Discriminative Region-based Multi-Label Zero-Shot Learning

Sanath Narayan*[1] Akshita Gupta*[1] Salman Khan[2] Fahad Shahbaz Khan[2,3]
Ling Shao[1] Mubarak Shah[4]
[1]Inception Institute of Artificial Intelligence, UAE [2]Mohamed Bin Zayed University of AI, UAE
[3]Linköping University, Sweden [4]University of Central Florida, USA

Abstract

Multi-label zero-shot learning (ZSL) is a more realistic counter-part of standard single-label ZSL since several objects can co-exist in a natural image. However, the occurrence of multiple objects complicates the reasoning and requires region-specific processing of visual features to preserve their contextual cues. We note that the best existing multi-label ZSL method takes a shared approach towards attending to region features with a common set of attention maps for all the classes. Such shared maps lead to diffused attention, which does not discriminatively focus on relevant locations when the number of classes are large. Moreover, mapping spatially-pooled visual features to the class semantics leads to inter-class feature entanglement, thus hampering the classification. Here, we propose an alternate approach towards region-based discriminability-preserving multi-label zero-shot classification. Our approach maintains the spatial resolution to preserve region-level characteristics and utilizes a bi-level attention module (BiAM) to enrich the features by incorporating both region and scene context information. The enriched region-level features are then mapped to the class semantics and only their class predictions are spatially pooled to obtain image-level predictions, thereby keeping the multi-class features disentangled. Our approach sets a new state of the art on two large-scale multi-label zero-shot benchmarks: NUS-WIDE and Open Images. On NUS-WIDE, our approach achieves an absolute gain of 6.9% mAP for ZSL, compared to the best published results. Source code is available at https://github.com/akshitac8/BiAM.

1. Introduction

Multi-label classification strives to recognize all the categories (labels) present in an image. In the standard multi-label classification [32, 39, 15, 4, 21, 40, 41] setting, the category labels in both the train and test sets are identical. In contrast, the task of multi-label zero-shot learning (ZSL) is to recognize multiple new unseen categories in images at test time, without having seen the corresponding visual examples during training. In the generalized ZSL (GZSL) setting, test images can simultaneously contain multiple seen and unseen classes. GZSL is particularly challenging in the large-scale multi-label setting, where several diverse categories occur in an image (*e.g.*, maximum of 117 labels per image in NUS-WIDE [5]) along with a large number of unseen categories at test time (*e.g.*, 400 unseen classes in Open Images [16]). Here, we investigate this challenging problem of multi-label (generalized) zero-shot classification.

Existing multi-label (G)ZSL methods tackle the problem by using global image features [20, 46], structured knowledge graph [17] and attention schemes [13]. Among these, the recently introduced LESA [13] proposes a shared attention scheme based on region-based feature representations and achieves state-of-the-art results. LESA learns multiple attention maps that are shared across all categories. The region-based image features are weighted by these shared attentions and then spatially aggregated. Subsequently, the aggregated features are projected to the label space via a joint visual-semantic embedding space.

While achieving promising results, LESA suffers from two key limitations. Firstly, classification is performed on features obtained using a set of attention maps that are shared across all the classes. In such a shared attention framework, many categories are observed to be inferred from only a few dominant attention maps, which tend to be diffused across an image rather than discriminatively focusing on regions likely belonging to a specific class (see Fig. 1). This is problematic for large-scale benchmarks comprising several hundred categories, *e.g.*, more than $7k$ seen classes in Open Images [16] with significant inter and intra-class variations. Secondly, attended features are spatially pooled before projection to the label space, thus entangling the multi-label information in the collapsed image-level feature vectors. Since multiple diverse labels can appear in an image, the class-specific discriminability within such a collapsed representation is severely hampered.

*Equal contribution

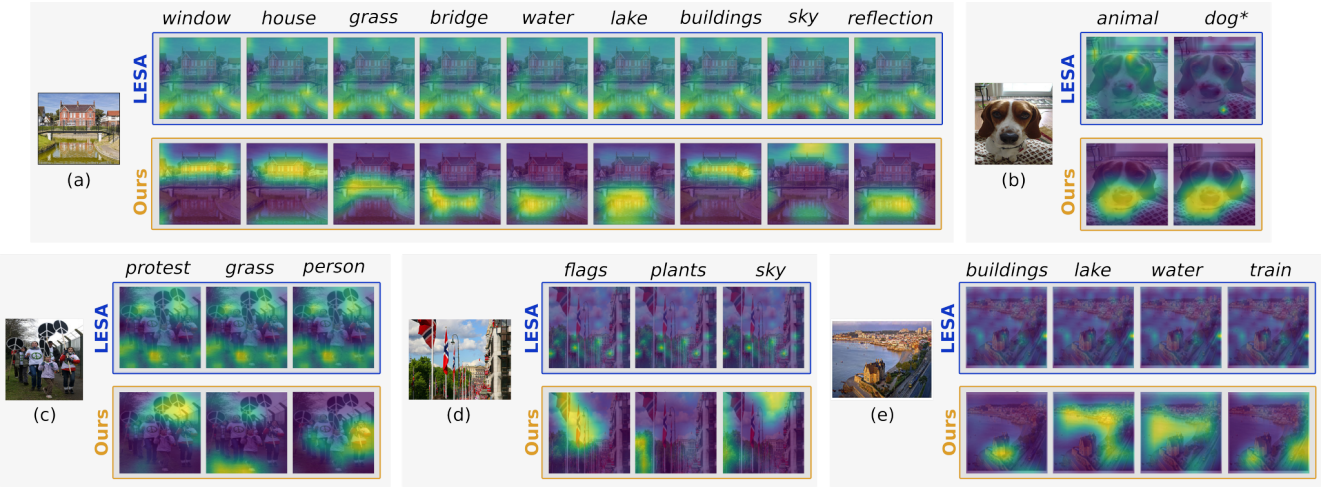

Figure 1. **Comparison, in terms of attention visualization, between shared attention-based LESA [13] and our approach on example NUS-WIDE test images**. For each image, the visualization of attentions of positive labels within that image are shown for LESA (top row) and our approach (bottom row). In the case of LESA, all classes in these examples are inferred from the eighth shared attention module except for *dog* class in (b), which is inferred from the ninth module. As seen in these examples, these dominant attention maps struggle to discriminatively focus on relevant (class-specific) regions. In contrast, our proposed approach based on a bi-level attention module (BiAM) produces attention maps by preserving class-specific discriminability, leading to an enriched feature representation. Our BiAM effectively captures region-level semantics as well as global scene-level context, thereby enabling it to accurately attend to object class (*e.g.*, *window* class in (a)) *and* abstract concepts (*e.g.*, *reflection* class in (a)). Best viewed zoomed in.

1.1. Contributions

To address the aforementioned problems, we pose large-scale multi-label ZSL as a region-level classification problem. We introduce a simple yet effective region-level classification framework that maintains the spatial resolution of features to keep the multi-class information disentangled for dealing with large number of co-existing classes in an image. Our framework comprises a bi-level attention module (BiAM) to contextualize and obtain highly discriminative region-level feature representations. Our BiAM contains region and global (scene) contextualized blocks and enables reasoning about all the regions together using pair-wise relations between them, in addition to utilizing the holistic scene context. The region contextualized block enriches each region feature by attending to all regions within the image whereas the scene contextualized block enhances the region features based on their congruence to the scene feature representation. The resulting discriminative features, obtained through our BiAM, are then utilized to perform region-based classification through a compatibility function. Afterwards, a spatial *top-k* pooling is performed over each class to obtain the final predictions.

Experiments are performed on two challenging large-scale multi-label zero-shot benchmarks: NUS-WIDE [5] and Open Images [16]. Our approach performs favorably against existing methods, setting a new state of the art on both benchmarks. Particularly, on NUS-WIDE, our approach achieves an absolute gain of 6.9% in terms of mAP for the ZSL task, over the best published results [13].

2. Proposed Method

Here, we introduce a region-based discriminability-preserving multi-label zero-shot classification framework aided by learning rich features that explicitly encodes both region as well as global scene contexts in an image.

Problem Formulation: Let $\mathbf{x} \in \mathcal{X}$ denote the feature instances of a multi-label image $i \in \mathcal{I}$ and $\mathbf{y} \in \{0,1\}^S$ the corresponding multi-hot labels from the set of S seen class labels $\mathcal{C}^s$. Further, let $\mathbf{A}_S \in \mathbb{R}^{S \times d_a}$ denote the d_a-dimensional attribute embeddings, which encode the semantic relationships between S seen classes. With n_p as the number of positive labels in an image, we denote the set of attribute embeddings for the image as $a_\mathbf{y} = \{\mathbf{A}_j, \forall j : \mathbf{y}[j]=1\}$, where $|a_\mathbf{y}| = n_p$. The goal in (generalized) zero-shot learning is to learn a mapping $f(\mathbf{x}) : \mathcal{X} \rightarrow \{0,1\}^S$ aided by the attribute embeddings $a_\mathbf{y}$, such that the mapping can be adapted to include the U unseen classes (with embeddings $\mathbf{A}_U \in \mathbb{R}^{U \times d_a}$) at test time, *i.e.*, $f(\mathbf{x}) : \mathcal{X} \rightarrow \{0,1\}^U$ for ZSL and $f(\mathbf{x}) : \mathcal{X} \rightarrow \{0,1\}^C$ for the GZSL setting. Here, $C = S + U$ represents the total number of seen and unseen classes.

2.1. Region-level Multi-label ZSL

As discussed earlier, recognizing diverse and wide range of category labels in images under the (generalized) zero-shot setting is challenging. The problem arises, primarily, due to the entanglement of features of the various different classes present in an image. Fig. 2(a) illustrates this feature entanglement in the shared attention-based classification pipeline [13] that integrates multi-label features by

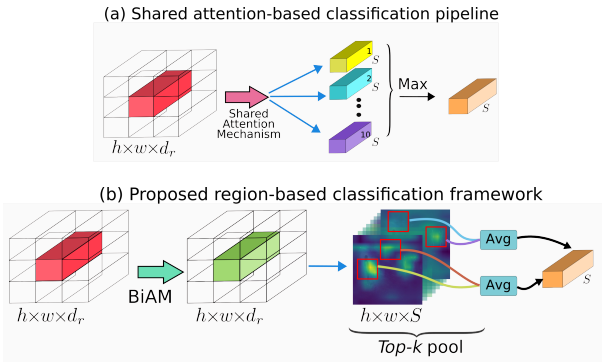

Figure 2. **Comparison of our region-level classification framework (b) with the shared attention-based classification pipeline (a) in [13].** The shared attention-based pipeline performs an attention-weighted spatial averaging of the region-based features to generate a feature vector per shared attention. These (spatially pooled) features are then classified to obtain S class scores per shared attention, which are max-pooled to obtain image-level class predictions. In contrast, our framework minimizes inter-class feature entanglement by enhancing the region-based features through a feature enrichment mechanism, which preserves the spatial resolution of the features. Each region-based enriched feature representation is then classified to S seen classes. Afterwards, *per class top-k* activations are aggregated to obtain image-level predictions.

performing a weighted spatial averaging of the region-based features based on the shared-attention maps. In this work, we argue that entangled feature representations are sub-optimal for multi-label classification and instead propose to alleviate this issue by posing large-scale multi-label ZSL as a region-level classification problem. To this end, we introduce a simple but effective region-level classification framework that first enriches the region-based features by the proposed feature enrichment mechanism. It then classifies the enriched region-based features followed by spatially pooling the *per-class* region-based scores to obtain the final image-level class predictions (see Fig. 2(b)). Consequently, our framework minimizes inter-class feature entanglement and enhances the classification performance.

Fig. 3 shows our overall proposed framework. Let $\mathbf{e}_f \in \mathbb{R}^{h\times w\times d_r}$ be the output region-based features, which are to be classified, from our proposed enrichment mechanism (*i.e.*, BiAM). Here, h, w denote the spatial extent of the region-based features with $h \cdot w$ regions. These features $\mathbf{e}_f$ are first aligned with the class-specific attribute embeddings of the seen classes. This alignment is performed, *i.e.*, a joint visual-semantic space is learned, so that the classifier can be adapted to the unseen classes at test time. The aligned region-based features are classified to obtain class-specific response maps $\mathbf{m} \in \mathbb{R}^{h\times w\times S}$ given by,

$$\mathbf{m} = \mathbf{e}_f \mathbf{W}_a \mathbf{A}_S^\top, \qquad \text{s.t., } \mathbf{A}_S \in \mathbb{R}^{S\times d_a}, \tag{1}$$

where $\mathbf{W}_a \in \mathbb{R}^{d_r\times d_a}$ is a learnable weight matrix that is used to reshape the visual features to attribute embeddings of seen classes ($\mathbf{A}_S$). The response maps are then *top-k* pooled along the spatial dimensions to obtain image-level *per-class* scores $\mathbf{s} \in \mathbb{R}^S$, which are then utilized for training the network (in Sec. 2.3). Such a region-level classification, followed by a score-level pooling, helps to preserve the discriminability of the features in each of the $h \cdot w$ regions by minimizing the feature entanglement of different positive classes occurring in the image.

The aforementioned region-level multi-label ZSL framework relies on discriminative region-based features. Standard region-based features $\mathbf{x}$ only encode local region-specific information and do not explicitly reason about all the regions together. Moreover, region-based features do not possess image-level holistic scene information. Next, we introduce a bi-level attention module (BiAM) to enhance feature discriminability and generate enriched features $\mathbf{e}_f$.

2.2. Bi-level Attention Module

Here, we present a bi-level attention module (BiAM) that enhances region-based features by incorporating both region and scene context information, without sacrificing the spatial resolution. Our BiAM comprises region and scene contextualized blocks, which are described next.

2.2.1 Region Contextualized Block

The region-contextualized block (RCB) enriches the region-based latent features $\mathbf{h}_r$ by capturing the contexts from different regions in the image. We observe encoding the individual contexts of different regions in an image to improve the discriminability of standard region-based features, *e.g.*, the context of a region with *window* can aid in identifying other possibly texture-less regions in the image as *house* or *building*. Thus, inspired by the multi-headed self-attention [30], our RCB allows the features in different regions to interact with each other and identify the regions to be paid more attention to for enriching themselves (see Fig. 3(b)). To this end, the input features $\mathbf{x}_r \in \mathbb{R}^{h\times w\times d_r}$ are first processed by a 3×3 convolution layer to obtain latent features $\mathbf{h}_r \in \mathbb{R}^{h\times w\times d_r}$. These latent features are then projected to a low-dimensional space ($d'_r = {}^{d_r}/_{H}$) to create query-key-value triplets using a total of H projection heads,

$$\mathbf{q}_h^r = \mathbf{h}_r \mathbf{W}_h^Q, \quad \mathbf{k}_h^r = \mathbf{h}_r \mathbf{W}_h^K, \quad \mathbf{v}_h^r = \mathbf{h}_r \mathbf{W}_h^V, \tag{2}$$

where $h \in \{1, 2, .., H\}$ and $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V$ are learnable weights of 1×1 convolution layers with input and output channels as d_r and d'_r, respectively. The *query* vector (of length d'_r) derived from each region feature[1] is used to find its correlation with the *keys* obtained from all the region features, while the *value* embedding holds the status of the current form of each region feature.

[1] Query $\mathbf{q}_h^r \in \mathbb{R}^{h\times w\times d'_r}$ can be considered as $h\cdot w$ queries represented by d'_r features each. Similar observation holds for keys, values, *etc.*

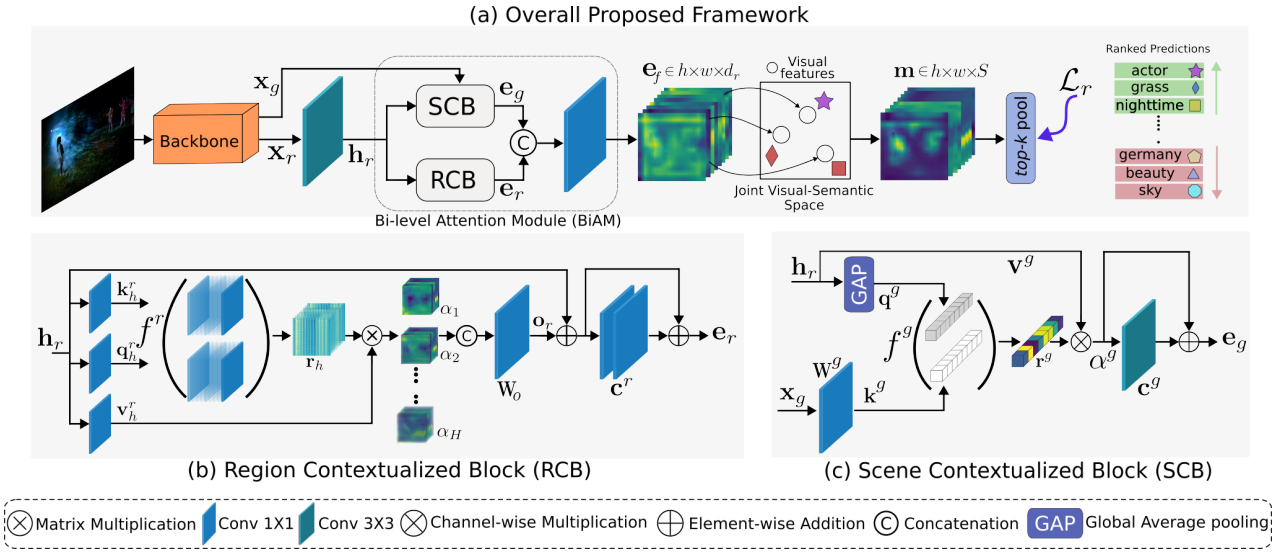

Figure 3. **Our region-level multi-label (G)ZSL framework:** The top row shows an overview of our network architecture. Given an image, the region-level features $\mathbf{x}_r$ are first obtained using a backbone. The region features are enriched using a Bi-level Attention Module (BiAM). This module incorporates region (b) and scene (c) contextualized blocks which learn to aggregate region-level and scene-specific context, respectively, which is in turn used to enhance the region features. The enriched features $\mathbf{e}_f$ are mapped to the joint visual-semantic space to relate them with class semantics, obtaining $\mathbf{m}$. Per-class region-based prediction scores are then spatially pooled to generate final image-level predictions. Notably, our design ensures *region-level feature enrichment* while preserving the spatial resolution uptil class predictions are made, which *minimizes inter-class feature entanglement*, a key requisite for large-scale multi-label (G)ZSL.

Given these triplets for each head, first, an intra-head processing is performed by relating each query vector with '*keys*' derived from the $h \cdot w$ region features. The resulting normalized relation scores ($\mathbf{r}_h \in \mathbb{R}^{hw \times hw}$) from the softmax function (σ) are used to reweight the corresponding '*value*' vectors. Without loss of generality[1], the attended features $\boldsymbol{\alpha}_h \in \mathbb{R}^{h \times w \times d'_r}$ are given by,

$$\boldsymbol{\alpha}_h = \mathbf{r}_h \mathbf{v}_h^r, \quad \text{where } \mathbf{r}_h = \sigma\left(\frac{\mathbf{q}_h^r \mathbf{k}_h^{r\top}}{\sqrt{d'_r}}\right). \tag{3}$$

Next, these low-dimensional self-attended features from each head are channel-wise concatenated and processed by a convolution layer $\mathbf{W}_o$ to generate output $\mathbf{o}_r \in \mathbb{R}^{h \times w \times d_r}$,

$$\mathbf{o}_r = [\boldsymbol{\alpha}_1; \boldsymbol{\alpha}_2; \dots \boldsymbol{\alpha}_H]\mathbf{W}_o. \tag{4}$$

To encourage the network to selectively focus on adding complimentary information to the 'source' latent feature $\mathbf{h}_r$, a residual branch is added to the attended features $\mathbf{o}_r$ and further processed with a small residual sub-network $c^r(\cdot)$, comprising two 1×1 convolution layers, to help the network first focus on the local neighbourhood and then progressively pay attention to the other-level features. The enriched region-based features $\mathbf{e}_r \in \mathbb{R}^{h \times w \times d_r}$ from the RCB are given by,

$$\mathbf{e}_r = c^r(\mathbf{h}_r + \mathbf{o}_r) + (\mathbf{h}_r + \mathbf{o}_r). \tag{5}$$

Consequently, the discriminability of the latent features $\mathbf{h}_r$ is enhanced by self-attending to the context of different regions in the image, resulting in enriched features $\mathbf{e}_r$.

2.2.2 Scene Contextualized Block

As discussed earlier, the RCB captures the regional context in the image, enabling reasoning about all regions together using pair-wise relations between them. In this way, RCB enriches the latent feature inputs $\mathbf{h}_r$. However, such a region-based contextual attention does not effectively encode the global scene-level context of the image, which is necessary for understanding abstract scene concepts like *night-time*, *protest*, *clouds*, *etc*. Understanding such labels from local regional contexts is challenging due to their abstract nature. Thus, in order to better capture the holistic scene-level context, we introduce a scene contextualized block (SCB) within our BiAM. Our SCB attends to the region-based latent features $\mathbf{h}_r$, based on their congruence with the global image feature $\mathbf{x}_g$ (see Fig. 3(c)). To this end, the learnable weights $\mathbf{W}^g$ project the features $\mathbf{x}_g$ to a d_r-dimensional space to obtain the global '*key*' vectors $\mathbf{k}^g \in \mathbb{R}^{d_r}$, while the latent features $\mathbf{h}_r$ are spatially average pooled to create the '*query*' vectors $\mathbf{q}^g \in \mathbb{R}^{d_r}$,

$$\mathbf{q}^g = \text{GAP}(\mathbf{h}_r), \quad \mathbf{k}^g = \mathbf{x}_g \mathbf{W}^g, \quad \mathbf{v}^g = \mathbf{h}_r. \tag{6}$$

The region-based latent features $\mathbf{h}_r$ are retained as '*value*' features $\mathbf{v}^g$. Given these query-key-value triplets, first, the *query* $\mathbf{q}^g$ is used to find its correlation with the *key* $\mathbf{k}^g$. The resulting relation score vectors $\mathbf{r}^g \in \mathbb{R}^{d_r}$ are then used to reweight the corresponding channels in *value* features to obtain the attended features $\boldsymbol{\alpha}^g \in \mathbb{R}^{h \times w \times d_r}$, given by,

$$\boldsymbol{\alpha}^g = \mathbf{v}^g \otimes \mathbf{r}^g, \quad \text{where } \mathbf{r}^g = \text{sigmoid}(\mathbf{q}^g * \mathbf{k}^g), \tag{7}$$

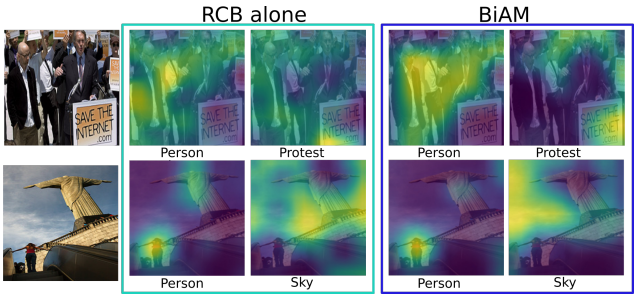

Figure 4. **Effect of enhancing the region-based features through our feature enrichment mechanism: BiAM**. The two complementary RCB and SCB blocks in BiAM integrate region-level semantics and global scene-level context, leading to a more discriminative feature representation. While RCB alone (on the left) is able to capture the region-level semantics of *person* class, it confuses those related to *protest* label. However, encoding the global scene-level context from the SCB in BiAM (on the right) improves the semantic recognition of scene-level concepts like *protest*.

where $\otimes$ and $*$ denote channel-wise and element-wise multiplications. The channel-wise operation is chosen here since we want to use the global contextualized features to dictate kernel-wise importance of the feature channels for aggregating relevant contextual cues without disrupting the local filter signature. Similar to RCB, to encourage the network to selectively focus on adding complimentary information to the 'source' $\mathbf{h}_r$, a residual branch is added after processing the attended features through a 3×3 convolution layer $c^g(\cdot)$. The scene-context enriched features $\mathbf{e}_g \in \mathbb{R}^{h\times w\times d_r}$ from the SCB are given by,

$$\mathbf{e}_g = c^g(\boldsymbol{\alpha}^g) + \mathbf{h}_r. \tag{8}$$

In order to ensure the enrichment due to both region and global contexts are well captured, the enriched features ($\mathbf{e}_r$ and $\mathbf{e}_g$) from both region and scene contextualized blocks are channel-wise concatenated and processed through a 1×1 channel-reducing convolution layer $c^f(\cdot)$ to obtain the final enriched features $\mathbf{e}_f \in \mathbb{R}^{h\times w\times d_r}$, given by,

$$\mathbf{e}_f = c^f([\mathbf{e}_r; \mathbf{e}_g]). \tag{9}$$

Fig. 4 shows that encoding scene context into the region-based features improves the attention maps of scene level labels (*e.g.*, *protest*), which were hard to attend to using only the region context. Consequently, our bi-level attention module effectively reasons about all the image regions together using pair-wise relations between them, while being able to utilize the whole image (holistic) scene as context.

2.3. Training and Inference

As discussed earlier, discriminative region-based features $\mathbf{e}_f$ are learned and region-wise classified to obtain class-specific response maps $\mathbf{m} \in \mathbb{R}^{h\times w\times S}$ (using Eq. 1). The response maps $\mathbf{m}$ are further *top-k* pooled spatially to compute the image-level *per-class* scores $\mathbf{s} \in \mathbb{R}^S$. The network is trained using a simple, yet effective ranking loss $\mathcal{L}_{rank}$ on the predicted scores $\mathbf{s}$, given by,

$$\mathcal{L}_{rank} = \sum_i \sum_{p\in\mathbf{y}_p, n\notin\mathbf{y}_p} \max(\mathbf{s}_i[n] - \mathbf{s}_i[p] + 1, 0). \tag{10}$$

Here, $\mathbf{y}_p = \{j : \mathbf{y}[j]=1\}$ denotes the positive labels in image i. The ranking loss ensures that the predicted scores of the positive labels present in the image rank ahead, by a margin of at least 1, of the negative label scores.

At test time, for the multi-label ZSL task, the unseen class attribute embeddings $\mathbf{A}_U \in \mathbb{R}^{U\times d_a}$ of the respective unseen classes are used (in place of $\mathbf{A}_S$) for computing the class-specific response maps $\mathbf{m} \in \mathbb{R}^{h\times w\times U}$ in Eq. 1. As in training, these response maps are then *top-k* pooled spatially to compute the image-level *per-class* scores $\mathbf{s} \in \mathbb{R}^U$. Similarly, for the multi-label GZSL task, the concatenated embeddings ($\mathbf{A}_C \in \mathbb{R}^{C\times d_a}$) of all the classes $C = S+U$ are used to classify the multi-label images.

3. Experiments

Datasets: We evaluate our approach on two benchmarks: NUS-WIDE [5] and Open Images [16]. 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 [13, 46], the 925 and 81 labels are used as seen and unseen classes, respectively. The **Open Images** (v4) is a large-scale dataset comprising nearly 9 million training images along with 41,620 and 125,456 images in validation and test sets. It has annotations with human and machine-generated labels. Here, 7,186 labels, with at least 100 training images, are selected as seen classes. The most frequent 400 test labels that are absent in the training data are selected as unseen classes, as in [13].

Evaluation Metrics: We use F1 score at *top-K* predictions and mean Average Precision (mAP) as evaluation metrics, as in [31, 13]. The model's ability to correctly rank labels in each image is measured by the F1, while the its image ranking accuracy for each label is captured by the mAP.

Implementation Details: Pretrained VGG-19 [28] is used to extract features from multi-label images, as in [46, 13]. The region-based features (of size h, w=14 and d_r=512) from *Conv*$_5$ are extracted along with the global features of size d_g=4,096 from FC7. As in [13], ℓ_2-normalized 300-dimensional GloVe [25] vectors of the class names are used as the attribute embeddings $\mathbf{A}_C$. The two 3×3 convolutions (input and output channels are set to 512) are followed by ReLU and batch normalization layers. The k for *top-k* pooling is set to 10, while the heads H=8. For training, we use the ADAM optimizer with (β_1, β_2) as (0.5, 0.999) and a gradual warm-up learning rate scheduler with an initial lr of $1e^{-3}$. Our model is trained with a mini-batch size of 32 for 40 epochs on NUS-WIDE and 2 epochs on Open Images.

Table 1. **State-of-the-art comparison for multi-label ZSL and GZSL tasks on NUS-WIDE**. We report the results in terms of mAP and F1 score at $K \in \{3, 5\}$. Our approach outperforms the state-of-the-art for both ZSL and GZSL tasks, in terms of mAP and F1 score. Best results are in bold.

Method	Task	mAP	F1 (K = 3)	F1 (K = 5)
CONSE [23]	ZSL	9.4	21.6	20.2
	GZSL	2.1	7.0	8.1
LabelEM [1]	ZSL	7.1	19.2	19.5
	GZSL	2.2	9.5	11.3
Fast0Tag [46]	ZSL	15.1	27.8	26.4
	GZSL	3.7	11.5	13.5
Attention per Label [14]	ZSL	10.4	25.8	23.6
	GZSL	3.7	10.9	13.2
Attention per Cluster [13]	ZSL	12.9	24.6	22.9
	GZSL	2.6	6.4	7.7
LESA [13]	ZSL	19.4	31.6	28.7
	GZSL	5.6	14.4	16.8
Our Approach	ZSL	**26.3**	**33.1**	**30.7**
	GZSL	**9.3**	**16.1**	**19.0**

Table 2. **State-of-the-art comparison for multi-label ZSL and GZSL tasks on Open Images**. Results are reported in terms of mAP and F1 score at $K \in \{10, 20\}$. Our approach sets a new state of the art for both tasks, in terms of mAP and F1 score. Best results are in bold.

Method	Task	mAP	F1 (K = 10)	F1 (K = 20)
CONSE [23]	ZSL	40.4	0.4	0.3
	GZSL	43.5	2.6	2.4
LabelEM [1]	ZSL	40.5	0.5	0.4
	GZSL	45.2	5.2	5.1
Fast0Tag [46]	ZSL	41.2	0.7	0.6
	GZSL	45.2	16.0	12.9
Attention per Cluster [13]	ZSL	40.7	1.2	0.9
	GZSL	44.9	16.9	13.5
LESA [13]	ZSL	41.7	1.4	1.0
	GZSL	45.4	17.4	14.3
Our Approach	ZSL	**73.6**	**8.3**	**5.5**
	GZSL	**84.5**	**19.1**	**15.9**

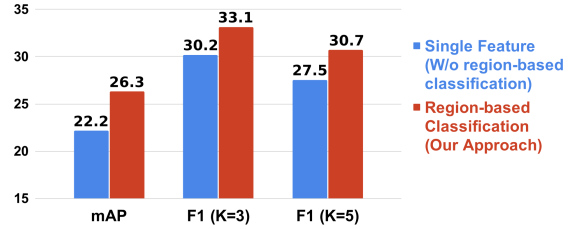

Figure 5. **Impact of region-based classification for the ZSL task on NUS-WIDE**, in terms of mAP and F1 at $K \in \{3, 5\}$. Classifying spatially pooled features (blue bars) entangles the features of the different classes resulting in sub-optimal performance. In contrast, our proposed approach, which classifies each region individually and then spatially pools the *per region* class scores (red bars), minimizes the inter-class feature entanglement and achieves superior classification performance.

3.1. State-of-the-art Comparison

NUS-WIDE: The state-of-the-art comparison for zero-shot (ZSL) and generalized zero-shot (GZSL) classification is presented in Tab. 1. The results are reported in terms of mAP and F1 score at *top-K* predictions ($K \in \{3, 5\}$). The approach of Fast0Tag [46], which finds principal directions in the attribute embedding space for ranking the positive tags ahead of negative tags, achieves 15.1 mAP on the ZSL task. The recently introduced LESA [13], which employs a shared multi-attention mechanism to recognize labels in an image, improves the performance over Fast0Tag, achieving 19.4 mAP. Our approach outperforms LESA with an absolute gain of 6.9% mAP. Furthermore, our approach achieves consistent improvement over the state-of-the-art in terms of F1 ($K \in \{3, 5\}$), achieving gains as high as 2.0% at $K=5$.

Similarly, on the GZSL task, our approach achieves an mAP score of 9.3, outperforming LESA with an absolute gain of 3.7%. Moreover, consistent performance improvement in terms of F1 is achieved over LESA by our approach, with absolute gains of 1.5% and 2.2% at $K=3$ and $K=5$.

Open Images: Tab. 2 shows the state-of-the-art comparison for multi-label ZSL and GZSL tasks. The results are reported in terms of mAP and F1 score at *top-K* predictions ($K \in \{10, 20\}$). We follow the same evaluation protocol as in the concurrent work of SDL [2]. Since Open Images has significantly larger number of labels, in comparison to NUS-WIDE, ranking them within an image is more challenging. This is reflected by the lower F1 scores in the table. Among existing methods, LESA obtains an mAP of 41.7% for the ZSL task. In comparison, our approach outperforms LESA by achieving 73.6% mAP with an absolute gain of 31.9%. Furthermore, our approach performs favorably against the best existing approach with F1 scores of 8.3 and 5.5 at $K=10$ and $K=20$. It is worth noting that the ZSL task is challenging due to the high number of unseen labels (400). As in ZSL, our approach obtains a significant gain of 39.1% mAP over the best published results for GZSL and also achieves favorable performance in F1. Additional details and results are presented in the supplementary.

3.2. Ablation Study

Impact of region-based classification: To analyse this impact, we train our proposed framework without region-based classification, where the enriched features $\mathbf{e}_f$ are spatially average-pooled to a single feature representation (of size d_r) per image and then classified. Fig. 5 shows the performance comparison between our frameworks trained with and without region-based classification in terms of mAP and F1. Since images have large and diverse set of positive labels, spatially aggregating features without the region-based classification (blue bars), leads to inter-class feature entanglement, as discussed in Sec. 2.1. Instead, preserving the spatial dimension by classifying the region-based features, as in the proposed framework (red bars), mitigates

Table 3. **Impact of the proposed BiAM comprising RCB and SCB blocks**. Note that all results here are reported with the same region-level classification framework and only the features utilized within the classification framework differs. Both RCB alone and SCB alone achieve consistently improved performance over standard region features. For both ZSL and GZSL tasks, the best performance is obtained when utilizing the discriminative features obtained from the proposed BiAM. Best results are in bold.

Method	Task	mAP	F1 (K = 3)	F1 (K = 5)
Standard region features	ZSL	21.1	28.0	26.9
	GZSL	6.8	12.0	14.5
RCB alone	ZSL	23.7	31.9	29.0
	GZSL	7.6	14.7	17.6
SCB alone	ZSL	23.2	29.4	27.8
	GZSL	8.6	14.0	16.7
BiAM (RCB + SCB)	ZSL	**26.3**	**33.1**	**30.7**
	GZSL	**9.3**	**16.1**	**19.0**

Table 4. **ZSL comparison on NUS-WIDE with attention variants:** our attention (left) and other attentions [33, 12] (right).

Method	mAP
BiAM: RCB w/ LayerNorm	25.0
BiAM: RCB w/ *sigmoid*	24.6
BiAM: SCB w/ *softmax*	24.3
BiAM: Final	**26.3**

Method	mAP
Non-Local [33]	23.1
Criss-Cross Atn [12]	23.9
BiAM (Ours)	**26.3**

the inter-class feature entanglement to a large extent. This leads to a superior performance for the region-based classification on both multi-label ZSL and GZSL tasks. These results suggest the importance of region-based classification for learning discriminative features in large-scale multi-label (G)ZSL tasks. Furthermore, Fig. 6 presents a t-SNE visualization showing the impact of our region-level classification framework on 10 unseen classes from NUS-WIDE.

Impact of the proposed BiAM: Here, we analyse the impact of our feature enrichment mechanism (BiAM) to obtain discriminative feature representations. Tab. 3 presents the comparison between region-based classification pipelines based on standard features $\mathbf{h}_r$ and discriminative features $\mathbf{e}_f$ obtained from our BiAM on NUS-WIDE. We also present results of our RCB and SCB blocks alone. Both RCB alone and SCB alone consistently improve the (G)ZSL performance over the standard region-based features. This shows that our region-based classification pipeline benefits from the discriminative features obtained through the two complementary attention blocks. Furthermore, best results are obtained with our BiAM that comprises both RCB and SCB blocks, demonstrating the importance of encoding both region *and* scene context information. Fig. 8 shows a comparison between the standard features-based classification and the proposed classification framework utilizing BiAM on example unseen class images.

Varying the attention modules: Tab. 4 (left) shows the comparison on NUS-WIDE when ablating RCB and SCB modules in our BiAM. Including LayerNorm in RCB or replacing its *softmax* with *sigmoid* or replacing *sigmoid* with *softmax* in SCB result in sub-optimal performance compared to our final BiAM. Similarly, replacing our BiAM with existing Non-Local [33] and Criss-cross [12] attention blocks also results in reduced performance (see Tab. 4 (right)). This shows the efficacy of BiAM, which integrates both region and holistic scene context.

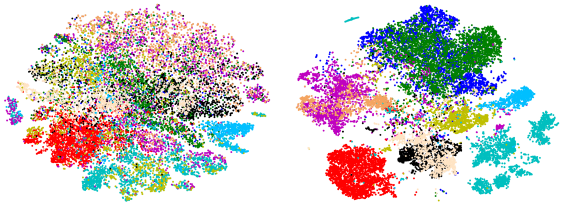

Figure 6. **t-SNE visualization showing the impact of the proposed region-level classification framework on the inter-class feature entanglement.** We present the comparison on 10 unseen classes of NUS-WIDE. On left: the single feature representation-based classification pipeline, where the enriched features are spatially aggregated to obtain a feature vector (of length d_r) and then classified. On right: the proposed region-level classification framework, which classifies the region-level features first and then spatially pools the class scores to obtain image-level predictions. Our classification framework maintains the spatial resolution to preserve the region-level characteristics, thereby effectively minimizing the inter-class feature entanglement.

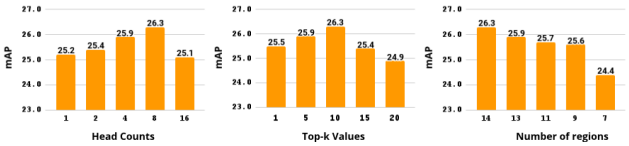

Figure 7. **ZSL comparison on NUS-WIDE when varying H, *top-*k and $h{\cdot}w$ regions**. Results improve slightly as heads H increases till 8 and drops beyond 8, likely due to overfitting to seen classes. A similar trend is observed when *top-*k increases. Decreasing $h{\cdot}w$ regions from 14x14 to 9x9 does not affect much.

Varying the hyperparameters: Fig. 7 shows the ZSL performance of our framework when varying heads H, k in *top-*k and number of regions ($h{\cdot}w$). Performance improves as H is increased till 8 and drops beyond 8, likely due to overfitting to seen classes. Similarly, as *top-*k increases beyond 10, features of spatially-small classes entangle and reduce the discriminability. Furthermore, decreasing the regions leads to multiple classes overlapping in the same regions causing feature entanglement and performance drop.

Compute and run-time complexity: Tab. 5 shows that our approach achieves significant performance gains of 6.7% and 31.3% over LESA with *comparable* FLOPs, memory cost, training and inference run-times, on NUS-WIDE and Open Images, respectively. For a fair comparison, both methods are run on the same Tesla V100.

Additional examples w.r.t. failure cases of our model such as confusing abstract classes (*e.g.*, *sunset vs. sunrise*) and fine-grained classes are provided in the supplementary.

Table 5. **Comparison of our BiAM with LESA in terms of ZSL performance (mAP), train and inference time, FLOPs and memory cost on NUS-WIDE (NUS) and Open Images (OI).** Our BiAM achieves significant gain in performance with comparable compute and run-time complexity, over LESA.

Method	mAP (NUS / OI)	Train (NUS / OI)	Inference	FLOPs	Memory
LESA [10]	19.4 / 41.7	9.1 hrs / 35 hrs	1.4 ms	0.46 G	2.6 GB
BiAM (Ours)	26.1 / 73.0	7.5 hrs / 26 hrs	2.3 ms	0.59 G	2.8 GB

Figure 8. **Qualitative comparison on four test examples from NUS-WIDE, between the standard region features and our discriminative features**. *Top*-3 predictions per image for both approaches are shown with true positives and false positives. Compared to the standard region-based features, our approach learns discriminative region-based features and performs favorably.

3.3. Standard Multi-label Classification

In addition to multi-label (generalized) zero-shot classification, we evaluate our proposed region-based classification framework on the standard multi-label classification task. Here, image instances for all the labels are present in training. The state-of-the-art comparison for the standard multi-label classification on NUS-WIDE with 81 human annotated labels is shown in Tab. 6. Among existing methods, the work of [14] and LESA [13] achieve mAP scores of 32.6 and 31.5, respectively. Our approach outperforms all published methods and achieves a significant gain of 15.2% mAP over the state of the art. Furthermore, our approach performs favorably against existing methods in terms of F1.

4. Related Work

Several works [38, 27, 18, 19, 42, 22, 44] have researched the conventional single-label ZSL problem. In contrast, a few works [20, 46, 17, 13, 11] have investigated the more challenging problem of multi-label ZSL. Mensink *et al*. [20] propose an approach based on using co-occurrence statistics for multi-label ZSL. Zhang *et al*. [46] introduce a method that utilizes linear mappings and non-linear deep networks to approximate principal direction from an input image. The work of [17] investigates incorporating knowledge graphs to reason about relationships between multiple labels. Recently, Huynh and Elhamifar [13] introduce a shared attention-based multi-label ZSL approach, where the shared attentions are label-agnostic and are trained to focus on relevant foreground regions by utilizing a formulation based on multiple loss terms.

Context is known to play a crucial role in several vision problems, such as object recognition [24, 8, 36, 45]. Studies [9, 3] have shown that deep convolutional networks-based visual recognition models implicitly rely on contextual information. Recently, self-attention models have achieved promising performance for machine translation and natural language processing [30, 37, 6, 7]. This has inspired studies to investigate self-attention and related ideas for vision tasks, such as object recognition [26], image synthesis [43] and video prediction [34]. Self-attention strives to learn the relationships between elements of a sequence by estimating the relevance of one item to other items. Motivated by its success in several vision tasks, we introduce a multi-label zero-shot region-based classification approach that utilizes self-attention in the proposed bi-level attention module to reason about all regions together using pair-wise relations between these regions. To complement the self-attentive region features with the holistic scene context information, we integrate a global scene prior which enables us to enrich the region-level features with both region and scene context information.

Table 6. **State-of-the-art performance comparison for the standard multi-label classification on NUS-WIDE**. The results are reported in terms of mAP and F1 score at $K \in \{3, 5\}$. Our proposed approach achieves superior performance compared to existing methods, with gains as high as 15.2% in terms of mAP. Best results are in bold.

Method	mAP	F1 (K = 3)	F1 (K = 5)
WARP [10]	3.1	54.4	49.4
WSABIE [35]	3.1	53.8	49.2
Logistic [29]	21.6	51.1	46.1
Fast0Tag [46]	22.4	53.8	48.6
CNN-RNN [32]	28.3	55.2	50.8
LESA [13]	31.5	58.0	52.0
Attention per Cluster [13]	31.7	56.6	50.7
Attention per Label [14]	32.6	56.8	51.3
Our Approach	**47.8**	**59.6**	**53.4**

5. Conclusion

We proposed a region-based classification framework comprising a bi-level attention module for large-scale multi-label zero-shot learning. The proposed classification framework design preserves the spatial resolution of features to retain the multi-class information disentangled. This enables to effectively deal with large number of co-existing categories in an image. To contextualize and enrich the region features in our classification framework, we introduced a bi-level attention module that incorporates both region and scene context information, generating discriminative feature representations. Our simple but effective approach sets a new state of the art on two large-scale benchmarks and obtains absolute gains as high as 31.9% ZSL mAP, compared to the best published results.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 2015. 6

[2] Avi Ben-Cohen, Nadav Zamir, Emanuel Ben Baruch, Itamar Friedman, and Lihi Zelnik-Manor. Semantic diversity learning for zero-shot multi-label classification. *arXiv preprint arXiv:2105.05926*, 2021. 6

[3] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2019. 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, 5

[6] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019. 8

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 8

[8] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *CVIU*, 2010. 8

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 8

[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] Akshita Gupta, Sanath Narayan, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Joost van de Weijer. Generative multi-label zero-shot learning. *arXiv preprint arXiv:2101.11606*, 2021. 8

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 7

[13] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *CVPR*, 2020. 1, 2, 3, 5, 6, 8

[14] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 6, 8

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

[16] 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. 1, 2, 5

[17] 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, 8

[18] Jingren Liu, Haoyue Bai, Haofeng Zhang, and Li Liu. Near-real feature generative network for generalized zero-shot learning. In *ICME)*, 2021. 8

[19] 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. 8

[20] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 1, 8

[21] 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

[22] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, 2020. 8

[23] 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. 6

[24] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends Cogn Sci*, 7, 2007. 8

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 5

[26] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019. 8

[27] Yuming Shen, Jie Qin, Lei Huang, Li Liu, Fan Zhu, and Ling Shao. Invertible zero-shot recognition flows. In *ECCV*, 2020. 8

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

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

[30] 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. 3, 8

[31] 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. 5

[32] 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

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 7

[34] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 8

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

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 8

[37] Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *ICLR*, 2019. 8

[38] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 8

[39] 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

[40] 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

[41] 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

[42] Haofeng Zhang, Haoyue Bai, Yang Long, Li Liu, and Ling Shao. A plug-in attribute correction module for generalized zero-shot learning. *Pattern Recognition*, 2021. 8

[43] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 8

[44] Haofeng Zhang, Li Liu, Yang Long, Zheng Zhang, and Ling Shao. Deep transductive network for generalized zero shot learning. *Pattern Recognition*, 2020. 8

[45] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *CVPR*, 2020. 8

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