New Perspectives for the Deep Learning Based Photography Aesthetics Assessment

Vernon Asuncion and Yan Zhang

Western Sydney University, Australia
{v.asuncion,yan.zhang}@westernsydney.edu.au

Abstract. Image aesthetics assessment (IAA) has been an important topic in computer vision research. In recent years, many deep learning based approaches have been developed for machine automatic image (photograph) assessment. In this paper, we study the IAA for photography assessment, which captures the most significant aesthetic features in photography. In particular, we re-examine the multi-column deep convolutional neural network architecture, in which photographs are assessed based on their global and local views to make the assessment more precise. Our experiments are conducted with a completely new dataset we have built - Curated Photography Dataset (CPD) which contains over 500,000 photographs crossing eight different categories, and all of these photos have been curated by professional photographers and curators. We show that our approach outperforms the state of the art approaches in the area, and sheds a new light for developing practical AI photography curators in real world domains.

Keywords: Machine Learning · Classification · Image Classification · Image Aesthetics Assessment · Convolutional Neural Networks.

1 Introduction

Image Aesthetics Assessment (IAA) has been an important topic in computer vision research. In recent years, AI researchers started to explore the use of deep learning based approaches for IAA model development, and have made important progress, e.g., [1,2,26].

1.1 Related work

While Image Aesthetics Assessment (IAA) has been a focused research in computer vision for a long time, e.g., [10], deep learning based IAA has only become a predominant methodology from mid 2010s, e.g., [4,16,18,22,24]. In the following, we provide a brief overall of current deep learning based IAA models. We classify these models into three major different approaches: multi-column IAA models, image aspect ratio oriented IAA models, and image category based IAA models, because these three different approaches are mostly related to our work presented in this paper.

Multi-column CNN models. Lu *et al*'s influential paper [15] probably represents one of the earliest works of using deep learning for image aesthetics assessment. In their work, they firstly proposed a single-column deep convolutional neural network (DCNN) architecture for image aesthetics predictions, where the model was trained by input images with various normalisations, such as warping, padding and cropping, individually. Based on this single column model architecture, they then developed a double-column DCNN architecture, in which two different image features, such as global and local views, can be processed simultaneously and independently. In this way, their model was able to further improve aesthetic categorisation using style attributes and semantic attributes. AVA [19] was used as the major dataset for their development.

Besides Lu *et al*'s work [15], there were some other works also using multiple column DCNN architecture for image aesthetics predictions. Gou and Li studied various combinations to find an optimal multiple column DCNN architecture for aesthetics assessment [7]. In their approach, they used the classical AlexNet [13] as the baseline model, and through experiments, they found an optimal number of DCNNs which has the lowest error rate. However, one major issue of their work is that they only used PhotoQuality and CUHK two datasets in their model training, and these two datasets were rather restricted, compared to the much more commonly used dataset AVA. Another approach is from [27] through their "StereoQA-Net" that uses as inputs the left and right view distorted images patches as the input to the two-columns.

Similarly to [5,7], Fu, Yan and Fan also used pre-trained DCNNs as the baseline models to build their neural network architecture [6]. In their three column DCNNs, images' global views, local views and scene recognitions are learned via three identical pre-trained models, and the corresponding features are extracted from certain layers of these models, respectively, and then concatenated and input to SVM for making a final decision. In their work, AlexNet, VGG-16 and ResNet-50 were used for the baseline models, and CUHKPQ and AVA2[1] datasets were employed for conducting experiments.

Image aspect ratio oriented CNN models. Another line of research of IAA in recent years is to take the image aspect ratio into account. The main argument is that the aesthetics of an original image is very likely different from the aesthetics of its transformed image, i.e., cropping and warping, as aesthetics information in the original image may be lost from such transformation [3]. An early effort in this aspect is due to He *et al*'s work [8] , where by applying the adaptive spatial pooling strategy, it enables the ConvNet to operate on an image in its original form during both training and testing. In their work presented in [17], Mai, Jin and Liu designed a network architecture where images are input into n-column models, each of them consists of lower level convolution and pooling layers, followed by a variable dimensional adaptive spatial pooling layer, and then fully connected layers. The final assessment result is the average results from these

[1] AVA2 is a subset of AVA containing about 10% of the "highest quality" images and 10% of the "lowest quality" images of AVA, respectively [7].

number of n models. Their experiments were conducted on three VGG-net based architecture on AVA dataset. Nevertheless, we have re-done their experiments using InceptionResNet and NASNetLarge CNN architectures, and on both AVA and a new dataset CPD (see next sections), and the results seemed not as effective as we expected. We will explain why this is the case in the next section.

Attention based CNN models. All approaches described above share one common feature - images in different categories are assessed together without differentiations. However, in photography, it is well known that images in different categories, i.e., landscape, portrait, wildlife, etc., are taken based on different rules. Therefore, image aesthetics should be also evaluated in a different way for each image category.

From this observation, Le *et al* proposed an alternative IAA approach based on image classification and region segmentation [14]. Instead of assess different category images with different aesthetics criteria, the authors proposed a notion called Region of Interests (ROIs) and assumed that an image aesthetics is closely related the the aesthetics of ROIs in an image. Following this intuition, they considered two specific ROIs image classifications: the large field and close-up classifications, and developed three deep learning models for ROI based image aesthetics assessment. However, one limitation of their work was that they only used a very restricted dataset - 406 images from CUHKPQ dataset and 750 images from the flickr.com site. So it is hard to know whether their approach will perform well for any large dataset such as AVA.

There are other attention based approaches for IAA, e.g., [21,22,25]. In [22], the authors proposed an attention-based multi-patch aggregation for image aesthetics assessment. In their approach, they developed an attention-based mechanism that adaptively adjusts the weight of each patch during the training process to improve learning efficiency. They proposed a set of objectives with three attention mechanisms and evaluated their effectiveness for aesthetics assessment.

Content and scene based CNN models. Content and scene based models have been one of the major approaches for IAA in last a few years. One common observation for image aesthetics assessment is that an image's aesthetics may be closely related to the image's content (scene) [15]. For instance, an image of mountain sunrise scene is likely more visually attracted than an image with a sleeping dog.

Kao, He and Huang proposed a multi-task single CNN model, where the model can jointly learn both image aesthetics and image semantics, and their correlations [11]. In their approach, aesthetics quality assessment has not been taken as an isolation problem, while semantic information influence on aesthetics task was studied, and four multi-task CNN architectures have been explored to learn the aesthetic representation jointly with the supervision of aesthetic and semantic labels.

Another work alone this line was due to Kong *et al.*'s [12], in which the authors developed a CNN model to learn jointly image aesthetic attributes such as symmetry, shallow depth of field, etc.[2], and image content information. In particular, Their architecture contains an attribute-adaptive model and a content-adaptive model, and the final image aesthetics prediction was made from the concatenation of attribute features and content-specific features. In this way, the authors claimed that their underlying model achieved the state-of-the-art performance on existing aesthetics classification benchmark. However, one main limitation of their approach is the dataset they used for the model training. They built a dataset called Aesthetics and Attributes Database (AADB) with various annotations they needed. But AADB only contains 10,000 images, which greatly restricted the generalization of their approach for other large scale datasets such as AVA.

In this paper, we re-examine the multi-column model approach as originally proposed in [15] that considered an image's two features corresponding to its local and global views. Moreover, we further trained and evaluated the resulting model from the CPD dataset. This paper is organised as follows: Section 2 we discuss the quality and limits of current datasets used in automatic IAA and further discuss why on real professionally curated datasets considering the accuracy measure is not sufficient enough. Section 3 defines our approach that includes a more detailed discussions on the image transformation and single/multi-column approaches. Section 4 then shows our experiments and comparison between the single and multi-column approach with a more detailed look at the CPD dataset used to train and evaluate our models. Finally, Section 5 discusses the concluding remarks.

2 The Quality of Datasets and Model Accuracy

Datasets for training an IAA model play a critical role, the volume and quality of a photography dataset directly influence the final model performance. In previous years, researchers used several different photography datasets for IAA model developments, such as AVA, TID2013, CUHK-PQ, AADB [6,15,19]. Among these datasets, AVA is the primary dataset that most researchers have used for their IAA model training, while other existing datasets were less popular and rarely used in recent years due to their small number of images. For instance, CUHK-PQ contains 17,613 images and AADB only contains about 10,000 images, which are not sufficient for undertaking an effective model training.

However, from a professional photography viewpoint, there are three major issues in using AVA as a basis for develop photography aesthetics assessment models.

- Firstly, most images in AVA are in poor photography quality, even if many of them have been rated with high scores. AVA contains 255,530 photos, and each photo is associated with a rating score between 0 to 10, from which

[2] In particular, with consulting professional photographers, the authors identified eleven aesthetic attributes.

each image is then classified as either “high quality” or “low quality”. In particular, if an image has a score of 6 or above, we may view this image as “high quality”, otherwise, as “low quality”. Under this scoring method, AVA contains 81,989 “high quality” images, which is about 33% of the total images, and the rest images are viewed as in “low quality”.

Nevertheless, many such “high quality” images in AVA are actually aesthetically poor, e.g., see the following collage of AVA “high quality” images in Figure 1. Consequently, many models trained from such dataset cannot really make rational aesthetics predictions for the real world photography communities.

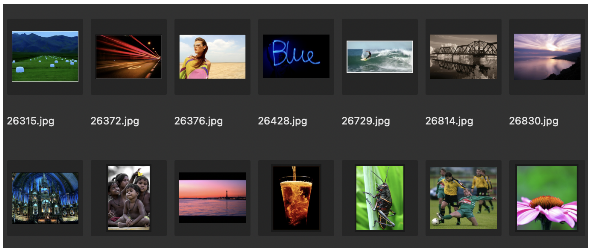


Fig. 1: Some of AVA images with high rating scores.

- Secondly, although all images in AVA are explicitly labelled with a certain photography categories such as landscape, portrait, animal, etc., for all current deep learning based IAA approaches, model trainings were based on the entire AVA or AVA2, where all images with different categories were mixed together during the training process.

  The advantage for such training method is of two fold: (1) since all categories of images are trained together, the outcome model is then able to make aesthetics predictions for all category images without any restriction; (2) putting all images together can also form a comprehensive dataset where sufficient amount of images are necessary for the underlying training.

  However, a mixed dataset has one critical drawback: since for different photography categories, the amount of images may vary significantly, for instance, the landscape photos usually much more than the photos with abstract category, as a result, the final model’s performance in aesthetics predictions may be not balanced - it may perform well for some specific photography categories (such as landscape category) but not for other photography categories (such as for abstract category).
- Finally, for some images, AVA provides not only aesthetics rating scores, but also rating scores on 14 different style attributes, such as Long Exposure, Complementary Colors, Macro, Soft Focus, etc.. These style attributes were used by many IAA approaches as important aggregations in the model training, in order to improve the final model performance for image assessment.

  While these essential technique attributes play an important role to distinguish generally good photos from poor photos which, technically, are not properly taken, they are hardly the critical factors to determine excellent images standing out from ordinary images.

In current research, the model accuracy against the underlying test dataset is the primary measure for evaluating an IAA model. That is, the (overall) model accuracy is defined as

$$Acc = \frac{TP + TN}{m}, \tag{1}$$

where TP and TN are the numbers of samples that the model correctly predicts positive and negative cases, respectively, and m is the number of total data samples of the dataset.

However, using the (overall) model accuracy to evaluate an IAA model can be problematic in practice. This is because that in general, photography datasets are highly imbalanced in terms of the amounts of "good" and "poor" photos. For example, for the well-known online photography art gallery 1x.com, all photos published on this site have been selected by professional curators, and only less than 5% of the total submitted photos may be finally accepted for publishing on 1X.COM's online gallery. As a consequence of this imbalanced dataset, quite often, the model accuracy may not really reflect the actual model's aesthetics prediction in practice.

Consider in a photography competition event, we have an imbalanced photography dataset, where the ratio of "good" and "poor" photos is 2:8. With a total 20,000 photo submissions, for example, 4,000 images are considered to be good and 16,000 images to be poor. Then the model may perform poorly in predicting true positive cases, but perform well in predicting true negative cases, say TP = 2,000, and TN = 13,000. In this case, we have the model accuracy $Acc = \frac{2000+13000}{20000} = 0.75$. In this case, the model has a 75% overall accuracy, which may be considered to be high enough for an IAA application. But in reality, the model can only make about 50% correct predictions for good photos. This is far less satisfied if we use the model to screen all submitted photos in a real photography competition.

As mentioned earlier, most training datasets are usually highly imbalanced in the IAA field, but achieving balanced accuracies is critical. One way to deal with this issue is to use F1-score as a metric to evaluate a model, instead of using the accuracy described in (1). To define F1-score, let us first introduce *Precision* and *Recall*, respectively, as:

$$Prec = \frac{TP}{TP + FP} \text{ and } Rec = \frac{TP}{TP + FN}. \tag{2}$$

Then the F1-score is defined as follows:

$$F1\text{-}score = 2 \times \frac{Prec \times Rec}{Prec + Rec}. \tag{3}$$

Now let us revisit the above example. In our case, we will have $Prec = \frac{2000}{2000+2000} = 0.5$ and $Rec = \frac{2000}{2000+3000} = 0.4$. Finally, we have $F1\text{-}score = 2 \times \frac{0.5 \times 0.4}{0.5+0.4} = 0.44$. That is, although the model achieves 75% overall accuracy, its F1-score is only 0.44, which reveals the model's poor performance in actual predictions for the underlying dataset.

Nevertheless, one drawback of using F1-score to evaluate a model is that F1-score does not provide concrete accuracy information for the underlying model performance, which, especially for IAA applications, is very important.

In this paper, we argue that when we evaluate an IAA model, instead of solely using the model accuracy or F1-score individually, we should use more comprehensive metrics for model evaluations. In particular, in addition to consider both the model accuracy, we also take into account of *true positive accuracy* and *true negative accuracy.*

Then true positive accuracy and true negative accuracy, denoted as *TP Acc* and *TN Acc*, respectively, are defined as:

$$TP\ Acc = \frac{TP}{TP + FP} \quad \text{and} \quad TN\ Acc = \frac{TN}{TN + FN}. \tag{4}$$

Now let us consider our previous photography competition example again. We have $TP\ Acc = \frac{2000}{2000+2000} = 0.50$ and $TN\ Acc = \frac{13000}{16000} = 0.8125$. So, for this example, we use the proposed combined metrics to represent the model's performance as follows:

Acc	*TP Acc*	*TN Acc*
75%	50%	81.25%

3 The Approach

We view a *dataset* D as a set of pairs of 2D feature maps and labels: $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ such that $\mathbf{x}_i \in \mathbb{R}^{h\times w\times c}$, where h, w and c denotes the *height*, *width* and *channels* of the feature maps, respectively, and $y_i \in \{\text{yes, no}\}$. The main problem of interest in this work is that for a given dataset D, we wish to fit an approximating function: $f : \mathbb{R}^{h\times w\times c} \longrightarrow \{\text{yes, no}\}$ that maximises the (accuracy) utility function:

$$g(D, f) = \frac{\sum\limits_{\substack{f(\mathbf{x}_i) = y_i,\\ (\mathbf{x}_i, y_i) \in D}} 1}{m},$$

i.e., the ratio of correctly predicted instance labels (i.e., where $f(\mathbf{x}_i) = y_i$ holds) with the size of dataset m. Ideally, we further assume a partitioning of the dataset D into disjoint subsets D_{train} and D_{test} such that the function f can be parameterised only by D_{train} (i.e., $f_{D_{train}}$) but where the utility function $g(D_{test}, f_{D_{train}})$ is also being maximised.

In this work, the dataset D contains the pairs $(\mathbf{x}_i, y_i)$ such that "$\mathbf{x}_i$" is a feature map corresponding to an image file (e.g., a landscape image) and y_i is its target label (e.g., "yes" for aesthetic or "no" for non-aesthetic photographs). The approximating function f (as mentioned above) takes the form of a *convolutional neural network* (CNN) [20].

3.1 Warp and Crop Image Transformations

Using the notions from [15], assuming we have an initial set of (image) dataset $I = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ such that $\mathbf{x}_i \in \mathbb{R}^{h\times w\times c}$, we further assume two transformations $\mathcal{C}_{h' \times w'} : \mathbb{R}^{h \times w \times c} \longrightarrow \mathbb{R}^{h' \times w' \times c}$ (or just $\mathcal{C}$) and $\mathcal{W}_{h' \times w'} : \mathbb{R}^{h \times w \times c} \longrightarrow \mathbb{R}^{h' \times w' \times c}$ (or just $\mathcal{W}$) such that $\mathcal{C}_{h' \times w'}(\mathbf{x}_i)$ and $\mathcal{W}_{h' \times w'}(\mathbf{x}_i)$ denotes the *random crop* (or just *crop*) and *warp* transformations of the original $h \times w \times c$ image into an $h' \times w' \times c$ image, respectively. For convenience, for the given dataset I above, we denote by $\mathcal{C}_{h' \times w'}(I)$ and $\mathcal{W}_{h' \times w'}(I)$ as the transformed datasets $\big\{(\mathcal{C}_{h' \times w'}(\mathbf{x}_1), y_1),\ \ldots,\ (\mathcal{C}_{h' \times w'}(\mathbf{x}_m), y_m)\big\}$ and $\big\{(\mathcal{W}_{h' \times w'}(\mathbf{x}_1), y_1),\ \ldots,\ (\mathcal{W}_{h' \times w'}(\mathbf{x}_m), y_m)\big\}$, respectively.

3.2 Single Column (1-col) Approach

In this *single column* (1-col) strategy, we trained a single CNN exclusively for each of the warp and crop transformed image dataset $I = \big\{(\mathbf{x}_1, y_1),\ \ldots,\ (\mathbf{x}_m, y_m)\big\}$. More specifically, for a CNN $\mathcal{F}$, we trained two CNNs: $\mathcal{F}_{C_{train}}$ and $\mathcal{F}_{W_{train}}$, where C_{train} and W_{train} are the (transformed) datasets $\mathcal{C}(I_{train})$ and $\mathcal{W}(I_{train})$, respectively, and where they were evaluated against $C_{test} = \mathcal{C}(I_{test})$ and $W_{test} = \mathcal{W}(I_{test})$, respectively.

3.3 Double Column (2-col) Approach

Due to a space reason, we only report on the *double column* (2-col) approach for the NASNetLarge *convolutional neural network* (CNN) architecture [28]. We have also implemented this 2-col approach for other well known CNN architectures such as ResNet-50 [9] and InceptionResNet-V2 [23] but these were outperformed by the NASNetLarge architecture. NASNetLarge achieves, among the published works, state-of-the-art accuracy of 82.7 top-1 and 96.2 top-5 on ImageNet [28].

Assuming a NASNetLarge CNN $\mathcal{F}$ is of the form $\mathcal{F} = \sigma \circ \mathcal{G}$ such that σ is its output (*softmax*) activation layer, we define another CNN $\mathcal{F'}_{X_{train}}$:

$$\mathcal{F'}_{X_{train}} = \sigma' \circ (\mathcal{G} \otimes \mathcal{G}),$$

where $\otimes$ is the concatenation operator on the output layers of the two CNNs $\mathcal{G}$ and σ' is the output (*softmax*) activation function of $\mathcal{F'}_{X_{train}}$. The training set X_{train} now corresponds to the multi-input dataset:

$$\big\{\big(\,[\mathcal{C}(\mathbf{x}_1), \mathcal{W}(\mathbf{x}_1)], y_1\big), \ldots, \big(\,[\mathcal{C}(\mathbf{x}_m), \mathcal{W}(\mathbf{x}_m)], y_m\big)\big\}$$

such that $\big\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\big\}$ is still viewed as the original dataset. In Figure 2, the two transformed images are depicted by the inputs "$\mathcal{W}(\mathbf{x})$" and "$\mathcal{C}(\mathbf{x})$" with $\mathbf{x}$ the feature map corresponding to the input image.

4 Experiments and Comparisons

4.1 CPD Dataset

The *curated photography dataset* (CPD) is a dataset of images containing half a million (566,384) high quality professionally curated photographs from the

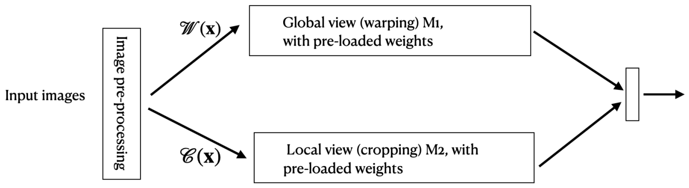

Input images
Image pre-processing
𝒲(x)
Global view (warping) M1, with pre-loaded weights
𝒞(x)
Local view (cropping) M2, with pre-loaded weights

Fig. 2: The 2-col approach.

Abstract	Architecture	Creative	Landscape	Nude	Overall	Portrait	Sill	Street	Wildlife
37,520	74,207	59,179	39,870	23,887	38,871	108,507	55,993	67,273	61,077
(total) 566,384									

Table 1: CPD dataset composition.

following areas of artistic photography as shown on Table 1. One source of this high quality dataset include the curated online photography gallery site 1x.com (https://1x.com/), which publishes only 5% of the submitted images as deemed to be "aesthetic" in the sense of professional photography.

In this work, we focused particularly on the Landscape photography set. For this dataset, in addition to the 39,870 good quality photographs, we also incorporated 117,767 low aesthetic photographs for our training data[3].

In our experiments, for a given feature map $\mathbf{x}$ corresponding to an image from the Landscape dataset, the crop and warp transformations: $\mathcal{C}_{h\times w}(\mathbf{x})$ and $\mathcal{W}_{h\times w}(\mathbf{x})$, have $h = w = 224$, i.e., the input images to the CNNs are of size $224 \times 224 \times 3$ feature maps after the crop and warp transformations.

4.2 The AiCUS System and Experimental Results

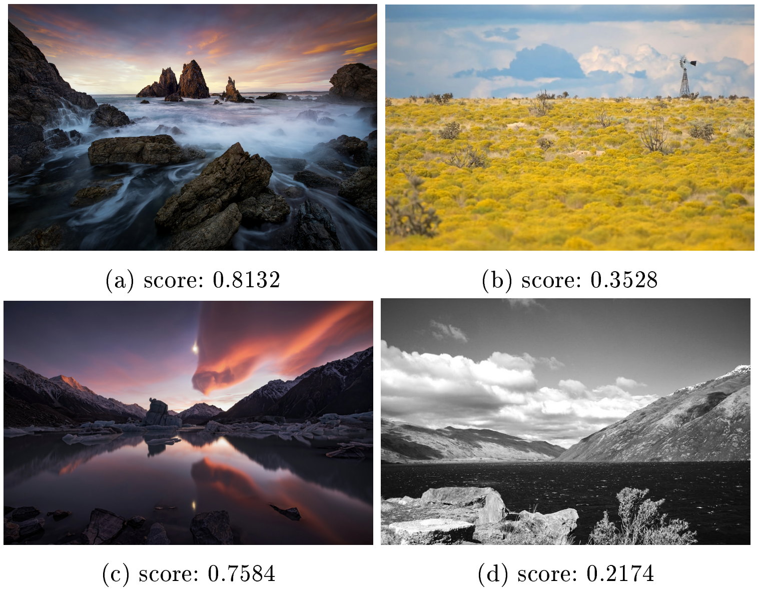
We can see from Table 2 that although the accuracies are quite high for the 1-col or 2-col approach, the balance between the TP and TN accuracies are quite low. So for this reason, we chose the results that not only considers the accuracy but also the balance between the TP and TN accuracies. Moreover, we further observe that the 2-col approach outperformed the 1-col exclusive crop and warp approaches. Indeed, for the 1-col approaches, we have that the average of the TP and TN (i.e., $(TP + TN)/2$) for the balanced results are $(65 + 73)/2 = 69\%$ and $(68 + 75)/2 = 71.5\%$ for crop and warp, respectively, while it is $(73+72)/2 = 72.5\%$ for the 2-col approach. We have implemented the 2-col approach that is trained on the CPD dataset in our system called AiCUS: https://airg.cdms.westernsydney.edu.au/aicus/.

In Figure 3, 3a and 3c are good scores (score ≥ 0.5, "aesthetic") while 3b and 3d are bad scores (score < 0.5, "non-aesthetic").

[3] Source of those "low" photographs includes AVA and data from other professional photography sites like the rejected images from 1x.com.

Base model	Approach	Dataset	Acc Type	Acc (%)	TP Acc (%)	TN Acc (%)
NASNetLarge	1-col	crop	best	76	33	96
			balanced	70	65	73
		warp	best	80	46	96
			balanced	73	68	75
	2-col	crop + warp	best	81	39	97
			balanced	72	73	72

Table 2: Experimental results.



(a) score: 0.8132

(b) score: 0.3528

(c) score: 0.7584

(d) score: 0.2174

Fig. 3: Returned scores by AiCUS.

5 Concluding Remarks

In this work, we have trained a CNN model on the CPD dataset that contains high quality and high resolution curated photographs. We further showed that the 2-col approach on both the crop and warp transformations of an image map is superior in performance than with the 1-col (exclusive crop and warp) approach. Moreover, although real high quality photography datasets are usually not balanced in the sense that there will be a lot more "no" (low aesthetic) than "yes" (high aesthetic) photographs (i.e., a highly-skewed dataset), we have carefully considered the balance between TP and TN accuracies in our evaluation of the resulting model's performance.

In spite of there being already some works on using CNN for evaluating a photograph's aesthetic quality (e.g., [15]), these works were mainly based on the AVA/AVA2 datasets, which we observed are not truly representative of the precise distinction between aesthetic and non-aesthetic photographs. For this reason, we argue that our approach is the first one on a high quality photography dataset. Finally, in the same way the AVA and AVA2 datasets have played a significant role in the use of CNN for machine (automatic) IAA [19], we believe that our contribution of the high quality CPD dataset would also play a significant role in the future of research within this area as well.

References

1. Bianco, S., Celona, L., Napoletano, P., Schettini, R.: Predicting image aesthetics with deep learning. In: Proceedings of 2016 International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS). vol. 10016, pp. 117–125 (10 2016) 1
2. Bosse, S., Maniry, D., Müller, K., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans. Image Process. **27**(1), 206–219 (2018) 1
3. Chen, Q., Zhang, W., Zhou, N., Lei, P., Xu, Y., Zheng, Y., Fan, J.: Adaptive fractional dilated convolution network for image aesthetics assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 2
4. Deng, Y., Loy, C.C., Tang, X.: Image aesthetic assessment: An experimental survey. IEEE Signal Process Magazine **34**, 80–106 (2017) 1
5. Doshi, N., Shikkenawis, G., Mitra, S.K.: image aesthetics assessment using multi channel convolutional neural networks. In: Proceedings of 2019 International Conference on COmputer Vision and Image Processing (CVIP). pp. 15–24 (2019) 2
6. Fu, X., Yan, J., Fan, C.: Image aesthetics assessment using composite features from off-the-shelf deep models. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 3528–3532 (2018) 2, 4
7. Guo, L., Li, F., Liew, A.W.C.: Image aesthetic evaluation using parallel deep convolution neural network. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–5 (2016) 2
8. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Proceedings of 2014 European Conference on Computer Vision (ECCV). pp. 346–361 (2014) 2
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016) 8
10. Jiang, W., Alexander C., L., Cathleen Daniels, C.: Automatic aesthetic value assessment in photographic images. In: Proceedings of 2010 IEEE International Conference on Multimedia and Expo. pp. 920–925 (2010) 1
11. Kao, Y., He, R., Huang, K.: Deep aesthetic quality assessment with semantic information. IEEE Transactions on Image Processing **26**(3), 1482–1495 (2017) 3
12. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: ECCV (2016) 3
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: **25** (2012) 2

14. Le, Q.T., Ladret, P., Nguyen, H.T., Caplier, A.: Image aesthetic assessment based on image classification and region segmentation. Journal of Imaging **7**(3), 1–33 (2021) 3
15. Lu, X., Lin, Z., Lin, H.J., Yang, J., Wang, J.Z.: Rating image aesthetics using deep learning. IEEE Transactions on Multimedia **17**(11), 2021–2034 (2015) 2, 3, 4, 7, 10
16. Lu, X., Lin, Z., Shen, X., Mech, R., Wang, J.Z.: Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). pp. 990–998 (2015) 1
17. Mai, L., Jin, H., Liu, F.: Composition-preserving deep photo aesthetics assessment. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 497–506 (2016) 2
18. Malu, G., Bapi, R.S., Indurkhya, B.: Learning photography aesthetics with deep CNNs. In: Proceedings of MAICS (2017) 1
19. Murray, N., Marchesotti, L., Perronnin, F.: AVA: A large-scale database for aesthetic visual analysis. In: Proceedings of Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2408–2415 (2012) 2, 4, 11
20. O'Shea, K., Nash, R.: An introduction to convolutional neural networks. CoRR **abs/1511.08458** (2015) 7
21. Qu, Q., Chen, X., Chung, Y.Y., Cai, W.: Lfacon: Introducing anglewise attention to no-reference quality assessment in light field space. IEEE Trans. Vis. Comput. Graph. **29**(5), 2239–2248 (2023) 3
22. Sheng, K., Dong, W., Ma, C., Mei, X., Huang, F., Hu, B.G.: Attention-based multi-patch aggregation for image aesthetic assessment. In: Proceedings of the ACM International Conference on Multimedia. pp. 879–886 (2018) 1, 3
23. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Singh, S., Markovitch, S. (eds.) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. pp. 4278–4284. AAAI Press (2017) 8
24. Talebi, H., Milanfar, P.: NIMA: Neural image assessment. IEEE Transactions on Image Processing **27**(8), 3998–4011 (2018) 1
25. Wang, W., Shen, J.: Deep cropping via attention box prediction and aesthetics assessment. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2205–2213 (2017) 3
26. Zhang, J., Miao, Y., Yu, J.: A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges. IEEE Access **9**, 77164–77187 (2021) 1
27. Zhou, W., Chen, Z., Li, W.: Dual-stream interactive networks for no-reference stereoscopic image quality assessment. IEEE Trans. Image Process. **28**(8), 3946–3958 (2019) 2
28. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 8697–8710. Computer Vision Foundation / IEEE Computer Society (2018) 8