Photozilla: A Large-Scale Photography Dataset and Visual Embedding for 20 Photography Styles

Trisha Singhal[1]
trisha_singhal@sutd.edu.sg

Junhua Liu[1,2]
junhua_liu@mymail.sutd.edu.sg, j@forth.ai

Lucienne T.M. Blessing[1]
lucienne_blessing@sutd.edu.sg

Kwan Hui Lim[1]
kwanhui_lim@sutd.edu.sg

[1]Singapore University of Technology and Design, Singapore
[2]Forth AI, Singapore

Abstract

The advent of social media platforms has been a catalyst for the development of digital photography that engendered a boom in vision applications. With this motivation, we introduce a large-scale dataset termed 'Photozilla', which includes over $990k$ *images belonging to* 10 *different photographic styles. The dataset is then used to train* 3 *classification models to automatically classify the images into the relevant style which resulted in an accuracy of* $\sim 96\%$*. With the rapid evolution of digital photography, we have seen new types of photography styles emerging at an exponential rate. On that account, we present a novel Siamese-based network that uses the trained classification models as the base architecture to adapt and classify unseen styles with only* 25 *training samples. We report an accuracy of over* 68% *for identifying* 10 *other distinct types of photography styles.* ***This dataset can be found at https://trisha025.github.io/Photozilla/***

1. Introduction

"*A picture is worth a thousand words.*" - *Henrik Ibsen*

Photography has become an integral part of people's everyday life for both professional and recreational purposes. Professionally, businesses use photographs for marketing and advertising purposes and for sharing news about significant political events. For recreational purposes, people use photographs to signify a purpose, such as telling a story, recording an event and recording moments of memory, and with the advent of social media to share their lives. Figure 1 shows the award-winning photos of 2020[1] from different realms of life. The era of Instagram and Flickr has seen an unprecedented overload of digital photography. According to the latest statistics shared by Omnicore Agency, $50B+$ photos have been uploaded on the platform and 995 photos are uploaded every other second[2], $350M$ photos are uploaded every day on Facebook[3], and a decade is needed to view all photos on Snapchat[4]. Such an upsurge in the proliferation of data has given rise to various computer vision applications.

Image classification and recognition have revolutionized the data visualization happening digitally. The ability to automatically identify objects has led to the development of various applications like image tagging, organization, categorization, behavioral analysis, recommendation systems, and so on. A dataset of images with various photographic styles is useful to carry out style-based image retrieval that can narrow-down image search, image recommendation based on preferences, generation of royalty-free synthetic images to avoid copyright restrictions, and many more applications.

In this paper, we present *Photozilla*, a large-scale dataset of $990k$ million images comprising 10 different photography styles. To demonstrate the usefulness of this dataset, we propose 3 photo-style classification models that outperform state-of-the-art classifiers and achieve over $\sim 96\%$ accuracy for predicting the correct photography style.

[1] https://www.photoawards.com/winner/?compName=IPA+2020
[2] https://www.omnicoreagency.com/instagram-statistics/
[3] https://www.omnicoreagency.com/facebook-statistics/
[4] https://www.omnicoreagency.com/snapchat-statistics/

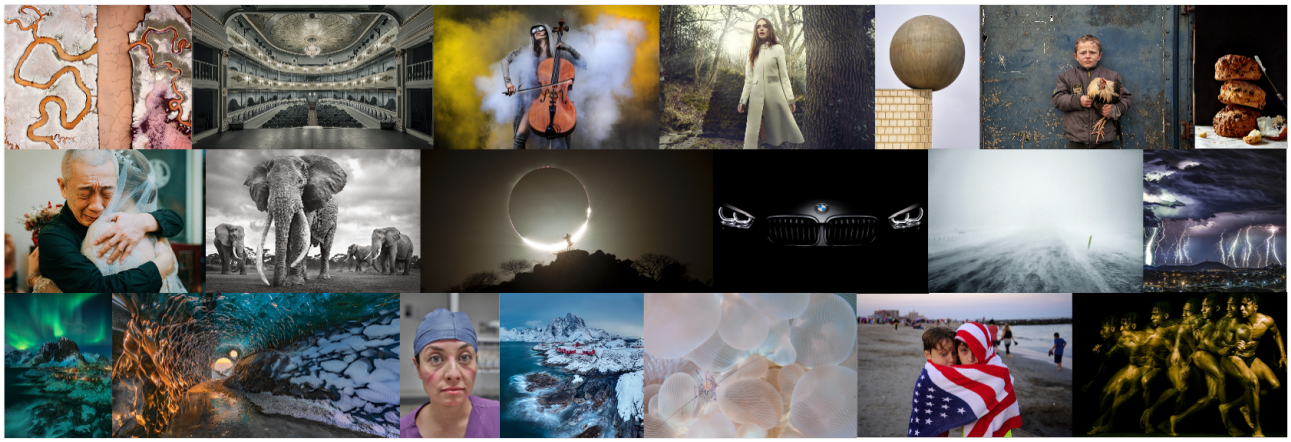


Figure 1: Awarded photos in different photographic styles.
(Source: International Photography Awards, 2020)

While photography is a rapidly evolving subject with many new styles being introduced over the years, our models must be able to quickly adapt to identify these new styles. The conventional deep learning classification models require large data sets to achieve a high level of accuracy. To overcome this issue and to enable our models to be able to quickly adapt to a new photography style even with a few samples, we propose to use few-shot learning. We implemented a Siamese network where we use our classification models as base architecture to quickly adapt to new styles with only 25 training samples.

Henceforth, the following are the contributions from this work:

1. We present a large-scale photography dataset comprising over $990k$ images under 10 different photography styles.
2. By using three state-of-the-art classification models, we achieve an accuracy of $\sim$ 96% on classifying to 10 these styles.
3. We further extend our dataset with 10 additional photography styles but with a small number of samples (25 per class). By using the pre-trained $10-$class classification model and a Siamese architecture, we show that our model can learn to classify new photography styles even with even a low number of samples. We achieve an accuracy of $\sim$ 68% for the same.

The remainder of the paper is organised as follows. Section 2 presents a review of the literature. Section 3 discusses the procedure of dataset accumulation. Section 4 describes the methodology in detail, while experimental results are shown in Section 5. The future work is discussed in Section 6 and concluding remarks are given in Section 7.

2. Related Works

With the breakthroughs in Deep Learning technology, photography has turned out to be an area of constant interest in the research community. From analyzing images for interpretation to synthetic image generation, an immense amount of applications has been developed. With these applications, many researchers have published various kinds of datasets in photography for specific uses and developed many state-of-the-art computer vision techniques. Some of the notable works are mentioned below.

Photography Datasets. [17] published a large-scale dataset, AVA, to perform Aesthetic Visual Analysis (AVA) with $\sim$ $250k+$ images containing aesthetic scores of all data points, semantic labels for over 60 categories, and labels for 14 photographic styles based on light, color, and composition. [8] created two datasets: (a) $80k$ photographs from Flickr [2] with corresponding style annotation, and (b) $85k$ paintings with 25 styles/genre annotation in order to predict image style and aesthetic standard. The photographic styles were categorized based on photographic technique, composition style, mood, genre, and type of scenes. The work aimed to show the significance of an image style in the age of digital photo-overload.

Some researchers have worked in specific genres, like [9], who proposed a dataset of digital paintings with more than $4.2k$ paintings from 91 painters having two annotation labels: artist name and style. [25] focused on contemporary artwork and featured an artistic imagery dataset, BAM, collected from the Behance website. The collected dataset was annotated with labels for content, emotions, and artistic media. Other notable datasets that primarily emphasized artworks are [1, 29].

Food imagery has gained substantial attention because of

its multi-purpose applications including health, food selection and recommendation, culture, and so on [16]. With this motivation, [22] presented a Gourmet Photography Dataset (GPD) having $12k$ photographs to assess the aesthetics in food images. [21] introduced a large-scale food image dataset with 152 categories consisting of generic types of food and 756 visual food items constituting a total of ∼400k+ images.

Some other genre-specific datasets include scenery database [19], aerial images dataset [26, 3], and street-level images dataset [18].

Image Classification. A considerable increase in the popularity of image classification tasks can be noticed since the introduction of the revolutionary ImageNet dataset [4]. Developed with the aim of visual object recognition, ImageNet at present contains $14M+$ annotated images of $20k+$ specific object categories with bounding boxes in $1M+$ images. Another large-scale visual dataset is the COCO (Common Object in Context) dataset [12]. It comprises instance segmentation of 80 common objects. The dataset is containing $328K$ images of $2.5M$ labeled instances.

The aforementioned datasets engendered state-of-the-art neural network architectures for image classification. A $152-$layered network, ResNet (Residual Network) [5] won the 2015 ImageNet challenge with the introduction of a novel concept of skip connection. These connections were designed to solve the problem of vanishing gradients caused by skipping more than one layer in the network and hence, use the output of one layer as the input for others. This enables the easier flow of the gradient from layer to layer. One of the popular variants of ResNet is DenseNet [7] 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.

The classification models we use, [28, 27, 23] are discussed in Section 4.1.

Classification Similarity Learning. Similarity-based classification takes pairwise input images and predicts a similarity score, instead of classifying an image directly as any target class. This score can be in a form of binary value i.e. 0 or 1, or any real number. To train a network to learn similarity, Siamese Neural Networks [10] were introduced. More details can be found in 4.2.1. 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 [11], one-shot [15], and few-shot learning (FSL) in which zero, one, and a few training samples are provided respectively. More details can be found in 4.2.

3. Dataset Collection

For our model evaluation, we first collected a large-scale dataset of Flickr[5] images using Flickr API to automatically extract the images belonging to the 10 photography styles. These images of specific styles were collected using tags of the same. For example, to crawl images in travel photography, *'travel'* was used as the specific tag. Each class of 10 has approximately $\sim 100k$ number of images. Figure 2 shows samples of each photography style.

Figure 2: Samples images from 10 classes of the training dataset for classification models.

We extended our dataset with 10 more classes to our dataset, as shown in figure 3, but with only a limited number of data samples i.e. 25 per class. This extended dataset was then used for our evaluations of few-shot learning of unseen photography classes.

4. Methodology

4.1. Classification Models

In our proposed work, three state-of-the-art classification models are used i.e. Wide ResNet [28], ResNext [27],

[5] https://www.flickr.com/

Figure 3: Samples images from Siamese network' testing dataset of 10 classes.

and EfficientNet [23] to train the baseline visual embeddings with the 10 classes with a larger number of datapoints. For all classification models, we have used identical hyperparameters as listed in Table 1. The classification accuracies on the curated dataset are shown in the Table 2.

Table 1: Hyperparameters used for classification models.

Hyparameter	Value
No. of Epochs	1
Optimizer	Stochastic Gradient Descent
Learning Rate	0.01
Momentum	0.9
Batch Size	64
Loss Function	Cross Entropy Loss

4.1.1 Wide ResNet

In the earlier work on Deep Neural Networks, performance would improve when more layers were stacked up. However, performance gains appear to saturate beyond a certain point. To mitigate this issue of performance saturation, Deep Residual Networks (ResNet) [5] were introduced. These deeper networks achieved superior performance while being able to scale up to thousands of layers. The key intuition of such architectures is the usage of residual blocks, which is primarily 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 architectures were able to achieve superior performances in comparison to the shallower models, these performance gains came at the cost of increased training and inference times. To overcome these issues, Wide Residual Networks (WRNs) [28] were introduced. WRNs have a nearly similar model architecture as ResNet except for the increased number of feature maps and shallower architecture. They have explored the benefit of increasing the width of the network by a hyper-parameter k instead of a deeper architecture.

4.1.2 ResNeXt

Instead of having higher depth and width, ResNext [27] uses the concept of having higher cardinality, a new dimension introduced to improve 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 aggregrated transformation is then used as the residual function similar to equation 1.

4.1.3 EfficientNet

EfficientNet [23] uses a novel scaling method that considers three dimensions of a neural network: depth, width, and resolution. EfficientNet performs compound scaling that combines the scaling of all three dimensions to optimize for accuracy while meeting 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.

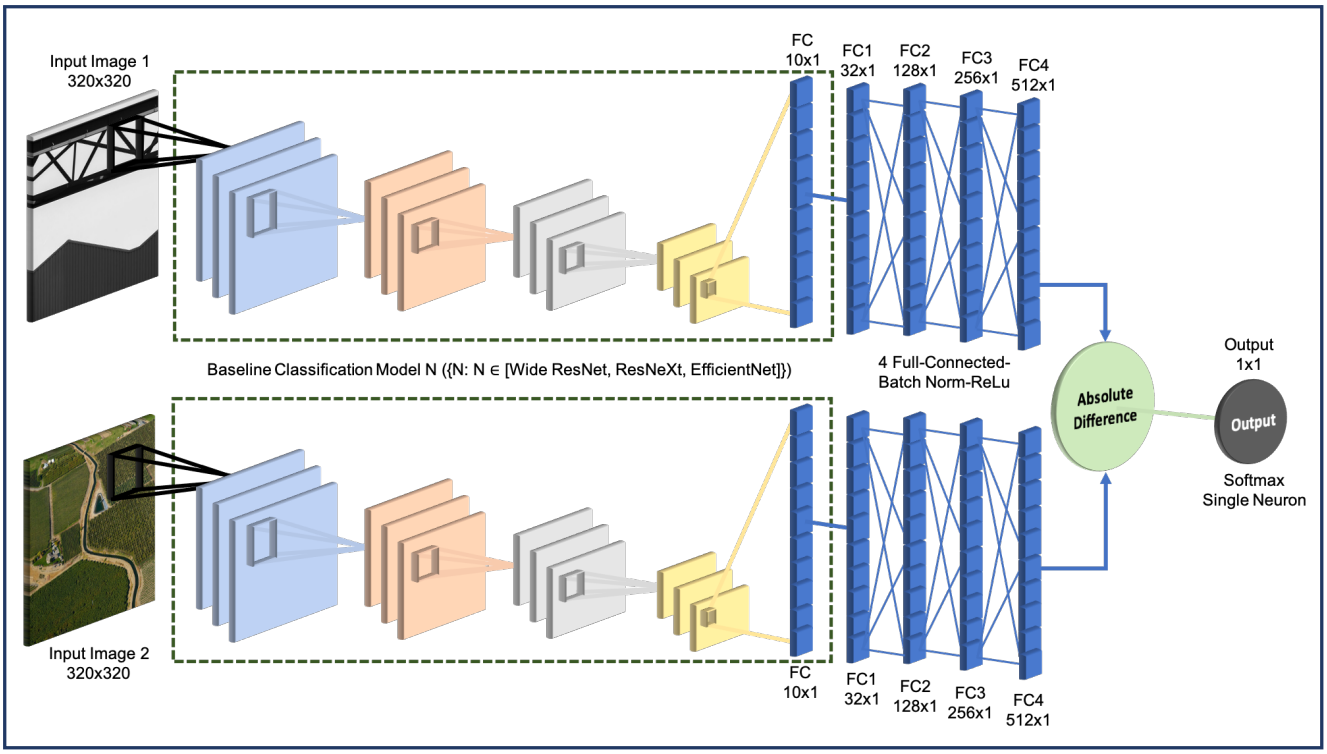


Figure 4: Proposed siamese-based neural network architecture.

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

4.2. Few-Shot Learning

Existing deep neural network architectures generally contain billions of parameters. Thus, training such DNN architectures requires a large training set which is often challenging due to the data collection and annotation of the ground truth. Few-Shot learning, as the name suggests, is a class of techniques that allows a DNN to learn with a smaller number of training samples. In our canonical case, photography is a rapidly evolving industry. Over time, various styles of photography have evolved and it is infeasible to modify our base classification model to adapt to a newer photography style. Instead, we use a network architecture called Siamese network to quickly adapt our base classification model to an unseen class of photography even with a low number of annotated training samples.

4.2.1 Siamese Network

Siamese networks [10] 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 have trained three state-of-the-art classification models for predicting the style of an image as being 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 more 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 siamese. 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.

$$P(I_1, I_2) = Softmax(W_{out} \cdot |\mathcal{N}(I_1) - \mathcal{N}(I_2)| + B_{out})$$
$$Where \;\; W_{out} \in \mathcal{R}^{512\times1} \;\; and \;\; B_{out} \in \mathcal{R}^{1\times1} \tag{4}$$

Furthermore, we use the cross-entropy loss in our Siamese network and a learning rate of 0.05 with SGD [20] as our optimizer to train for 30 epochs with each class having only 25 training, validation, and test samples each.

Figure 4 represents the complete model architecture of the proposed work.

5. Experimental Results

5.1. Classification Accuracy

As explained in Section 4.1, we used 10 photography classes to evaluate the classification accuracy of our dataset on 3 state-of-the-art models (see 2). 70% of the dataset was used for training and 30% as the test dataset. All three models achieve over 96% accuracy on the test dataset. ResNeXt marginally outperforms the other two with an accuracy of 96.35%.

Table 2: Classification models' accuracies.

Classification Model	Accuracy (%)
Wide ResNet	96.23
ResNeXt	96.35
Efficient Net	95.71

5.2. 10-Way Few-Shot Evaluation Metric

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

For the evaluation of 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 3 depicts the $10-$Way few-shot evaluation accuracies for siamese with different base networks. In summary, the Siamese network with ResNext performs better than the other two variants (68.34%). EfficientNet is currently the state-of-the-art network for image classification. Surprisingly, EfficientNet only gives an accuracy of 60.84%. The other two variants give an $8 - 12\%$ higher accuracy.

Table 3: Siamese network' accuracies with different classification models.

Model (with Siamese)	Accuracy (%)
Wide ResNet	64.17
ResNext	68.34
EfficientNet	56.25

5.3. 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) [24]. t-SNE is a non-linear and unsupervised dimensionality reduction approach to visualize high-dimensional data. Intuitively, t-SNE allows one to visualize how the data is arranged in the higher dimensional space.

We use the 512-dim feature embedding output as an input to generate the t-SNE transform. Figure 5 shows the t-SNE plots to visualize the 512-dim visual embedding arranged in a 2-dim space.

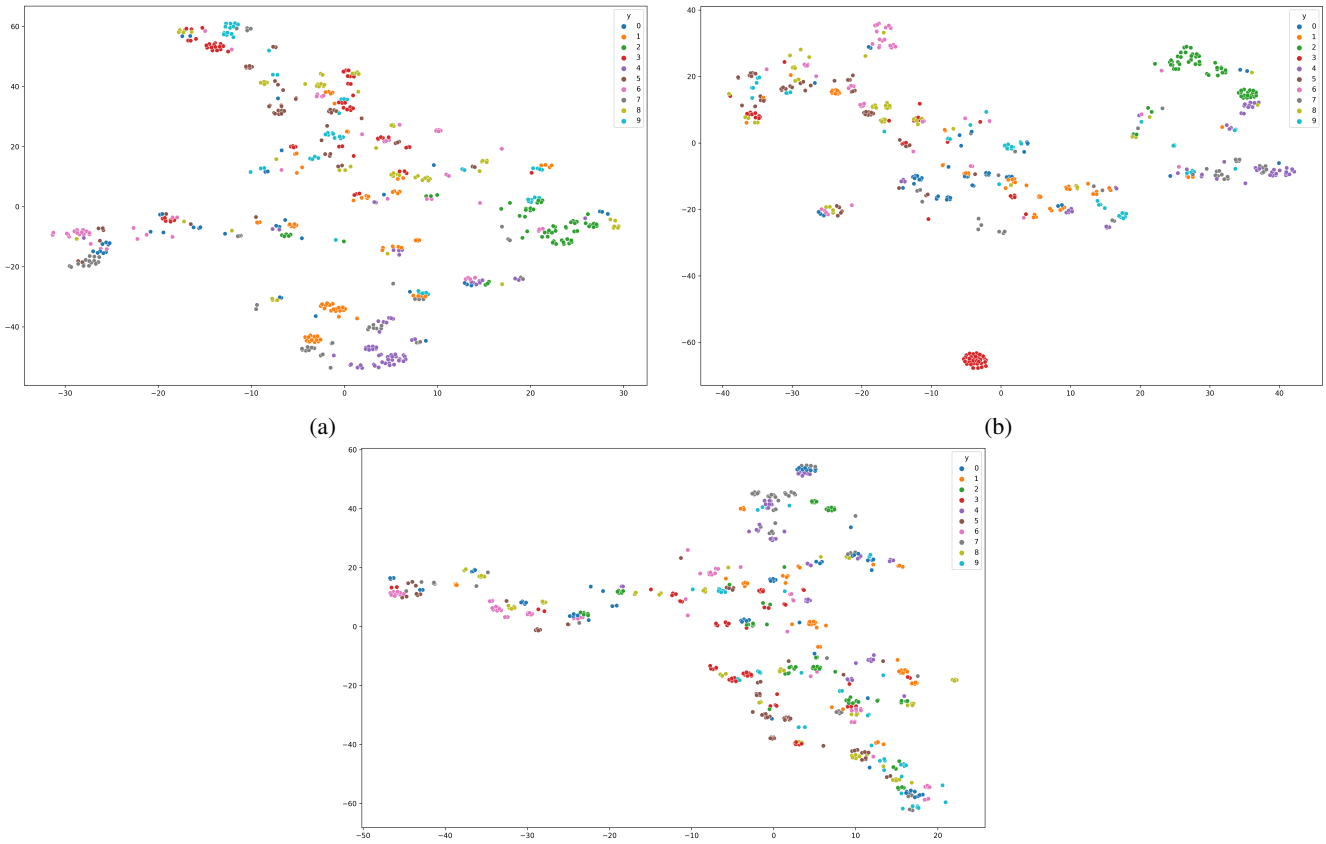
Based on this, we can observe that the Siamese variant of ResNeXt and Wide ResNet can generate better clusters for images of the same photography style. However, we do not observe any distinct clusters for EfficientNet. This is also proven by the performance comparison of Siamese networks where ResNeXt and Wide ResNet perform comparatively better than EfficientNet's Siamese network.

6. Discussion and Future Work

In this section, we discuss some potential applications of this work and highlight future directions for research.

6.1. Photozilla as a Service

Our proposed model can automatically identify 10 different classes and even adapt to new and unseen photography styles with merely 25 training samples. We envision our model as an API service to be used by digital photography platforms. We believe that these platforms will greatly benefit from such a service as it allows the automatic curating of their collection with relevant tags, enables photograph recommendation systems, etc.



(a) (b)

(c)

Figure 5: Clusters Visualization using t-SNE for (a) Wide ResNet, (b) ResNeXt, and (c) EfficientNet. Here, 'y' represents the different photography styles.

6.2. Siamese Visual Embedding for Photography Recommendation

In our proposed Siamese network, we use a $512-$dim visual feature embedding to quantify the similarity between two images. We could instead use this embedding for recommending images based on user's preferences. Such a system could use user-specific features, e.g. user's biographies, previous views, and likes, to generate another feature embedding. This feature embedding can be compared with the image embedding to recommend users with photographs of their preference. Mathematically, let us define $\mathcal{N}_V(I)$ to be the network that takes image $I \in \mathcal{I}$ as an input to generate visual feature embedding. Let $\mathcal{N}_T(U)$ be the network that takes the user's features U as an input to generate user-specific feature embedding. The system could recommend an image I to the user if, $dist(\mathcal{N}_V(I), \mathcal{N}_T(U)) \leq \mathcal{T}$. Here, $\mathcal{T}$ is a pre-determined similarity threshold.

6.3. Royalty-Free Images

Currently, with Photozilla, we have collected a large-scale dataset of images under Creative Commons license and copyright-free images. Moving forward, we intend to utilize this dataset to generate synthetic images using state-of-the-art generative models. Such royalty and copyright-free synthetic images could be beneficial for the academic communities to carry out various researches.

6.4. Next-POI Recommendation

We intend to collect more information and fuse different types of features, such as captions, locations, and hashtags [6, 13], and perform multi-modal analysis and sequence prediction tasks for various applications of the dataset, such as next-POI (Point-Of-Interest) recommendations [14].

6.5. Image Quality and Aesthetic Assessment

Finally, we aspire to collaborate with experts from the design community to propose models and conduct experiments to address the non-trivial issue of aesthetic and quality assessment.

7. Conclusion

This work presents a large-scale dataset termed *'Photozilla'* comprising over $990k$ images belonging to 10 photography styles. We used this dataset as a canonical example to train 3 different classification model architectures to automatically identify the photography style. These models achieve superior 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. The proposed Siamese network achieves an accuracy of over 68% on identifying 10 new photography styles with merely 25 training samples.

References

[1] Website: Wikiart, https://www.wikiart.org/. 2

[2] Wesbite: Flickr, https://www.flickr.com/. 2

[3] Lyujie Chen, Feng Liu, Yan Zhao, Wufan Wang, Xiaming Yuan, and Jihong Zhu. Valid: A comprehensive virtual aerial image dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2009–2016. IEEE, 2020. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4

[6] Jerome Heng, Junhua Liu, and Kwan Hui Lim. Urban crowdsensing using social media: An empirical study on transformer and recurrent neural networks. In *Proceedings of the 2020 IEEE International Conference on Big Data*, 2020. 7

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3

[8] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013. 2

[9] Fahad Shahbaz Khan, Shida Beigpour, Joost Van de Weijer, and Michael Felsberg. Painting-91: a large scale database for computational painting categorization. *Machine vision and applications*, 25(6):1385–1397, 2014. 2

[10] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015. 3, 5

[11] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008. 3

[12] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[13] Junhua Liu, Trisha Singhal, Lucienne T.M. Blessing, Kristin L. Wood, and Kwan Hui Lim. Epic30m: An epidemics corpus of over 30 million relevant tweets. In *Proceedings of the 2020 IEEE International Conference on Big Data*, 2020. 7

[14] Junhua Liu, Kristin L Wood, and Kwan Hui Lim. Strategic and crowd-aware itinerary recommendation. In *Proceedings of the 2020 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'20)*, 2020. 7

[15] Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 464–471. IEEE, 2000. 3

[16] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys (CSUR)*, 52(5):1–36, 2019. 3

[17] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012. 2

[18] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017. 3

[19] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012. 3

[20] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 6

[21] Doyen Sahoo, Wang Hao, Shu Ke, Wu Xiongwei, Hung Le, Palakorn Achananuparp, Ee-Peng Lim, and Steven CH Hoi. Foodai: Food image recognition via deep learning for smart food logging. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2260–2268, 2019. 3

[22] Kekai Sheng, Weiming Dong, Haibin Huang, Chongyang Ma, and Bao-Gang Hu. Gourmet photography dataset for aesthetic assessment of food images. In *SIGGRAPH Asia 2018 Technical Briefs*, pages 1–4. 2018. 3

[23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3, 4

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[25] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE international conference on computer vision*, pages 1202–1211, 2017. 2

[26] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 3

[27] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3, 4

[28] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 3, 4

[29] Chenyang Zhang, Christine Kaeser-Chen, Grace Vesom, Jennie Choi, Maria Kessler, and Serge Belongie. The imet collection 2019 challenge dataset. *arXiv preprint arXiv:1906.00901*, 2019. 2