Discriminative Region-based Multi-Label Zero-Shot Learning (Supplementary)

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

In this supplementary, we present additional quantitative and qualitative analysis of our region-based multi-label (generalized) zero-shot approach. The quantitative results are presented in Sec 1 followed by the qualitative analysis in Sec. 2.

1. Additional Quantitative Results

1.1. Standard Multi-Label Learning

Similar to Sec 3.3 of the main paper, where we evaluate our approach for the standard multi-label classification on the NUS-WIDE dataset [2], here, we also evaluate on the large-scale Open Images dataset [5]. Tab. A1 shows the state-of-the-art comparison for the standard multi-label classification on Open Images. Here, $7,186$ classes are used for both training and evaluation. Test samples with missing labels for these $7,186$ classes are removed during evaluation, as in [4]. Due to significantly larger number of labels in Open Images, ranking the labels within an image is more challenging. This is reflected by the lower F1 scores in the table. Among existing methods, Fast0Tag [10] and LESA [4] achieve an F1 score of 13.1 and 14.5 at K=20. Our approach achieves favorable performance against the existing approaches, achieving an F1 score of 17.3 at K=20. The proposed approach also achieves superior performance in terms of mAP score, compared to existing methods and obtains an absolute gain of 35.6% mAP over the best existing method.

1.2. Robustness to Backbone Variation

In the main paper, for a fair comparison with existing works such as Fast0Tag [10] and LESA [4], we employed a pretrained VGG-19 [6] as the backbone for extracting region-level and global-level features of images. However, such supervisedly pretrained backbone will not strictly conform with the zero-shot paradigm if there is any overlap between the unseen classes and the classes used for pretraining. To avoid using a supervisedly pre-trained network, we conduct an experiment by using the recent self-supervised DINO [1] ResNet-50 backbone trained on ImageNet *without any labels*. Tab. A2 shows that our approach (BiAM) significantly outperforms LESA [4] even with a self-supervised pretrained backbone on both benchmarks: NUS-WIDE [2] and Open Images [5]. Absolute gains as high as 6.9% mAP are obtained for NUS-WIDE on the ZSL task. Similar favorable gains are also obtained for the GZSL task on both datasets. These results show that irrespective of the backbone used for extracting the image features, our BiAM approach performs favorably against existing methods, achieving significant gains across different datasets on both ZSL and GZSL tasks.

*Equal contribution

Table A1. **State-of-the-art performance comparison for the standard multi-label classification on Open Images**. The results are reported in terms of mAP and F1 score at $K \in \{10, 20\}$. In comparison to existing approaches, our approach achieves favorable performance in terms of both mAP and F1. Best results are in bold.

Method	mAP	F1 (K = 10)	F1 (K = 20)
WARP [3]	46.0	7.7	7.4
WSABIE [9]	47.2	2.2	2.2
CNN-RNN [8]	41.0	9.6	10.5
Logistic [7]	49.4	13.3	11.8
Fast0Tag [10]	45.4	16.2	13.1
One Attention per Cluster [4]	45.1	16.3	13.0
LESA [4]	45.6	17.8	14.5
Our Approach	**85.0**	**20.4**	**17.3**

Table A2. **ZSL/GZSL performance comparison with LESA on NUS-WIDE and Open Images, when using the recent DINO ResNet-50 backbone** pretrained on ImageNet *without any labels*. Our BiAM outperforms LESA [4] with a large margin on both datasets.

Backbone	Task	NUS-WIDE (mAP)		Open Images (mAP)	
		LESA	**BiAM (Ours)**	LESA	**BiAM (Ours)**
DINO ResNet-50 [1]	ZSL	20.5	**27.4**	41.9	**74.0**
	GZSL	6.4	**10.2**	45.5	**84.8**

2. Additional Qualitative Results

Multi-label zero-shot classification: Fig. 1 shows the qualitative results for multi-label (generalized) zero-shot learning. Nine example images from the test set of

the NUS-WIDE dataset [2] are presented in each figure. The comparison is shown between the standard region-based features and our discriminative region-based features. Alongside each image, *top*-5 predictions for both approaches are shown with true positives and false positives. In general, our approach learns discriminative region-based features and achieves increased true positive predictions along with reduced false positives, compared to the standard region-based features. *E.g.*, categories such as *reflection* and *water* in Fig. 1(b), *ocean* and *sky* in Fig. 1(g), *boat* and *sky* in Fig. 1(j) along with *graveyard* and *england* in Fig. 1(k) are correctly predicted. Both approaches predict a few confusing classes such as *beach* and *surf* in Fig. 1(d) in addition to *sunrise* and *sunset* that are hard to differentiate using visual cues alone in Fig. 1(l). Moreover, false positives that are predicted by the standard region-based features, are reduced by our discriminative region-based features, *e.g.*, *vehicle* in Fig. 1(g), *soccer* in Fig. 1(h), *balloons* in Fig. 1(j), and *ocean* in Fig. 1(k). These results suggest that our approach based on discriminative region features achieves promising performance against the standard features, for multi-label (generalized) zero-shot classification.

Visualization of attention maps: Fig. 2 and 3 show the visualizations of attention maps for the ground truth classes in example test images from NUS-WIDE and Open Images, respectively. Alongside each example, class-specific maps for the unseen classes are shown with the corresponding labels on top. In general, we observe that these maps focus reasonably well on the desired classes. *E.g.*, promising class-specific attention is captured for *zebra* in Fig. 2(a), *vehicle* in Fig. 2(b), *buildings* in Fig. 2(d), *Keelboat* in Fig. 3(c), *Boeing 717* in Fig. 3(e) and *Exercise* in Fig. 3(i). Although we observe that the attention maps of visually similar classes overlap for *sky* and *clouds* in Fig. 2(d), these abstract categories, including *reflection* in Fig. 2(a) and *nighttime* in Fig. 2(c) are well captured. These qualitative results show that our proposed approach generates promising class-specific attention maps, leading to improved multi-label (generalized) zero-shot classification.

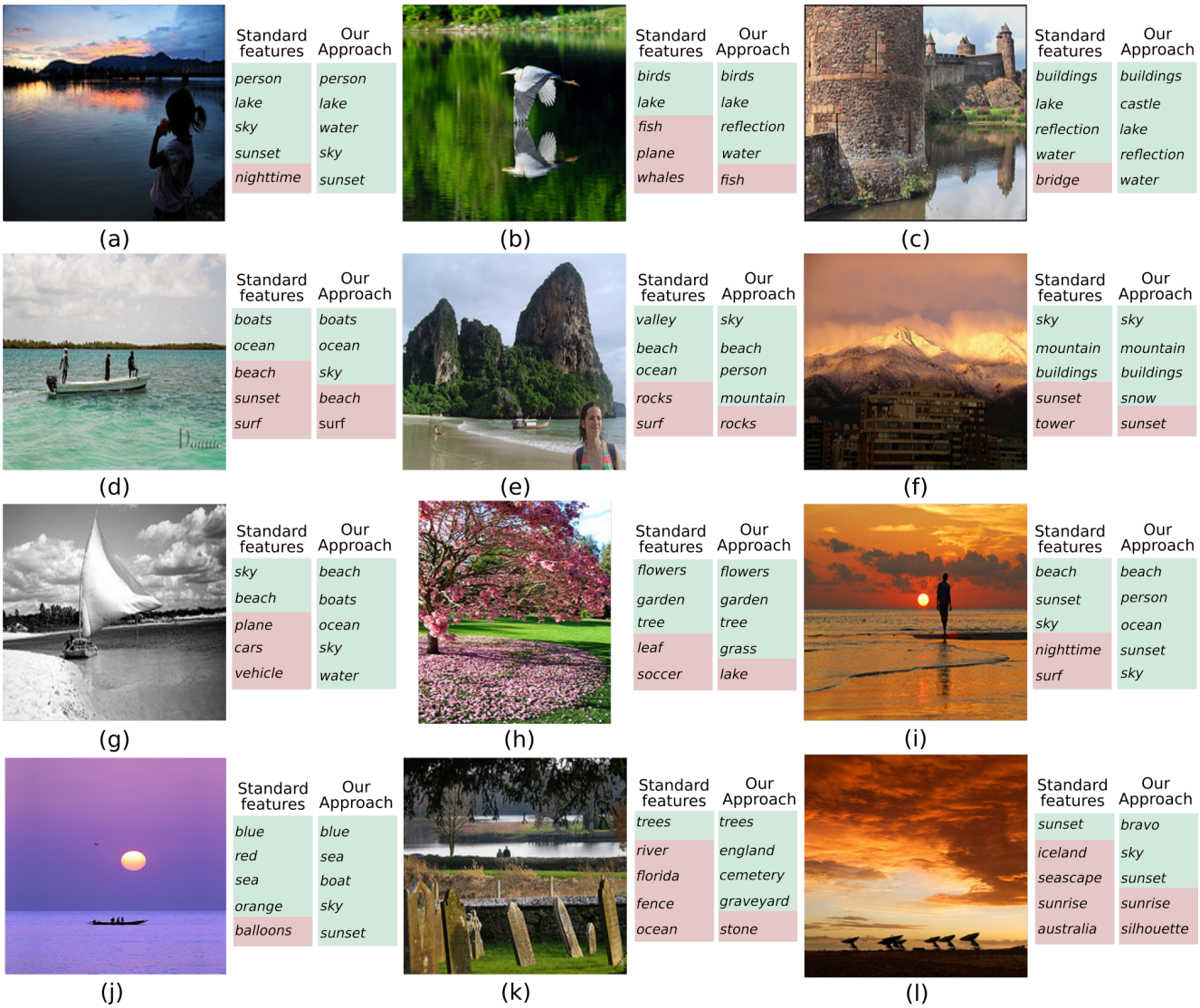

Figure 1. **Qualitative comparison for multi-label zero-shot classification on nine example images from the NUS-WIDE test set, between the standard region-based features and our discriminative features**. *Top*-5 predictions per image for both approaches are shown with true positives and false positives. Generally, in comparison to the standard region-based features, our approach learns discriminative region-based features and results in increased true positive predictions along with reduced false positives. *E.g.*, *reflection* and *water* in (b), *ocean* and *sky* in (g), *boat* and *sky* in (j) along with *graveyard* and *england* in (k) are correctly predicted. Though a few confusing classes are predicted (*e.g.*, *beach* and *surf* in (d)), the obvious false positives such as *vehicle* in (g), *soccer* in (h), *balloons* in (j) and *ocean* in (k) which are predicted by the standard region-based features, are reduced by our discriminative region-based features. These qualitative results suggest that our approach based on discriminative region features achieves promising performance in comparison to the standard features, for the task of multi-label (generalized) zero-shot classification.

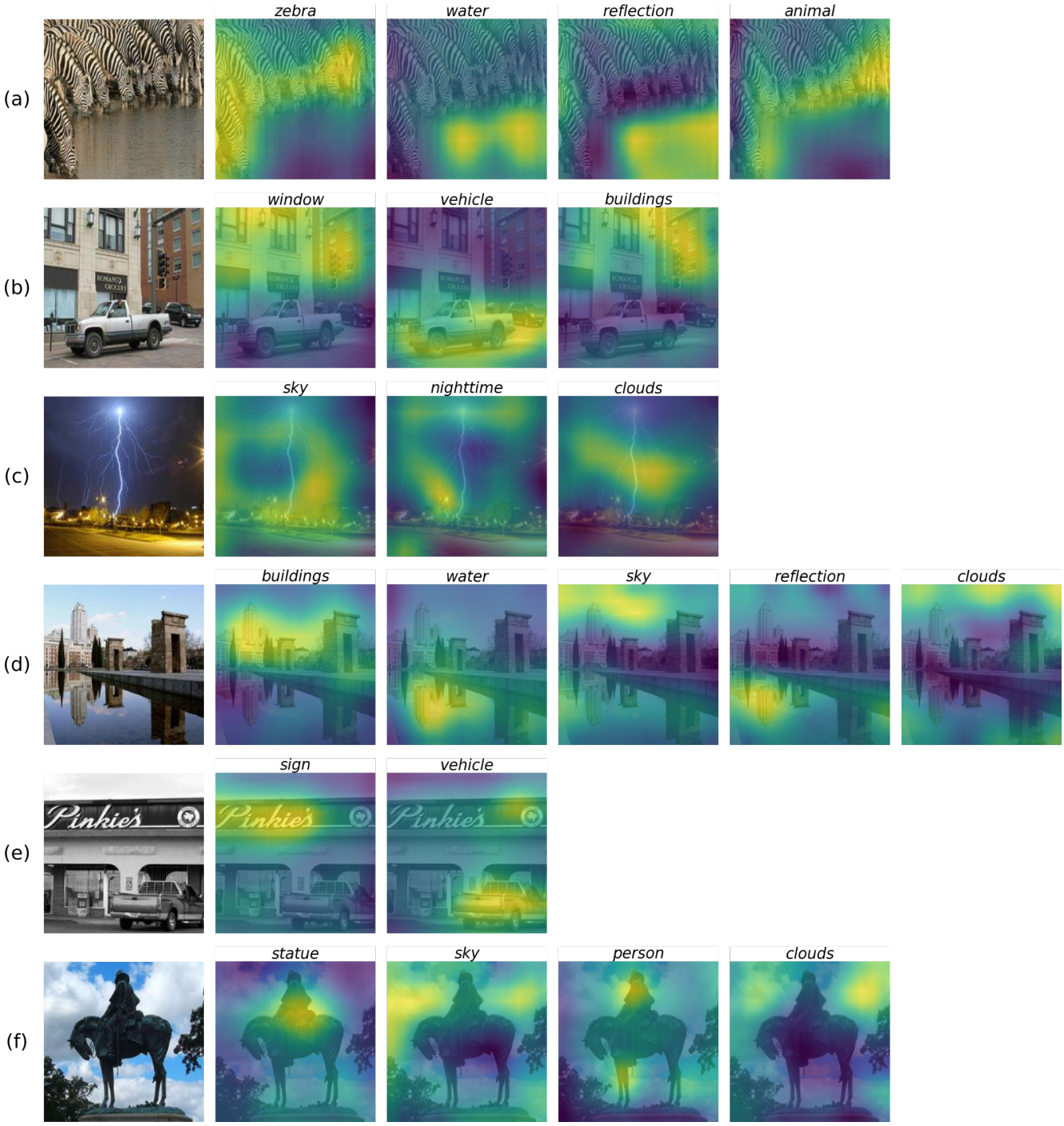

Figure 2. **Qualitative results with attention maps generated by our proposed approach, on example test images from the NUS-WIDE [2] dataset**. For each image, class-specific maps for the ground truth unseen classes are shown with the corresponding labels on top. Generally, we observe that these maps focus reasonably well on the desired classes. *E.g.*, promising attention/focus is observed on classes such as *zebra* in (a), *vehicle* in (b), *buildings* in (d) and *statue* in (f). Although we observe that the attention maps of visually similar classes such as *sky* and *clouds* overlap, as in (d), these abstract classes, including *reflection* in (a), (d) and *nighttime* in (c) are well captured. These qualitative results show that our proposed approach generates promising class-specific attention maps, leading to improved multi-label (generalized) zero-shot classification.

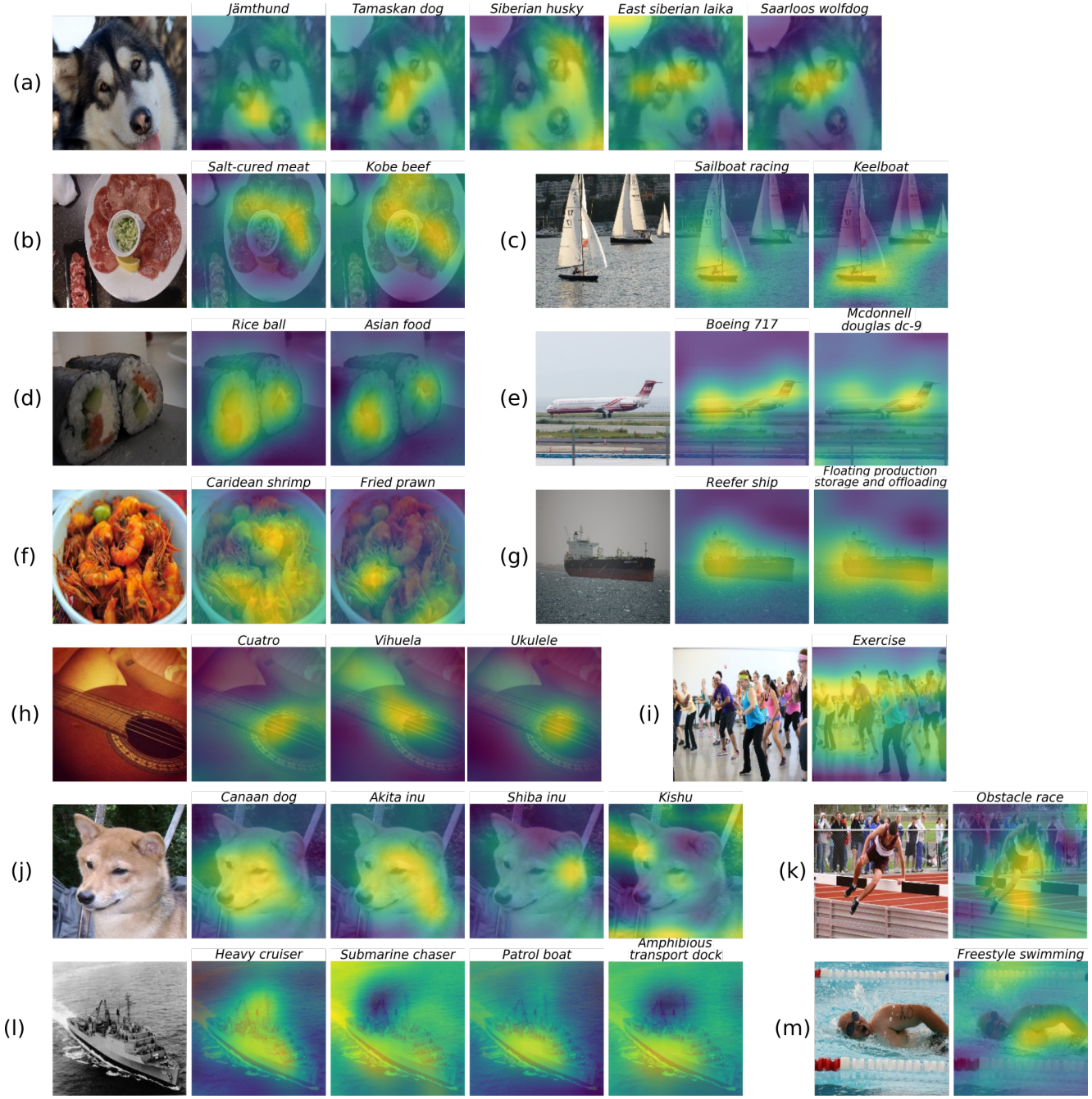

Figure 3. **Qualitative results with attention maps generated by our proposed approach, on example test images from the Open Images [5] dataset**. For each image, class-specific maps for the ground truth unseen classes are shown with the corresponding labels on top. Although there are overlapping attention regions for visually similar and fine-grained classes (*e.g.*, *Caridean shrimp* and *Fried prawn* in (f), *Canaan dog* and *Akita inu* in (j)), generally, these maps focus reasonably well on the desired classes. *E.g.*, promising class-specific attention is captured for *Keelboat* in (c), *Boeing 717* in (e) and *Exercise* in (i). These qualitative results show that our proposed approach generates promising class-specific attention maps, resulting in improved multi-label (generalized) zero-shot classification.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1

[2] 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, 4

[3] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. 1

[4] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *CVPR*, 2020. 1

[5] 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, 5

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[7] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *IJDWM*, 2007. 1

[8] 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

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

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