-of-the-art single-label zero-shot approaches [42, 7, 21, 13, 23, 43] are generative in nature, utilizing the power of generative models (*e.g.*, GAN [11], VAE [17]) to synthesize unseen class features. Here, each image is assumed to have a *single* object category label (*e.g.*, CUB [39], FLO [27] and AWA [41]).

Problem Formulation: Let $x \in \mathcal{X}$ denote the encoded feature instances of images and $y \in \mathcal{Y}^s$ the corresponding class labels from the set of S seen class labels $\mathcal{Y}^s = \{y_1, \ldots, y_S\}$. Let $\mathcal{Y}^u = \{y_{S+1}, \ldots, y_C\}$ denote the set of U unseen classes, which is disjoint from the seen classes $\mathcal{Y}^s$. Here, the total number of seen and unseen classes is denoted by $C{=}S{+}U$. The relationships among all the seen and unseen classes are described by the category-specific semantic embeddings $e(k) \in \mathcal{E}, \forall k \in \mathcal{Y}^s \cup \mathcal{Y}^u$. To learn the ZSL and GZSL classifiers, existing single-label GAN-based approaches [42, 21, 13, 43] first learn a generator using the seen class features x_s and corresponding class embeddings $e(y_s)$. Then, the learned generator and the unseen class embeddings $e(y_u)$ are used to synthesize the unseen class features $\tilde{x}_u$. The resulting synthesized features $\tilde{x}_u$, along with the real seen class features x_s, are further deployed to train the final classifiers $f_z : \mathcal{X} {\rightarrow} \mathcal{Y}^U$ and $f_{gz} : \mathcal{X} {\rightarrow} \mathcal{Y}^C$. Next, we briefly describe the feature synthesis stage employed in existing single-label GAN-based zero-shot frameworks, to generate unseen class features.

Typically, GAN-based zero-shot classification frameworks utilize a feature synthesizing generator $G : \mathcal{Z} {\times} \mathcal{E} {\rightarrow} \mathcal{X}$ and a discriminator D. Both G and D compete against

each other in a two player minimax game. While D attempts to accurately distinguish real image features x from generated features $\tilde{x}$, G attempts to fool D by generating features that are semantically close to real features. Since class-specific features are to be synthesized, a conditional Wasserstein GAN [2] is employed due to its more stable training, by conditioning both G and D on the embeddings $e(y)$. Here, G learns to synthesize class-specific features $\tilde{x}$ from the corresponding single-label embeddings $e(y)$, given by $\tilde{x} = G(z, e(y))$. Nevertheless, the generator only synthesizes single-label features in existing zero-shot learning frameworks. To the best of our knowledge, the problem of designing a feature synthesizing generator for the multi-label zero-shot learning paradigm is yet to be investigated.

3. Generative Multi-Label Zero-Shot Learning

As discussed earlier, most real-world tasks involve multi-label recognition, where an image can contain multiple and wide range of category labels (*e.g*., MS COCO [22], NUS-WIDE [5] and Open Images [19]). The multi-label classification problem becomes more challenging in the zero-shot learning setting, where the test set either contains only unseen classes (ZSL) or both seen and unseen classes (GZSL). In this work, we propose a generative multi-label zero-shot learning approach that exploits the capabilities of generative models to learn the underlying data distribution of seen classes. This helps to mimic the fully-supervised setting by synthesizing (fake) features for unseen classes. While generative approaches have been extensively studied for single-label zero-shot learning, we are the first to address the problem of multi-label feature synthesis in the multi-label (generalized) zero-shot setting. Next, we describe the problem formulation of generative multi-label (generalized) zero-shot learning.

Problem Formulation: In contrast to single-label zero-shot learning, here, $x \in \mathcal{X}$ denotes the encoded feature instances of multi-label images and $y \in \{0,1\}^S$ the corresponding multi-hot labels from the set of S seen class labels $\mathcal{Y}^s$. Let x denote the multi-label feature instance of an image and y, its multi-hot label with n positive classes in the image. Then, the set of attribute embeddings for the image can be denoted as $e(y) = \{e(y_j), \forall j : y[j] = 1\}$, where $|e(y)| = n$. Here, we use GloVe [29] vectors of the class names as the attribute embeddings $e(y_j)$, as in [14]. Now, the generator's task is to learn to synthesize the multi-label features x from the set of associated embedding vectors $e(y)$. Post-training of G, multi-label features corresponding to the unseen classes are synthesized. The resulting synthesized features along with the real seen class features are deployed to train the final ZSL/GZSL classifiers $f_z : \mathcal{X} \rightarrow \{0,1\}^U$ and $f_{gz} : \mathcal{X} \rightarrow \{0,1\}^C$. Next, we investigate three different approaches to synthesize unseen multi-label features.

3.1. Generative Multi-Label Feature Synthesis

To synthesize multi-label features, we introduce three fusion approaches: attribute-level fusion (ALF), feature-level fusion (FLF) and cross-level feature fusion (CLF).

3.1.1 Attribute-Level Fusion

In attribute-level fusion (ALF) approach, a global image-level embedding vector is obtained from a set of class-specific embedding vectors that correspond to multiple labels in the image. The image-level embedding represents the global description of the positive labels in an image. The global embedding e_μ is obtained by averaging the individual class embeddings $e(y)$. This embedding e_μ is then input to the generator G^a along with the noise z for synthesizing the feature $\tilde{x}^a$. The attribute-level fusion is then given by

$$ e_\mu = \frac{1}{n} \sum_{j:y[j]=1} e(y_j) \quad \text{and} \quad \tilde{x}^a = G^a(z, e_\mu). \quad (1) $$

Fig. 2 shows the feature generation of $\tilde{x}^a$ from the global embedding e_μ. The generator G^a in ALF performs global image-level feature generation, thereby capturing label dependencies (correlations among labels) in an image. However, such a feature generation from the global embedding $\tilde{x}^a$ has lower class-specific discriminability since it does not explicitly encode discriminative information with respect to individual classes.

3.1.2 Feature-Level Fusion

Here, we introduce a feature-level fusion (FLF) approach, which synthesizes the features from the class-specific embeddings individually. This allows the FLF approach to better preserve the class-specific discriminative information in the synthesized features. Different from the ALF that integrates the embedding vectors, the FLF approach first inputs the n class-specific embeddings $e(y_j)$ to G^f in order to generate n class-specific latent features $\tilde{x}_j$. These features are then integrated through an average operation to obtain the final synthesized feature $\tilde{x}^f$. The feature-level fusion (FLF) is denoted by

$$ \tilde{x}_j = G^f(z, e(y_j)) \quad \text{and} \quad \tilde{x}^f = \frac{1}{n} \sum_{j:y[j]=1} \tilde{x}_j. \quad (2) $$

Fig. 2 shows the feature generation of $\tilde{x}^f$ from the individual embeddings $e(y_j)$. We observe from Eq. 2 that for a fixed noise z, the generator G^f synthesizes a fixed latent feature $\tilde{x}_j$ for class j, regardless of the presence/absence of other classes in an image. Thus, while the generated latent features $\tilde{x}_j$ better preserve class-specific discriminative information of the positive classes j present in the image, G^f
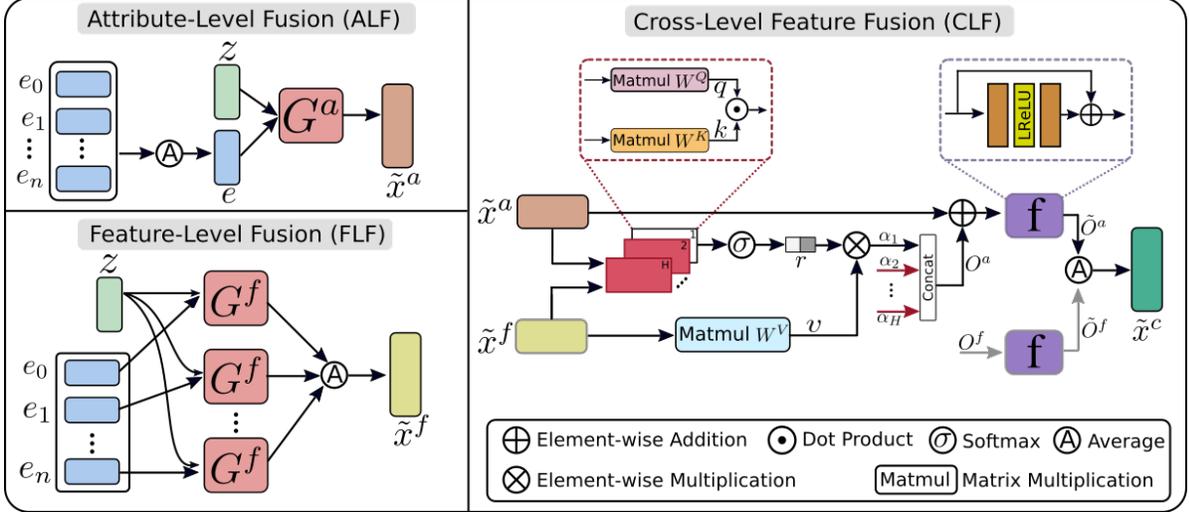
Figure 2. Overview of three different approaches to synthesize unseen multi-label features. The attribute-level fusion (ALF) generates a global image-level embedding vector from a set of class-specific embedding vectors corresponding to multiple labels in an image (Sec. 3.1.1). In ALF, the generator synthesizes global features $\tilde{x}^a$ that capture the correlations among the labels in the image. On the other hand, the feature-level fusion (FLF) synthesizes features from individual class-specific embeddings (Sec. 3.1.2). As a result, the generator produces class-specific latent features which are then integrated to obtain synthesized features $\tilde{x}^f$. The cross-level fusion (CLF) combines the advantages of ALF and FLF, during feature generation (Sec. 3.1.3). Specifically, it uses each individual-level feature ($\tilde{x}^a$, $\tilde{x}^f$) to attend to the bi-level context and adapt itself to generate $\tilde{o}^a$, $\tilde{o}^f$. These enriched features are then pooled to obtain the CLF output, which represents our final synthesized feature $\tilde{x}^c$.

synthesizes them independently of each other. As a result, the synthesized feature $\tilde{x}^f$ does not explicitly encode the label dependencies in an image.

As discussed above, both the aforementioned fusion approaches (ALF and FLF) have shortcomings when synthesizing multi-label features. Next, we introduce a fusion approach that combines the advantage of the label dependency of ALF and class-specific discriminability of FLF.

3.1.3 Cross-Level Feature Fusion

The proposed cross-level feature fusion (CLF) aims to combine the advantages of both ALF and FLF. The CLF approach (see Fig. 2) incorporates label dependency and class-specific discriminability in the feature generation stage as in the ALF and FLF, respectively. To this end, $\tilde{x}^f$ and $\tilde{x}^a$ are forwarded to a feature fusion block. Inspired by the multi-headed self-attention [35], the feature fusion block enriches each respective feature by taking guidance from the other branch. Specifically, we create a matrix $\tilde{x} \in \mathbb{R}^{2 \times d}$ by stacking the individual features $\tilde{x}^f$ and $\tilde{x}^a$. Then, these features are linearly projected to a low-dimensional space ($d' = {}^d\!/_H$) to create query-key-value triplets using a total of H projection heads,

$$q_h = \tilde{x}W_h^Q, k_h = \tilde{x}W_h^K, v_h = \tilde{x}W_h^V,$$

where $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d \times d'}$ and $h \in \{1, 2, .., H\}$. For each feature, a status of its current form is kept in the 'value'

embedding, while the *query* vector derived from each input feature is used to find its correlation with the *keys* obtained from both the features, as we elaborate below.

Given these triplets from each head, the features undergo two levels of processing (i) intra-head processing on the triplet and (ii) cross-head processing. For the first case, the following equation is used to relate each query vector with '*keys*' derived from both the features. The resulting normalized relation scores ($r_h \in \mathbb{R}^{2 \times 2}$) are used to reweight the corresponding value vectors, thereby obtaining the attended features $\alpha_h \in \mathbb{R}^{2 \times d'}$,

$$\alpha_h = r_h v_h, \quad \text{where} \ \ r_h = \sigma\left(\frac{q_h k_h^T}{\sqrt{d'}}\right).$$

To aggregate information across all heads, these attended low-dimensional features from each head are concatenated and processed by an output layer to generate the original d-dimensional output vectors $o_i \in \mathbb{R}^{2 \times d}$,

$$o_i = (\alpha_1 \oplus \alpha_2 \oplus \ldots \alpha_H)W^O, \tag{3}$$

where $W^O \in \mathbb{R}^{d \times d}$ is a learnable weight matrix and $\oplus$ denotes concatenation.

After obtaining the self-attended features o, a residual branch is added from the input to the attended features and further processed with a small residual sub-network $f(\cdot)$ to help the network first focus on the local neighbourhood and then progressively pay attention to the other-level features,

$$\tilde{o} = f(\tilde{x} + o) + (\tilde{x} + o), \quad \text{s.t.,} \ \tilde{o} \in \mathbb{R}^{2 \times d}. \tag{4}$$

This encourages the network to selectively focus on adding complimentary information to the source vectors $\tilde{x}$. Finally, we mean-pool the matrix $\tilde{o}$ along the row dimension to obtain a single cross-level fused feature $\tilde{x}^c \in \mathbb{R}^d$,

$$\tilde{x}^c = \frac{1}{2} \sum_{j \in \{1,2\}} \tilde{o}^{j,k}. \qquad (5)$$

The cross-level fused feature $\tilde{x}^c$ is obtained by effectively fusing the features generated from ALF and FLF. As a result, $\tilde{x}^c$ explicitly encodes the correlation among labels in the image in addition to the class-specific discriminative information of the positive classes present in the image. Next, we describe the integration of our CLF approach in two representative generative architectures for multi-label (generalized) zero-shot classification.

3.2. Multi-Label Zero-Shot Classification

We integrate our fusion approaches in two representative generative architectures (`f-CLSWGAN` [42] and `f-VAEGAN` [43]) for multi-label zero-shot classification. Both `f-CLSWGAN` and `f-VAEGAN` have been shown to achieve promising performance for single-label zero-shot classification. Since CLF integrates both ALF and FLF, we only describe here the integration of CLF in the two classification frameworks. Next, we describe the integration of CLF in `f-CLSWGAN`, followed by `f-VAEGAN`.

Briefly, `f-CLSWGAN` comprises a conditional WGAN (conditioned on the embeddings $e(y)$) and a seen class classifier f_{cls}. Here, we replace the standard generator of the `f-CLSWGAN` with our multi-label feature generation (CLF) to synthesize multi-label features $\tilde{x}^c$. The resulting multi-label WGAN loss is given by

$$\mathcal{L}_W = \mathbb{E}[D(x_s, e(y_s))] - \mathbb{E}[D(\tilde{x}_s^c, e(y_s))] - \qquad (6)$$
$$\lambda \mathbb{E}[(||\nabla_{\hat{x}} D(\hat{x}, e(y_s))||_2 - 1)^2],$$

where $\tilde{x}_s^c$ is synthesized using Eq. 5 for the seen classes, λ is the penalty coefficient and $\hat{x}$ is a convex combination of x_s and $\tilde{x}_s^c$. Furthermore, a classifier f_{cls}, trained on the seen classes, is employed to encourage the generator to synthesize features that are well suited for final ZSL/GZSL classification. The final objective for training our CLF-based `f-CLSWGAN` in a multi-label setting is given by

$$\min_G \max_D \mathcal{L}_W - \alpha \mathbb{E}[\text{BCE}(f_{cls}(\tilde{x}_s^c), y_s)], \qquad (7)$$

where $\text{BCE}(f_{cls}(\tilde{x}_s^c), y_s)$ denotes the standard binary cross entropy loss between the predicted multi-label $f_{cls}(\tilde{x}_s)$ and the ground-truth multi-label y_s of feature x_s.

In addition to `f-CLSWGAN`, we also integrate our CLF into `f-VAEGAN` [43] framework to perform multi-label feature synthesis. The `f-VAEGAN` extends `f-CLSWGAN` by combining a conditional VAE [17] along with a conditional
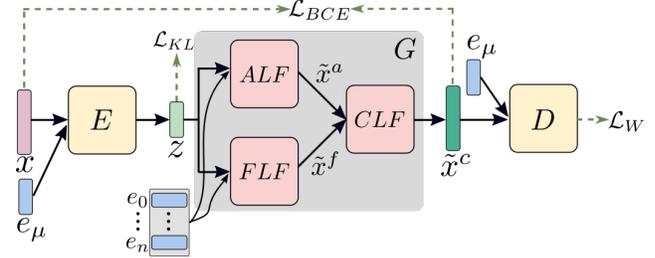

Figure 3. Our CLF-based `f-VAEGAN` architecture, which integrates the proposed multi-label feature synthesis CLF approach into the `f-VAEGAN`. The standard `f-VAEGAN` extends `f-CLSWGAN` by integrating a conditional VAE along with a conditional WGAN, utilizing a shared generator between them. `f-VAEGAN` comprises an encoder E, shared generator and a discriminator D, conditioned on the class-specific embeddings e. In the proposed CLF-based `f-VAEGAN` shown here, we replace the single-label shared generator in `f-VAEGAN` with our multi-label generator G for synthesizing multi-label $\tilde{x}^c$. The resulting synthesized features $\tilde{x}^c$ are then passed to the discriminator for training the generator using the loss terms: $\mathcal{L}_{KL}$, $\mathcal{L}_{BCE}$ and $\mathcal{L}_W$.

WGAN, utilizing a shared generator between them. For more details on `f-VAEGAN`, we refer to [43]. Similar to our CLF-based `f-CLSWGAN` described earlier, we replace the single-label shared generator in `f-VAEGAN` with our multi-label generator (CLF) for synthesizing multi-label $\tilde{x}^c$. The resulting CLF-based `f-VAEGAN` is trained similar to the standard `f-VAEGAN`, using the original loss formulation [43]. Fig. 3 shows our CLF-based `f-VAEGAN` architecture for multi-label (generalized) zero-shot classification.

4. Experiments

4.1. Experimental Setup

Datasets: We evaluate our generative multi-label zero-shot approach on three benchmarks: NUS-WIDE [5], Open Images [19] and MS COCO [22]. The **NUS-WIDE** dataset comprises nearly 270K images with 81 human-annotated categories, in addition to the 925 labels obtained from Flickr user tags. As in [14, 48], the 925 and 81 labels are used as seen and unseen classes, respectively. The **Open Images** (v4) is a large-scale dataset consisting of 9 million training images along with 41,620 and 125,456 validation and testing images, respectively. The scale of Open Images is larger than other multi-label datasets, such as NUS-WIDE and MS COCO. This dataset is partially annotated with human labels and machine-generated labels. Here, 7,186 labels, having at least 100 images in the training, are selected as seen classes. The most frequent 400 test labels, which are not present in the training data are selected as unseen classes, as in [14]. The **MS COCO** dataset is divided into training and validation sets with 82,783 and 40,504 images, respectively. It has been previously used for multi-label zero-shot object detection [3, 12] but not for multi-label zero-shot classifica-

tion. Here, we perform multi-label zero-shot classification experiments by using the same split (65 seen and 15 unseen classes), as in detection works [3, 12].

Evaluation Metrics: For evaluating our approach on multi-label (generalized) zero-shot classification, we use the mean Average Precision (mAP) and F1 score at *top-K* predictions, similar to [14, 36, 48]. While the mAP measure captures the image ranking accuracy of the model for each label, the F1 measure captures the model's ability to correctly rank the labels in each image.

Implementation Details: Following existing zero-shot classification works [48, 14], we use the pretrained VGG-19 backbone to extract features from multi-label images.

Image-level features of size $d=4,096$, corresponding to FC7 layer output are used as input to our GAN. We use the l_2 normalized 300 dimensional GloVe [29] vectors corresponding to the category names as the attribute embeddings $e(k) \in \mathcal{E}$, as in [14]. The encoder E, discriminator D and generators G^a and G^f are all two layer fully connected (FC) networks with 4,096 hidden units and have Leaky ReLU as the non-linearity. The number of heads H in our CLF is set to 8. The sub-network $f(\cdot)$ in CLF is a two layer FC network with 8,192 hidden units. The feature synthesizing network is trained with a learning rate of 10^{-4}. The WGAN is trained with (*batch size*, *epoch*) of $(64, 50)$, $(128, 1)$ and $(64, 70)$ on NUS-WIDE, Open Images and MS COCO, respectively. For the f-CLSWGAN variant, α is set to 0.1, while the coefficient λ is set to 10 in the f-VAEGAN variant. The ZSL and GZSL classifiers: f_z and f_{gz} are trained for 50 epochs with (*batch size*, *learning rate*) of $(300, 10^{-3})$, $(100, 10^{-4})$ and $(100, 10^{-3})$ on NUS-WIDE, Open Images and MS COCO, respectively. The WGAN and the classifiers are trained using ADAM with (β_1, β_2) as $(0.5, 0.999)$. All parameters are chosen via cross-validation.

4.2. Ablation Study

We first present an ablation study w.r.t. our fusion approaches: attribute-level (ALF), feature-level (FLF) and cross-level feature (CLF) on the NUS-WIDE dataset. We evaluate our fusion strategies with both architectures: f-CLSWGAN and f-VAEGAN. Tab. 1 shows the comparison, in terms of F1 score ($K \in \{3, 5\}$) and mAP for both ZSL and GZSL tasks. In the case of ZSL, the ALF-based f-CLSWGAN achieves mAP score of 22.1. The FLF-based f-CLSWGAN obtains similar performance with mAP score of 22.6. The CLF-based f-CLSWGAN achieves improved performance with mAP score of 23.7. Similarly, the CLF-based f-CLSWGAN obtains consistent improvement over both ALF and FLF-based f-CLSWGAN in terms of F1 score ($K \in \{3, 5\}$). In the case of GZSL, ALF and FLF obtain F1 scores ($K=3$) of 16.8 and 17.0, respectively. Our CLF approach achieves improved performance with F1 score of 17.9. Similarly, CLF performs favorably against both ALF

Table 1. Classification performance comparison of the three fusion approaches (ALF, FLF and CLF) for both ZSL and GZSL tasks on the NUS-WIDE dataset. The comparison is shown in terms of F1 score ($K \in \{3, 5\}$) and mAP. We present the evaluation of our fusion approaches with both architectures: f-CLSWGAN and f-VAEGAN. Regardless of the underlying generative architecture, our CLF approach achieves consistent improvement in performance, on all metrics, over both ALF and FLF for ZSL and GZSL tasks. Best results are in bold.

Fusion	GAN	Task	F1 (K=3)	F1 (K=5)	mAP
ALF	f-CLSWGAN	ZSL	29.8	26.7	22.1
		GZSL	16.8	19.7	7.1
	f-VAEGAN	ZSL	30.9	27.9	22.9
		GZSL	17.7	20.6	7.5
FLF	f-CLSWGAN	ZSL	29.8	26.8	22.6
		GZSL	17.0	19.6	7.2
	f-VAEGAN	ZSL	29.9	27.0	23.3
		GZSL	17.2	19.9	7.6
CLF	f-CLSWGAN	ZSL	31.1	27.6	23.7
		GZSL	17.9	20.7	8.0
	f-VAEGAN	ZSL	**32.8**	**29.3**	**25.7**
		GZSL	**18.9**	**22.0**	**8.9**

and FLF in terms of F1 at $K=5$ and mAP metrics.

As in f-CLSWGAN, we also observe the CLF approach to achieve consistent improvement in performance over both ALF and FLF, when integrated in the more sophisticated f-VAEGAN. In the case of ZSL, our CLF-based f-VAEGAN achieves absolute gains of 2.8% and 2.4% in terms of mAP over ALF and FLF, respectively. Similarly, CLF-based f-VAEGAN achieves consistent improvement in performance in terms of F1 score, over both ALF and FLF. Furthermore, CLF-based f-VAEGAN performs favorably against the other two fusion approaches in case of GZSL.

The aforementioned results show that our CLF approach achieves consistent improvement in performance over both ALF and FLF for both ZSL and GZSL, regardless of the underlying architecture. Furthermore, the best results are obtained when integrating our CLF in f-VAEGAN. Next, we compare our CLF-based f-VAEGAN, denoted as *our approach* hereafter, with state-of-the-art methods.

4.3. State-of-the-art Comparison

NUS-WIDE: Tab. 2 shows the state-of-the-art comparison for zero-shot (ZSL) and generalized zero-shot (GZSL) classification. The results are reported in terms of mAP and F1 score at *top-K* predictions ($K \in \{3, 5\}$). In addition, we also report the precision (P) and recall (R) for each F1 score. For the ZSL task, Fast0Tag [48] approach that finds principal directions in the word vector space for ranking the relevant tags ahead of the irrelevant tags, achieves mAP score of 15.1. The recently introduced LESA approach [14], which utilizes a shared multi-attention mechanism to predict all labels in an image, achieves improved performance over Fast0Tag [48], with mAP score of 19.4. LESA also ob-

Table 2. State-of-the-art comparison for ZSL and GZSL tasks on the NUS-WIDE and Open Images datasets. We report the results in terms of mAP and F1 score at $K\in\{3,5\}$ for NUS-WIDE and $K\in\{10,20\}$ for Open Images. Our approach outperforms the state-of-the-art for both ZSL and GZSL tasks, in terms of mAP and F1 score, on both datasets. Best results are in bold.

Method	Task	NUS-WIDE (#seen / #unseen = 925/81)							Open-Images (#seen / #unseen = 7186/400)						
		K = 3			K = 5			mAP	K = 10			K = 20			mAP
		P	R	F1	P	R	F1		P	R	F1	P	R	F1	
CONSE [28]	ZSL	17.5	28.0	21.6	13.9	37.0	20.2	9.4	0.2	7.3	0.4	0.2	11.3	0.3	40.4
	GZSL	11.5	5.1	7.0	9.6	7.1	8.1	2.1	2.4	2.8	2.6	1.7	3.9	2.4	43.5
LabelEM [1]	ZSL	15.6	25.0	19.2	13.4	35.7	19.5	7.1	0.2	8.7	0.5	0.2	15.8	0.4	40.5
	GZSL	15.5	6.8	9.5	13.4	9.8	11.3	2.2	4.8	5.6	5.2	3.7	8.5	5.1	45.2
Fast0Tag [48]	ZSL	22.6	36.2	27.8	18.2	48.4	26.4	15.1	0.3	12.6	0.7	0.3	21.3	0.6	41.2
	GZSL	18.8	8.3	11.5	15.9	11.7	13.5	3.7	14.8	17.3	16.0	9.3	21.5	12.9	45.2
One Attention per Label [16]	ZS	20.9	33.5	25.8	16.2	43.2	23.6	10.4	-	-	-	-	-	-	-
	GZSL	17.9	7.9	10.9	15.6	11.5	13.2	3.7	-	-	-	-	-	-	-
One Attention per Cluster (M=10) [14]	ZSL	20.0	31.9	24.6	15.7	41.9	22.9	12.9	0.6	22.9	1.2	0.4	32.4	0.9	40.7
	GZSL	10.4	4.6	6.4	9.1	6.7	7.7	2.6	15.7	18.3	16.9	9.6	22.4	13.5	44.9
LESA (M=10) [14]	ZSL	25.7	41.1	31.6	19.7	52.5	28.7	19.4	0.7	25.6	1.4	0.5	37.4	1.0	41.7
	GZSL	23.6	10.4	14.4	19.8	14.6	16.8	5.6	16.2	18.9	17.4	10.2	23.9	14.3	45.4
Our Approach	ZSL	**26.6**	**42.8**	**32.8**	**20.1**	**53.6**	**29.3**	**25.7**	**1.3**	**42.4**	**2.5**	**1.1**	**52.1**	**2.2**	**43.0**
	GZSL	**30.9**	**13.6**	**18.9**	**26.0**	**19.1**	**22.0**	**8.9**	**33.6**	**38.9**	**36.1**	**22.8**	**52.8**	**31.9**	**49.7**

tains better performance over both one attention per label and per cluster. Our approach achieves state-of-the-art results with an absolute gain of 6.3% in terms mAP, over the best existing approach LESA. Similarly, our approach obtains consistent improvement in classification performance over the state-of-the-art in terms of F1 score ($K\in\{3,5\}$).

For the GZSL task, Fast0Tag and one attention per label achieve similar performance in terms of mAP. LESA obtains improved classification results among existing methods with mAP score of 5.6. Our approach achieves mAP score of 8.9, outperforming LESA with an absolute gain of 3.3%. Similarly, our approach achieves consistent improvement in classification performance with absolute gains of 4.5% and 5.2% over LESA in terms of F1 score at $K{=}3$ and $K{=}5$, respectively.

Fig. 4 shows the classification mAP improvement of our approach w.r.t. LESA for 15 unseen labels with the largest performance improvement as well as 15 unseen labels with the largest drop. Our proposed approach significantly improves (more than 26%) on several unseen labels, such as *waterfall*, *bear*, *dog*, *protest* and *food*, while having relatively smaller (less than 12%) negative impact on labels, such as *sunset*, *person*, *sky*, *water* and *rocks*. We observe our approach to be particularly better on animal categories (11 out of 13). We also observe our approach to struggle in case of abstract concepts (*e.g. sunset*, *sky*, *reflection*). In total, our approach outperforms LESA on 57 out of 81 labels.

Open Images: Tab. 2 shows the state-of-the-art comparison for ZSL and GZSL classification. The results are reported in terms of mAP and F1 score at *top-K* predictions ($K\in\{10,20\}$). Compared to the NUS-WIDE dataset, Open Images has significantly larger number of labels. This makes the ranking problem within an image more challeng-
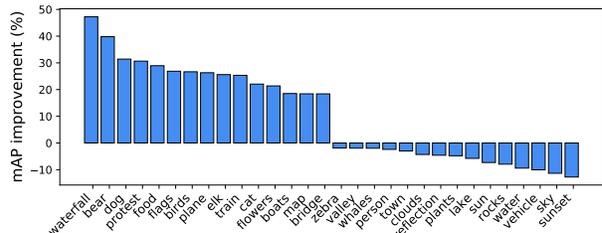

Figure 4. Classification mAP improvement comparison between our approach and LESA on NUS-WIDE. We show the comparison for 15 unseen labels with the largest gain as well 15 unseen labels with the largest drop. Our approach significantly improves (more than 26%) on several unseen labels (*e.g.*, *waterfall*, *bear*, *dog*, *protest* and *food*), while having relatively smaller (less than 12%) negative impact on other labels. Best viewed zoomed in.

ing, reflected by the lower F1 scores in the table. For the ZSL task, LESA obtains F1 scores of 1.4 and 1.0 at $K{=}10$ and $K{=}20$, respectively. Our approach performs favorably against LESA with F1 scores of 2.5 and 2.2 at $K{=}10$ and $K{=}20$, respectively. A similar performance gain is also observed for the mAP metric. It is worth noting that this dataset has 400 unseen labels, thereby making the problem of ZSL challenging. As in ZSL, our approach also achieves consistent improvement in performance, in terms of both F1 and mAP, over the state-of-the-art for the GZSL task.

MS COCO: Although the problem of zero-shot object detection is investigated, we are the first to evaluate zero-shot classification on this dataset. Tab. 3 shows the state-of-the-art comparison for ZSL and GZSL classification. Since the maximum number of unseen classes in an image is 3 in the validation set, we only report the F1 score at $K{=}3$. We re-implement LabelEM, CONSE and Fast0tag since their codes are not publicly available. For LESA, we obtain the results using the code-base from authors. Our approach ob-

Table 3. State-of-the-art comparison for ZSL and GZSL tasks on the MS COCO dataset having 65 seen and 15 unseen classes. We report the results in terms of mAP and F1 score at $K=3$. Our approach performs favorably against existing approaches for both ZSL and GZSL tasks. Best results are in bold.

Method	Task	F1 (K=3)	F1 (K=5)	mAP
CONSE [28]	ZSL	18.4	-	13.2
	GZSL	19.6	18.9	7.7
LabelEM [1]	ZSL	10.3	-	9.6
	GZSL	6.7	7.9	4.0
Fast0tag [48]	ZSL	37.5	-	43.3
	GZSL	33.8	34.6	27.9
LESA [14]	ZSL	33.6	-	31.8
	GZSL	26.7	28.0	17.5
Our Approach	ZSL	**43.5**	-	**52.2**
	GZSL	**44.1**	**43.4**	**33.2**

Table 4. Standard multi-label classification performance comparison on NUS-WIDE. The results are reported in terms of mAP and F1 score at $K \in \{3, 5\}$. Our approach achieves superior performance compared to existing methods. Best results are in bold.

Method	K=3			K=5			mAP
	P	R	F1	P	R	F1	
Logistic [34]	46.1	57.3	51.1	34.2	70.8	46.1	21.6
WARP [10]	49.1	61.0	54.4	36.6	75.9	49.4	3.1
WSABIE [40]	48.5	60.4	53.8	36.5	75.6	49.2	3.1
Fast0Tag [48]	48.6	60.4	53.8	36.0	74.6	48.6	22.4
CNN-RNN [37]	49.9	61.7	55.2	37.7	78.1	50.8	28.3
One Attention per Label [16]	51.3	63.7	56.8	38.0	78.8	51.3	32.6
One Attention per Cluster [14]	51.1	63.5	56.6	37.6	77.9	50.7	31.7
LESA [14]	52.3	65.1	58.0	38.6	80.0	52.0	31.5
Our Approach	**53.5**	**66.5**	**59.3**	**39.4**	**81.6**	**53.1**	**46.7**

tains superior (G)ZSL classification performance, in terms of F1 score and mAP, compared to existing methods.

4.4. Standard Multi-label Classification

In addition to zero-shot classification, we evaluate our approach for standard multi-label classification where all labels have training images. Tab. 4 shows the state-of-the-art comparison for standard multi-label classification on NUS-WIDE with 81 human annotated labels. As in [14], we remove all test samples without any label in the 81 label set. Among existing methods, CNN-RNN and LESA achieve mAP scores of 28.3 and 31.5, respectively. Our approach outperforms existing methods, achieving mAP score of 46.7. The proposed approach also performs favorably against existing methods in terms of F1 scores.

5. Extension to Zero-shot Object Detection

Lastly, we also evaluate our multi-label feature generation approach (CLF) for zero-shot object detection (ZSD). ZSD strives for simultaneous classification and localization of previously unseen objects. In the generalized settings, the test set contains both seen and unseen object categories

Table 5. State-of-the-art comparison for ZSD and GZSD tasks on MS COCO. The results are reported in terms of Recall. For GZSD, we report the harmonic mean (HM) between seen and unseen classes. Our approach performs favorably against existing methods. Best results are in bold.

Method	Seen/Unseen Split	ZSD	GZSD		
			Seen	Unseen	HM
SB [3]	48/17	24.39	-	-	-
DSES [3]	48/17	27.19	15.02	15.32	15.17
PL-48 [30]	48/17	43.56	38.24	26.32	31.18
PL-65 [30]	65/15	37.72	36.38	37.16	36.76
SUZOD [12]	65/15	54.40	57.70	54.40	55.74
Our Approach	65/15	**58.10**	**58.90**	**58.10**	**58.50**

(GZSD). Recently, the work of [12] introduce a zero-shot detection approach (SUZOD) where features are synthesized, conditioned upon class-embeddings, and integrated in the popular Faster R-CNN framework [31]. The feature generation stage is jointly driven by the classification loss in the semantic embedding space for both seen and unseen categories. Their approach addresses multi-label zero-shot detection by constructing single-label features for each region of interest (RoI). Different from SUZOD [12], we first generate a pool of multi-label RoI features by integrating random sets of individual single-label RoI features. These integrated multi-label RoI features are then used as real features to train our CLF-based classification architecture.

Tab. 5 shows the state-of-the-art comparison, in terms of recall, for ZSD and GZSD detection on MS COCO. For GZSD, Harmonic Mean (HM) of performances for seen and unseen classes are reported. Similar to SUZOD [12], we also report the results by using f-CLSWGAN as the underlying generative architecture. Our approach performs favorably against existing methods for both ZSD and GZSD.

6. Conclusion

We investigate the problem of multi-label feature synthesis in the zero-shot setting. We introduce three different fusion approaches (ALF, FLF and CLF) to synthesize multi-label features. Our ALF synthesizes features by integrating class-specific attribute embeddings at the generator input. On the other hand, FLF synthesizes features from class-specific embeddings individually and integrates them in feature space. Our CLF combines the advantages of both ALF and FLF, using each individual-level feature and attends to the bi-level context. Consequently, individual-level features adapt themselves producing enriched synthesized features that are pooled to obtain final output. We integrate our fusion approaches in two generative architectures. Our approach outperforms existing zero-shot methods on three multi-label classification benchmarks. Lastly, we also show the effectiveness of our approach for zero-shot detection.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 2015. 1, 7, 8

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

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 5, 6, 8

[4] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, 2019. 1

[5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 2009. 1, 2, 3, 5

[6] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *CVPR*, 2019. 1

[7] Rafael Felix, B. G. Vijay Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018. 1, 2

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013. 1

[9] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *TPAMI*, 2015. 1

[10] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. 8

[11] Ian Goodfellow, Jean PougetAbadie, Mehdi Mirza, Bing Xu, and David Warde-Farley. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[12] Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *ACCV*, 2020. 5, 6, 8

[13] He Huang, Changhu Wang, Philip S Yu, and Chang-Dong Wang. Generative dual adversarial network for generalized zero-shot learning. In *CVPR*, 2019. 1, 2

[14] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *CVPR*, 2020. 1, 3, 5, 6, 7, 8

[15] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *NeurIPS*, 2014. 1

[16] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 7, 8

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014. 1, 2, 5

[18] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1

[19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 2, 3, 5

[20] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. In *CVPR*, 2018. 1

[21] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *CVPR*, 2019. 1, 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 3, 5

[23] 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 *CVPR*, 2019. 1, 2

[24] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 1

[25] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. *NeurIPS*, 2017. 1

[26] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. *arXiv preprint arXiv:2003.07833*, 2020. 1

[27] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 2

[28] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 7, 8

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 3, 6

[30] Shafin Rahman, Salman Khan, and Nick Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018. 8

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 8

[32] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *NeurIPS*, 2013. 1

[33] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 1

[34] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *IJDWM*, 2007. 8

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[36] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, 2017. 6

[37] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, 2016. 1, 8

[38] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*, 2017. 1

[39] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. *Technical Report CNS-TR-2010-001, Caltech*, 2010. 2

[40] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011. 1, 8

[41] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018. 1, 2

[42] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 1, 2, 5

[43] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, 2019. 1, 2, 5

[44] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *CVPR*, 2020. 1

[45] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *ECCV*, 2020. 1

[46] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *CVPR*, 2017. 1

[47] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *AAAI*, 2020. 1

[48] Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In *CVPR*, 2016. 1, 5, 6, 7, 8