Photozilla: An Image Dataset of Photography Styles and its Application to Visual Embedding and Style Detection

Trisha Singhal, Junhua Liu, Wenchuan Mu, Lucienne T. M. Blessing and Kwan Hui Lim
Singapore University of Technology and Design
{trisha_singhal, junhua_liu wenchuan_mu, lucienne_blessing, kwanhui_lim}@sutd.edu.sg

***Abstract*—The widespread sharing of digital photography and images have led to the rapid development of various vision-related applications, such as photography style detection. Towards this effort, we introduce a photography style dataset termed Photozilla, which comprises over 990k images belonging to 10 different photographic styles. We used Photozilla to train 3 classification models for categorizing images into the relevant style and achieve an accuracy of $\sim$96%. To better detect new photography styles that are constantly emerging, we also present a Siamese-based network that uses the trained classification models as the base architecture to adapt and classify unseen styles with only 25 training samples. Experiment results show an accuracy of over 68% in terms of identifying 10 additional distinct categories of photography styles. This dataset can be found at https://trisha025.github.io/Photozilla/.**

***Index Terms*—Image Recognition, Visual Embedding, Image Classification, Neural Networks**

I. Introduction

The popularity of social media and photo-sharing platforms, such as Instagram and Flickr, have contributed to an abundance and variety of digital photography. For instance, more than 50B photos have been uploaded to these platforms and the upload is at a rate of 995 photos every second [1]. This large trove of imagery data and its rapid proliferation have facilitated numerous computer vision applications. As photography is evolving with new styles being rapidly introduced, any model needs to be able to rapidly adapt to identify these new styles. Conventional deep learning models are effective in image detection but require a large amount of training data to achieve a high level of accuracy. To overcome this issue, we utilize few-shot learning techniques to better enable our models to rapidly adapt to new photography styles with only a few samples of the new style. More specifically, we implemented few shot learning using a Siamese network where our classification models are used as base architecture to quickly adapt to new styles with only 25 training samples.

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
http://dx.doi.org/10.1145/3625007.3627476

In this paper, we make the following contributions:[1]

1) We curated and share a large-scale photography dataset, denoted *Photozilla*, comprising over 990k images comprising 10 different photography styles, with 10 additional photography styles with 25 images per class for few shot learning applications.
2) For image style detection, we implemented three state-of-the-art image classification models and experimental results show an accuracy of $\sim$96% in terms of the classification of these 10 photography styles.
3) For new style detection, we combined the previous pre-trained model in a Siamese network, and demonstrated how it is able to classify new photography styles even with a low number of samples. Experimental results show an accuracy of $\sim$68% for this task.

II. Related Works

Photography Datasets. There are various photography related datasets, such as the Aesthetic Visual Analysis (AVA) dataset [3], comprising $\sim$250k+ images along with aesthetic scores and labels for 14 photographic styles based on light, color, and composition. Karayev et al. [4] shared two datasets, namely a dataset of $80k$ photographs from Flickr with corresponding style annotation for each image, and another dataset of 85k paintings with 25 styles/genre annotation. One objective of these works is to highlight the role image style plays in the age of digital photo overload. In addition to differing in terms of the photography styles in the dataset, our Photozilla dataset is larger in scale with close to 1M images covering 20 photography styles. There are also datasets covering other different genres, such as painting styles [5], food photography [6] scenery photography [7], aerial images [8], and street-level images [9].

Image Classification. ImageNet [10], comprising 14M+ annotated images with 20k+ object categories, is one of the first large-scale image datasets tasks, alongside others like the COCO dataset [11] with 328K images of 2.5M labeled instances of 80 common objects. Various models were developed and trained on these datasets, such as the $152-$layered network, Residual Network (ResNet) [12] which introduced

[1]This work was previously presented as a poster in a workshop with no formal proceedings [2].

Fig. 1: Samples images from 10 classes of Photozilla dataset

the novel idea of skip connections to address the vanishing gradients problem. Another popular variant is DenseNet [13] where the concept of extra connections was introduced to resolve the same problem. All the layers with identical feature-maps in the network are directly interconnected with each other so that maximum information flow will take place. In our work, we build upon various other competitive models, such as Wide ResNet [14], ResNext [15], and EfficientNet [16], which we discuss further in Section IV-A.

Classification Similarity Learning. Similarity-based classification takes as input a pair of images and predicts a similarity score for that image pair, instead of classifying an image directly to a specific target class. For this purpose, Siamese Neural Networks [17] were introduced to train a network to learn these similarity scores. In our work, we also utilise the idea of Siamese Neural Networks for the task of new style detection using few images of the new photography style. We discuss more about this method later in Section IV-B1. Siamese networks can be further used to perform low-shot learning in which a limited number of samples are used to train the model. Some examples are zero-shot [18], one-shot [19], and few-shot learning (FSL) in which zero, one, and a few training samples are provided respectively. Similarly, more details are subsequently provided in Section IV-B.

III. Dataset Collection

To curate our dataset, we used the Flickr API [20] to collect a large-scale image dataset from Flickr comprising images belonging to 10 photography styles. The images of specific styles were collected using tags corresponding the the name of the various photography styles. For example, to collect images in travel photography, *'travel'* was used as the specific tag. Each class contains approximately $\sim$100k images. Figure 1 shows samples of each photography style.

To facilitate the task of few-shot new style detection, we further expanded our dataset with 10 additional classes but with a limited number of 25 images for each class. This expanded dataset was then used for our evaluations of few-shot learning of unseen photography classes.

IV. Methodology

A. Classification Models

We build upon three competitive classification models, i.e. Wide ResNet [14], ResNext [15], and EfficientNet [16], for training the baseline visual embeddings on the 10 classes with a larger number of data points. We performed an empirical experimentation and used identical hyper-parameters for all classification models, which are using Stochastic Gradient Descent as the optimizer, Cross Entropy Loss as the loss function, learning rate of 0.01, and batch size of 64. The results of the various classification models in terms of accuracy on the curated dataset are shown in Table I.

*1) **Wide ResNet**:* Early research on neural networks observed that performance would improve when more layers were stacked up but this performance improvement appear to stagnate after a certain point. Deep Residual Networks, such as ResNet [12], were introduced to overcome this problem of performance stagnation. These deeper neural networks were able to achieve superior performance while being able to scale up to thousands of layers. One of the main innovation in these architectures is the utilisation of residual blocks, which can be represented by the following equation:

$$x_{l+1} = x_l + \mathcal{F}(x_l, \mathcal{W}_l) \tag{1}$$

Here, x_l, $\mathcal{W}_l$ and x_{l+1} are the input, weights and the output of the l^{th} layer, respectively. $\mathcal{F}$ is the residual function governed by the architecture of the residual block.

Although ResNet and its variants were able to achieve superior performances in comparison to its shallower counterparts, these performance gains came at the cost of increased training and inference times. Wide Residual Networks (WRNs) [14] were then introduced to address this issue of lengthy training and inference time. WRNs have a similar model architecture as ResNet except for the increased number of feature maps and shallower architecture. Zagoruyko et. al. have also explored and noticed the benefits in increasing the width of the network by a hyper-parameter k, instead of implementing a deeper architecture.

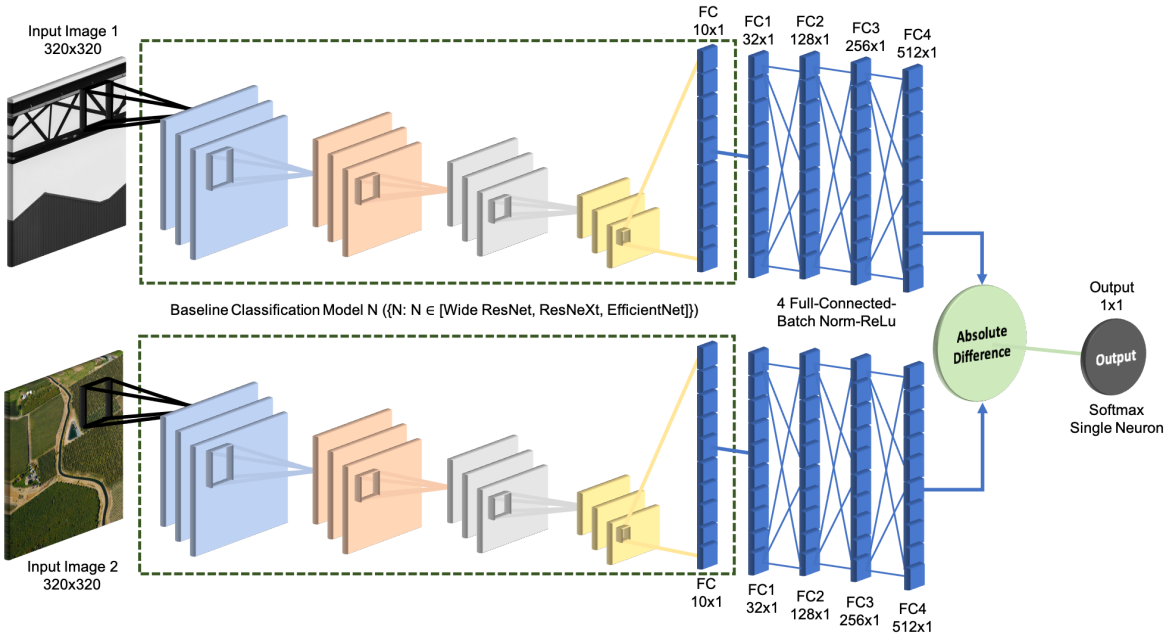


Fig. 2: Proposed Siamese-based neural network architecture.

2) ***ResNeXt****:* In contrast to using a deeper and wider architecture, ResNeXt [15] utilizes the idea of incorporating higher cardinality, which is a new dimension that is aimed at improving the performance of networks with less complex architectures. The cardinality, C of a model can be defined as the number of branches in a residual block to control more complex transformations. Mathematically, these C transformations are formulated as follows:

$$\mathcal{F}(x_l, W_l) = \sum_{i=1}^{C} \mathcal{T}_i(x_l, {W_l}^i) \tag{2}$$

Here, T_i is the transformation function for the branch i of the residual block. The aggregated transformation is the sum of all C branches. This aggregated transformation is then used as the residual function similar to equation 1.

3) ***EfficientNet****:* EfficientNet [16] uses a novel scaling method that considers three dimensions of a neural network, namely its depth, width, and resolution. EfficientNet performs compound scaling by combining the scaling of all three dimensions to optimize for accuracy while satisfying the memory and computational constraints. To reduce the design space, all layers must be scaled uniformly with a constant ratio. Let d, w and r be these constant ratios for depth, width and resolution, respectively. Let $\mathcal{N}(d, w, r)$ be the resulting neural network with these ratios. EfficientNet proposes the following optimization function to find optimal d, w and r:

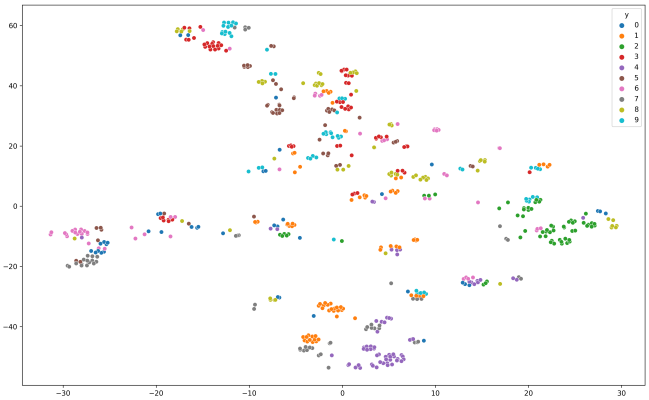
$$\begin{aligned} &\max_{d,w,r} Accuracy(\mathcal{N}(d,w,r)) \\ &Memory(\mathcal{N}) \leq target\_memory \\ &FLOPS(\mathcal{N}) \leq target\_flops \end{aligned} \tag{3}$$

B. Few-Shot Learning

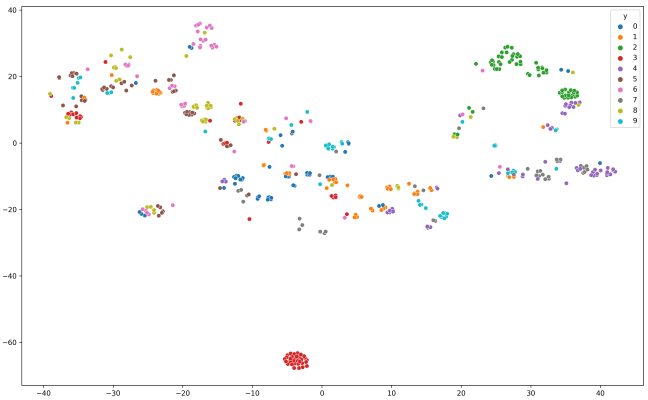
Existing deep neural network architectures typically contain billions of parameters. As a result, the training of such deep neural network architectures requires a large training set, which in turn is often challenging due to the efforts and resources involved in data collection and annotation of the ground truth. To address these limitations, few-shot learning was developed as a class of techniques that allows a deep neural network to learn with a smaller number of training samples. In the context of this work, the domain of photography we study is a rapidly evolving industry. Numerous styles of photography have evolved over time and continuously evolve, thus it is infeasible to modify our base classification model to adapt to a newer photography style. Instead, we use the Siamese network architecture to quickly adapt our base classification model to unseen classes of photography even with the limitations of a low number of annotated training samples.

1) ***Siamese Network****:* Siamese networks [17] are a class of deep neural networks that contains two identical sub-networks. Here, identical means that the sub-networks have the same architecture, parameters, and weights. Traditionally, classification networks learn to classify a training sample into multiple classes. Siamese networks, however, learn a deep similarity function that takes two different inputs and computes whether the inputs belong to the same class or different classes. The two identical sub-networks each take one input and compute a deep feature embedding for that input. This deep feature embedding is then used to compute a similarity score to determine whether the two inputs are from the same class or two different classes.

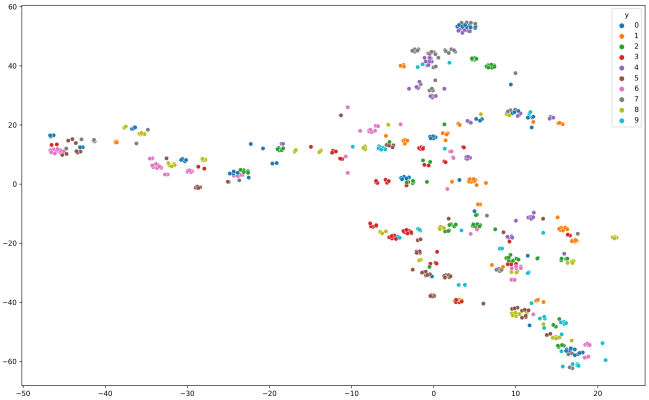
As mentioned in the previous section, we trained three



(a)



(b)



(c)

Fig. 3: Clusters Visualization using t-SNE for: (a) Wide ResNet; (b) ResNeXt; and (c) EfficientNet. Here, 'y' represents the different photography styles.

popular and competitive classification models for categorizing the style of an image into one of the 10 photography styles. We use the output of the last fully connected layer of these base classification models and then stack 4 additional fully connected layers to compute the 512-dimensional deep feature embedding. Finally, we compute the similarity score by computing the absolute difference between the two deep feature embeddings from the Siamese network. Let I_1 and I_2 be two images for which we measure the similarity score. Let $\mathcal{N}(I)$ be the 512-dim visual feature embedding output for image I. Then, we compute the similarity score $P(I_1, I_2)$ as follows.

$$P(I_1, I_2) = Softmax(W_{out} \cdot |\mathcal{N}(I_1) - \mathcal{N}(I_2)| + B_{out})$$
$$Where \;\; W_{out} \in \mathcal{R}^{512\times 1} \;\; and \;\; B_{out} \in \mathcal{R}^{1\times 1} \tag{4}$$

Furthermore, we use the cross-entropy loss in our Siamese network and a learning rate of 0.05 with SGD [21] as our optimizer to train for 30 epochs with each class having only 25 training, validation, and test samples each. Figure 2 shows an illustration of the proposed model architecture.

V. Experimental Results

A. *Classification Accuracy*

As previously described in Section IV-A, we used 10 photography classes to evaluate the classification accuracy of our dataset on 3 popular and competitive models (see Table I). We used 70% of the dataset for training and the remaining 30% for testing. All three models achieve over 96% accuracy on the test dataset. ResNeXt marginally outperforms the other two with an accuracy of 96.35%.

TABLE I: Model Performance on Standard (Full Dataset) and Few-shot Settings

Model	Accuracy (%) (Standard)	Accuracy (%) (Few-shot)
Wide ResNet	96.23	64.17
ResNeXt	96.35	68.34
Efficient Net	95.71	56.25

B. *10-Way Few-Shot Evaluation Metric*

We further extracted 75 images each for the 10 additional photography style classes. These additional classes were used for evaluating the performance of our proposed Siamese network for few-shot learning of new photography styles. Out of these 75 images, we used 25 images each for training, validation, and testing respectively.

To evaluate the test dataset, we used the 10-way few-shot evaluation metric. In this evaluation task, we take one image Q belonging to class c as the query image, and randomly pick 10 more images from each class. Let I_j be the randomly picked image for j^{th} class. For the pair of images Q and I_j, the Siamese network predicts a similarity probability $P(Q, I_j)$, which defines the similarity between two images. We run the Siamese network for image Q and $I_j \forall j \in [1, 10]$ and select the image with the highest $P(Q, I_j)$. A prediction is said to be correct if the following criterion is met:

$$c = argmax_{j\in[1,10]}(P(Q, I_j)) \tag{5}$$

Table I shows the results in terms of accuracy for the 10-way few-shot evaluation for various Siamese networks with different base classification models. In summary, the Siamese network with ResNeXt performs better than the other two variants (68.34%). EfficientNet is a popular model for image classification but surprisingly, EfficientNet only gives an accuracy of 56.25%. The other two variants give an accuracy score that is $8 - 12\%$ higher.

C. Qualitative Analysis of Visual Embedding for clustering

To further analyze the capability of our Siamese network to classify unseen photography styles with only a few samples, we use t-Distributed Stochastic Neighbor Embedding (t-SNE) [22]. t-SNE is a non-linear and unsupervised dimensionality reduction approach for visualizing high-dimensional data. Intuitively, t-SNE allows one to visualize how the data is arranged in the higher dimensional space. We use the 512-dimension feature embedding output as an input to generate the t-SNE transform. Figure 3 shows the t-SNE plots to visualize the 512-dim visual embedding arranged in a 2-dim space. Based on this, we can observe that the Siamese variant of ResNeXt and Wide ResNet generate better clusters for images of the same photography style. However, we do not observe any distinct clusters for EfficientNet. This is also evident in the performance comparison of Siamese networks where ResNeXt and Wide ResNet perform comparatively better than EfficientNet's Siamese network.

VI. Conclusion

In this work, we curated and shared a large-scale dataset termed Photozilla comprising over 990k images belonging to 10 photography styles. We then used this dataset for the task of photography style detection and train 3 different classification model architectures to automatically identify the photography style. These models achieved a strong performance of over 96% accuracy on our testing dataset. Digital photography is a rapidly evolving field, which requires that our models can adapt quickly to identify new photography styles. To facilitate this, we propose a novel Siamese network that learns from our base classification networks. Our proposed Siamese network achieves an accuracy of over 68% on identifying 10 new photography styles with merely 25 training samples.

For future work, we can explore the use of more advanced models, such as the Transformer architecture [23], which has shown superior performance in a variety of domains from Natural Language Processing [24], [25] to Recommendation Systems [26], [27]. For our task, we can explore the use of Vision Transformer [28] and its variants to replace the image representation models in our Siamese-based architecture.

Acknowledgment. This research is funded in part by the Singapore University of Technology and Design under grant RS-MEFAI-00005-R0201.

References

[1] Omnicore agency, https://www.omnicoreagency.com/instagram-statistics/.
[2] T. Singhal, J. Liu, L. T. Blessing, and K. H. Lim, "Photozilla: A Large-Scale Photography Dataset and Visual Embedding for 20 Photography Styles," in *WiCV*, 2021.
[3] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *CVPR*, 2012, pp. 2408–2415.
[4] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing image style," *arXiv:1311.3715*, 2013.
[5] C. Zhang, C. Kaeser-Chen, G. Vesom, J. Choi, M. Kessler, and S. Belongie, "The imet collection 2019 challenge dataset," *arXiv:1906.00901*, 2019.
[6] D. Sahoo, W. Hao, S. Ke, W. Xiongwei, H. Le, P. Achananuparp, E.-P. Lim, and S. C. Hoi, "Foodai: Food image recognition via deep learning for smart food logging," in *KDD*, 2019, pp. 2260–2268.
[7] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *CVPR*, 2012, pp. 2751–2758.
[8] L. Chen, F. Liu, Y. Zhao, W. Wang, X. Yuan, and J. Zhu, "Valid: A comprehensive virtual aerial image dataset," in *ICRA*, 2020, pp. 2009–2016.
[9] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017, pp. 4990–4999.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.
[14] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv:1605.07146*, 2016.
[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017, pp. 1492–1500.
[16] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114.
[17] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Workshop*, vol. 2, 2015.
[18] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks." in *AAAI*, vol. 1, no. 2, 2008, p. 3.
[19] E. G. Miller, N. E. Matsakis, and P. A. Viola, "Learning from one example through shared densities on transforms," in *CVPR*, vol. 1, 2000, pp. 464–471.
[20] Flickr, https://www.flickr.com/.
[21] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv:1609.04747*, 2016.
[22] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *JMLR*, vol. 9, no. 11, 2008.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.
[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.
[25] M. Li, K. H. Lim, T. Guo, and J. Liu, "A transformer-based framework for poi-level social post geolocation," in *ECIR*, 2023, pp. 588–604.
[26] X. Q. Ong and K. H. Lim, "Skillrec: A data-driven approach to job skill recommendation for career insights," in *ICCAE*, 2023, pp. 40–44.
[27] S. Halder, K. H. Lim, J. Chan, and X. Zhang, "Capacity-aware fair poi recommendation combining transformer neural networks and resource allocation policy," *Applied Soft Computing*, vol. 147, p. 110720, 2023.
[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.