Residual Attention Network vs Real Attention on Aesthetic Assessment of the Great Barrier Reef

Ranju Mandal*, Susanne Becken‡ , Rod M. Connolly†, and Bela Stantic*
*School of Information and Communication Technology, Griffith University, QLD 4222, Australia
†Australian Rivers Institute - Coast & Estuaries and
School of Environment and Science, Griffith University, QLD 4222, Australia
‡Griffith Institute for Tourism, QLD 4222, Australia
{r.mandal, s.becken,r.connolly,b.stantic}@griffith.edu.au

***Abstract*—Photo aesthetics assessment is a challenging problem. Deep Convolutional Neural Network (CNN)-based algorithms have achieved promising results for aesthetics assessment in recent times. Lately, few efficient and effective attention-based CNN architectures are proposed that improve learning efficiency by adaptively adjusts the weight of each patch during the training process. In this paper, we investigate how real human attention affects instead of CNN-based synthetic attention network architecture in image aesthetic assessment. A dataset consists of a large number of images along with eye-tracking information has been developed using an eye-tracking device[1] power by sensor technology for our research, and it will be the first study of its kind in image aesthetic assessment. We adopted a Residual Attention Network and standard ResNet architectures which achieve state-of-the-art performance image recognition tasks on benchmark datasets. We report and demonstrate our findings on photo aesthetics assessment with two sets of datasets consist of original images and images with masked attention patches.**

***Index Terms*—Photo Aesthetic Assesment, Image Aesthetic Evaluation, Great Barrier Reef, Aesthetic Scoring, Deep Learning**

I. Introduction

Image quality assessment and predict photo aesthetic values have been a challenging problem in image processing and computer vision, as aesthetic assessment is subjective (i.e. influenced by individual's feelings, tastes, or opinions) in nature. A significant number of existing photo aesthetics assessment methods are available ([1], [2], [3], [4], [5], [6], [7]) in the literature using extraction of visual features and then employ various machine learning algorithms to predict photo aesthetic values. Aesthetic assessment techniques aim to quantify semantic level characteristics associated with emotions and beauty in images, whereas technical quality assessment deals with measuring low-level degradations such as noise, blur, compression artifacts, etc.

Based on the available image assessment techniques in the literature, full-reference and no-reference approaches are the two main categories of image quality assessment. While the availability of a reference image is assumed in the former (metrics such as PSNR, SSIM [8], etc.), typically blind (no-reference) approaches rely on a statistical model of distortions to predict image quality. A quality score is to predict that relates well with human perception is the main goal in both cases.

Broadly, the task involved to distinguish computationally the aesthetic attributes of an image [9] for a related assumption. The literature proposes several methods to solve such challenging classification and scoring problems. The earlier approaches can be categorised into two groups, based on visual feature types (hand-crafted features and deep features based on Convolutional Neural Network), and evaluation criteria, dataset characteristics and evaluation metrics (examples include: Precision-recall, Euclidean distance, ROC curve, and mean Average Precision). More specifically, the term "hand-crafted" features refer to properties derived employing various algorithms using the information present in an image. As an example, edges and corners are two simple features that can be extracted from images. A basic edge detector algorithm works by finding areas where the image intensity "suddenly" changes. For example, the shell of a turtle can be identified as an edge. Likewise, the so-called Histogram of Gradients (HoG) [10] is another type of handcrafted feature that can be applied in many different ways.

Earlier proposed techniques designed hand-crafted aesthetics features according to aesthetics perception of people and photography rules [11], [12], [1] and obtained encouraging results while handcrafted feature design for aesthetic assessment is a very challenging task. More robust feature extraction techniques were proposed later on to leverage more generic image features (e.g. Fisher Vector [13], [14], [15] and the bag of visual words [6]) for photo aesthetics evaluation. Generic feature-based representation of images is not ideal for image aesthetic assessment as those features are designed to represent natural images in general, and not specifically for aesthetics assessment.

In contrast, Convolutional Neural Network (CNN)-based features are learning from the training samples, and they do this by using dimensionality reduction and convolutional filters. Recent approaches to image aesthetic assessment mostly apply more complex and robust deep Convolutional Neural Networks (CNN) architectures ([16], [17], [18], [19]). Availability of large-scaled labeled and scored images from online repositories have enabled CNN-based methods to perform bet-

[1]https://www.tobii.com/group/about/this-is-eye-tracking/

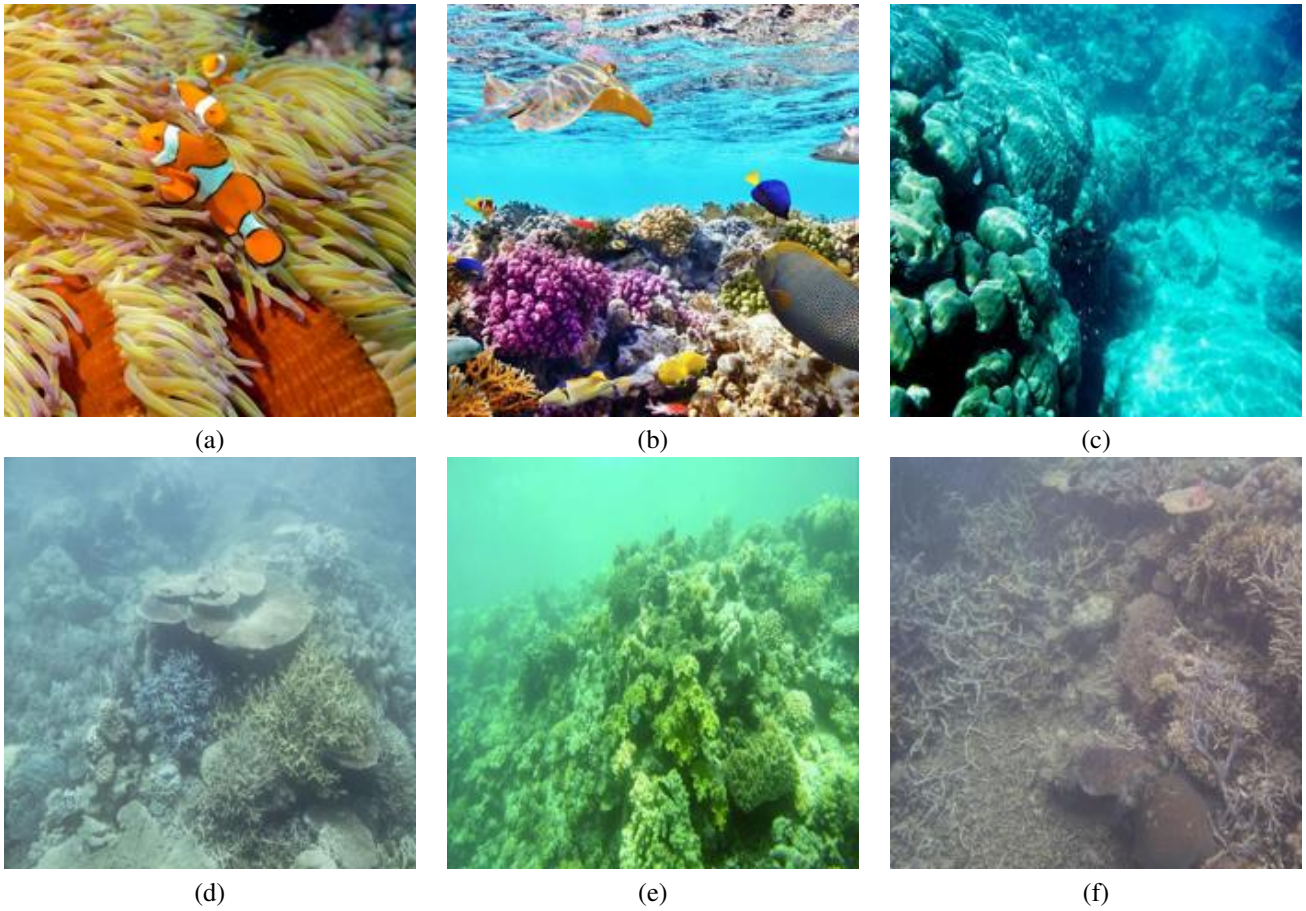

Fig. 1. Some example images from GBR dataset with score $\mu(\pm\sigma)$, where μ and σ represent Mean and Standard Deviation (SD) of score, respectively. (a) high aesthetics and low SD (μ = 9.583, SD = 0.64), (b) high aesthetics and low SD (μ = 9.462, SD = 0.746), (c) high aesthetics and high SD (μ = 5.916, SD = 3.773), (d) low aesthetics and low SD (μ = 2.273, SD = 1.42), (e) low aesthetics and low SD (μ = 3.0, SD = 1.0), (f) low aesthetics and high SD (μ = 3.454, SD = 3.23)

ter than previously proposed non-CNN approaches [20], [21]. Moreover, having access to pre-trained models (e.g. ImageNet [22]) for network training initialization and fine-tuned the network on subject data of image aesthetic assessment have been proven more effective technique for typical deep CNN approach.

[23]

II. Related Work

Recently, deep learning methods have shown great success in various computer vision tasks, such as object recognition [24], [22], [25], object detection [26], [27], and image classification [28]. Deep learning methods, such as deep convolutional neural network and deep belief network, have also been applied to photo quality/aesthetics assessment and have shown good results []. As most deep neural network architectures require fixed-size inputs, recent methods [] transform input images via cropping, scaling, and padding, and design dedicated deep network architectures, such as double-column or multi-column networks, to simultaneously take multiple transformed versions as input.

III. Methodology

We propose two novel Deep CNN architecture for image aesthetics assessment adapted from the recently published state-of-the-art image classification model ([29], [23]). Both the models used in our experiments have been well tested as image classifier for a large number of classes. In experiments, our aim for predictions with higher correlation with human ratings, instead of classifying images to low/high score or mean score regression, the distribution of ratings are predicted as a histogram [30]. The squared EMD (Earth Mover's Distance) loss-based assessment was proposed by Talebi et al.[30], which shows a performance boost inaccurate prediction of the mean score. All network models for aesthetic assessment are based on image classifier architectures. Two differnt architectures (with and without attention mechanism) state-of-the-art networks are explored for the proposed applications. Networks used in our experiments were first trained and then fine-tuned using the large-scale aesthetics assessment AVA dataset [20]. The AVA dataset has 250,000 images, which is very useful for training such a large deep neural network model. The complete architecture of this project consists of different sub-modules, and each of these sub-modules consists of building

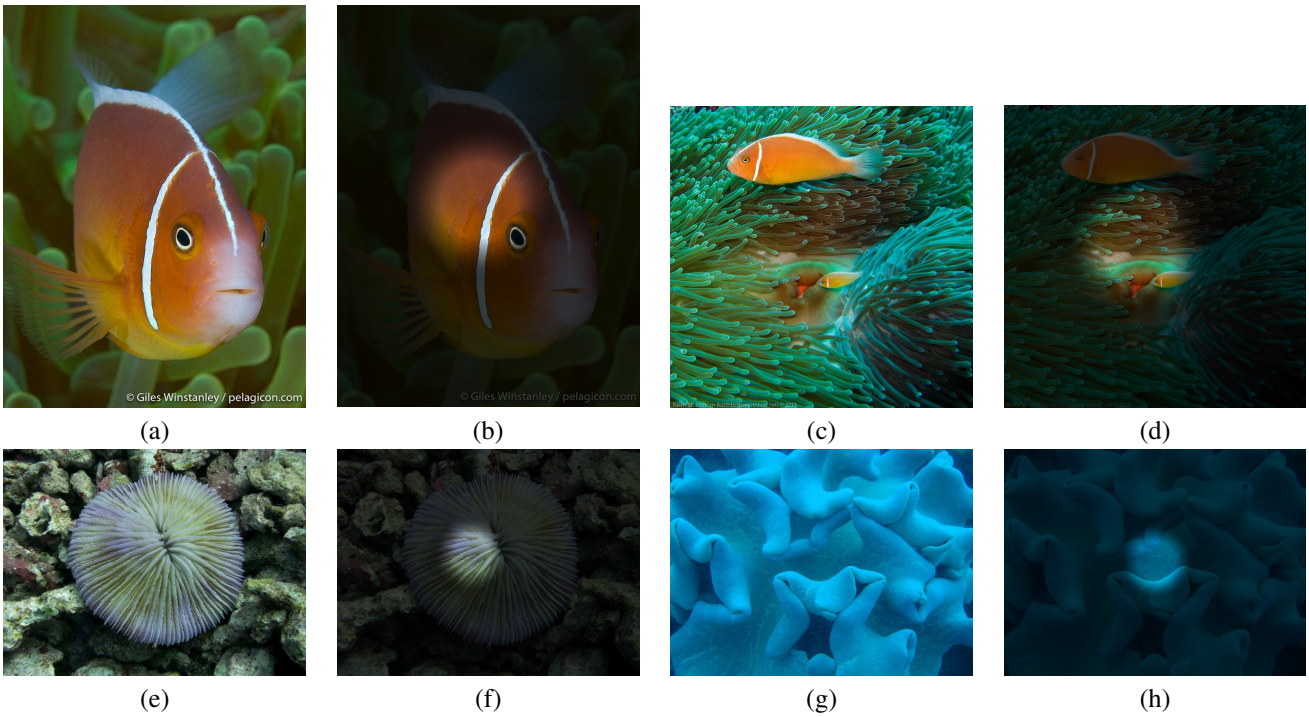

Fig. 2. Four example images from GBR dataset along with same processed images after studying eye movements. (a), (c), (e), (g) are original images. (b), (d),(f), (h) are masked images of (a), (c), (e), (g) respectively, after masked using attention patch information obtained from the Tobii eye-tracking device.

blocks, such as pooling, filters, activation functions, and so forth. The following sections provide more information on the sub-modules. A more detailed description of the architecture and different modules is provided below.

ResNet: Theoretically, neural networks should get better results as added more layers. A deeper network can learn anything a shallower version of itself can, plus possibly some more parameters. The intuition behind adding more layers to a deep neural network was that more layers progressively learn more complex features. The first, second, third, layers learn features such as edges, shapes, objects, respectively, and so on. He et al. [29] empirically presented that there is a maximum threshold for depth with the traditional CNN model. As more layers are added, the network gets better results until at some point; then as continue add extra layers, the accuracy starts to drop. The reason behind failures of the very deep CNN was mostly related to optimization function, network weights initialization, or the well-known vanishing/exploding gradient problem. Vanishing gradients are especially blamed, however, He et al. [29] argue that the use of Batch Normalization ensures that the gradients have healthy norms. In contrast, deeper networks are harder to optimize due to add more difficulty in the process of training; it becomes harder for the optimization to find the right parameters.

The problem of training very deep networks has been attenuated with the introduction of a new neural network layer -The Residual Block 3. Residual Networks attempt to solve this issue by adding the so-called skip connections. A skip connection is depicted in Fig 3. If, for a given dataset, there are no more things a network can learn by adding more layers

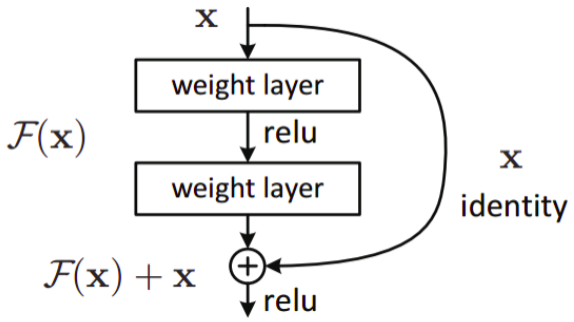

Fig. 3. Residual learning: a building block.

to it, then it can just learn the identity mapping for those additional layers. In this way, it preserves the information in the previous layers and can not do worse than shallower ones. So, the most important contribution to ResNet architecture is the 'Skip Connection' identity mapping.

This identity mapping does not have any parameters and is just there to add the output from the previous layer to the layer ahead. However, sometimes x and F(x) will not have the same dimension. Recall that a convolution operation typically shrinks the spatial resolution of an image, e.g. a 3x3 convolution on a 32 x 32 image results in a 30 x 30 image. The identity mapping is multiplied by a linear projection W to expand the channels of shortcuts to match the residual. This allows for the input x and F(x) to be combined as input to the next layer. A block with a skip connection as in the

image above is called a residual block, and a Residual Neural Network (ResNet) is just a concatenation of such blocks.

$$y = F(x, W_i) + W_s x$$

The above equation shows when F(x) and x have a different dimensionality such as 32x32 and 30x30. This Ws term can be implemented with 1x1 convolutions, this introduces additional parameters to the model. The Skip Connections between layers add the outputs from previous layers to the outputs of stacked layers. This results in the ability to train much deeper networks than what was previously possible. He et al. [29] proposed their network with 100 and 1,000 layers and tested on bench mark datasets such as CIFAR-10, ImageNet dataset with 152 layers and achieved state-of-the-art performance.

Residual Attention Network: Residual Attention Network (Fig 4), a convolutional neural network that incorporates both attention mechanism and residual units which can incorporate with state-of-art feed forward network architecture in an end-to-end training fashion. Residual Attention Network is constructed by stacking multiple Attention Modules which generate attention-aware features. Residual unit is a basic component that utilizes skip-connections to jump over few layers with nonlinearities and batch normalizations which is the prominent feature is the attention module.

Each Attention Module (see Fig. 4) is divided into two branches: mask branch and trunk branch. The trunk branch performs feature processing with Residual Units and can be adapted to any state-of-the-art network structures. Mask Branch uses bottom-up top-down structure softly weight output features with the goal of improving trunk branch features. The Bottom-Up step collects global information of the whole image by downsampling (i.e. max pooling) the image. The Top-Down step combines global information with original feature maps by upsampling (i.e. interpolation) to keep the output size the same as the input feature map. Inside each Attention Module, bottom-up top-down feedforward structure is used to unfold the feedforward and feedback attention process into a single feedforward process. The attention-aware features from different modules change adaptively as layers going deeper. Importantly, an attention residual learning to train very deep Residual Attention Networks which can be easily scaled up to hundreds of layers. The experiment also demonstrates that Residual Attention Network is robust against noisy labels.

In Residual Attention Network , a pre-activation Residual Unit [31], ResNeXt [32] and Inception [33] are used as basic unit to construct Attention Module. Given trunk branch output T(x) with input x, the mask branch uses bottom-up top-down structure [34], [35], [36], [35] to learn same size mask M(x) that softly weight output features T(x). The bottom-up top-down structure mimics the fast feedforward and feedback attention process. The output mask is used as control gates for neurons of trunk branch similar to Highway Network [37]. The output of Attention Module H is:

$$H_i, c(x) = M_i, c(x) \times T_i, c(x)$$

IV. Results and Discussions

Dataset: For experimental purposes, both publicly available datasets and dataset developed in-house were used. For the dataset specific to the Great Barrier Reef (i.e. the GBR dataset), we used 5,417 underwater GBR images, which were downloaded from the Flickr social media platform. These images were sorted based on the content and then rated by participants in an online survey for their aesthetic beauty. At least 10 survey participants provided an aesthetic score for each image and the mean score was calculated. Most of the images (i.e. 80%) served as training material to enable the proposed Neural Network model to learn key feature parameters. The remaining 20% of the data were used during the test and validation phases. The validation dataset helps to understand the system performance in terms of accuracy during the training phase, whereas the test set is normally used once the training phase is completed and ready for deployment. To better understand the distribution of 'beautiful' and 'ugly' pictures in the dataset, Figure 5 presents the number of images with scores above or equal/below 5. More than 2,080 images were scored as highly aesthetic ($score > 5$) and only 420 ($score <= 5$) images were scored as having low aesthetics. Figure 5 also shows how many images of high and low scores were used in each experimental stage.

The GBR dataset of 5,471 images is comparatively small for training a multi-layered deep Convolutional Neural Network. It was, therefore, necessary to complement the GBR data with a large-size, publicly available dataset (AVA, see Murry et al., 2012). This helped to train the system and allowed us to use the in-house GBR dataset for fine-tuning the algorithm. The detailed dataset description of the AVA [20] is described below.

- AVA1: We adopted the score of 0.5 (mean aesthetic score ranges between 0 and 1) as the threshold value to divide the dataset into high aesthetic value and low aesthetic value classes. By doing this, we obtained 74,056 images in the low aesthetic value class and 181,447 images in the high aesthetic value class. 229,954 and 25,549 (approximately 10% images) were used for testing system performance.
- AVA2: In a different experimental setup, and to increase the gap between images with high aesthetic and low aesthetic value, all images were sorted based on their mean scores. Then, the top 10% of images were considered as highly aesthetic and the bottom 10% images were classed as low aesthetic. Thus, 51,100 images (approximately, 20% of the full dataset) then formed the AVA dataset that was used for training.

V. Conclusion

The objective of the work was to study the effectiveness of real human attention obtained using an eye-tracking device on deep Convolutional Neural Network architecture for automatic image aesthetic assessment and predicting an aesthetic score for images. The significance of aesthetic evaluation system was studied in details from the literature. A state-of-the-art deep

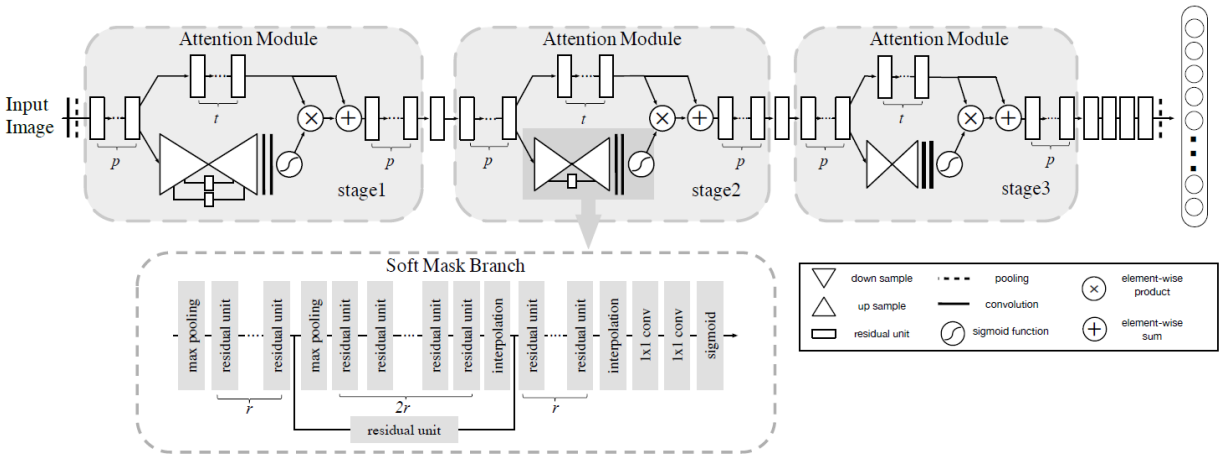

Fig. 4. An architecture the Residual Attention Network Architecture [23]. The output layer is modified to produce image aesthetic score instead class label.

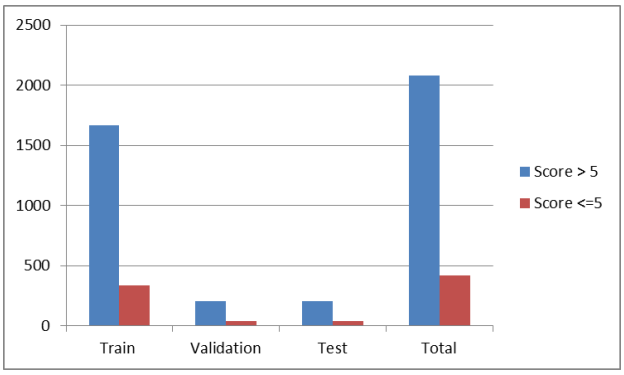

Fig. 5. Score distribution of the GBR dataset developed in-house, containing a total of 5417 images.

TABLE I

PERFORMANCE OF THE PROPOSED METHOD AND FEW RECENTLY PUBLISHED ARCHITECTURES IN PREDICTING IMAGES FROM GBR DATASET. ACCURACY VALUES ARE BASED ON CLASSIFICATION OF PHOTOS TO TWO CLASSES HIGH AND LOW AESTHETICS(COLUMN 2). LCC (LINEAR CORRELATION COEFFICIENT) AND ARE COMPUTED BETWEEN PREDICTED AND GROUND TRUTH MEAN SCORES (COLUMN 3) AND STANDARD DEVIATION OF SCORES (COLUMN 5). EMD MEASURES CLOSENESS OF THE PREDICTED AND GROUND TRUTH RATING DISTRIBUTIONS.

Model	Accuracy	LCC-Mean	LCC-std.dev	EMD
NIMA(MobileNet)	80.36%	0.518	0.152	0.081
NIMA(VGG16)	80.60%	0.610	0.205	0.052
NIMA (Inception-Resnet)	81.51%	0.636	0.233	0.050
ResNet	81.74%	0.641	0.211	0.045
Residual Attention Network	81.74%	0.641	0.211	0.045

Residual Attention Network architecture and ResNet were adapted to our need for the modeling of our GBR aesthetic assessment task. A wide range of experiments was conducted using a comprehensive analysis of performance is presented.

REFERENCES

[1] S. Dhar, V. Ordonez, and T. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *In Proc. Computer Vision and Pattern Recognition*, 2011, pp. 1657–1664.

[2] W. Jiang, A. Loui, and C. Cerosaletti, "Automatic aesthetic value assessment in photographic images," in *In Proc. International Conference on Multimedia and Expo (ICME)*, 2010, pp. 920–925.

[3] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *In Proc. ACM International Conference on Multimedia*, 2014, pp. 457–466.

[4] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *In Proc. Computer Vision and Pattern Recognition*, 2011, pp. 33–40.

[5] Y. Niu and F. Liu, "What makes a professional video? a computational aesthetics approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22(7), pp. 1037–1049, 2012.

[6] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien., "Scenic photo quality assessment with bag of aesthetics-preserving features," in *In Proc. ACM International Conference on Multimedia*, 2011, pp. 1213–1216.

[7] X. Tang, W. Luo, and X.Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15(8), pp. 1930–1943, 2013.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 600–612, 2004.

[9] A. Haas, M. Guibert, A. Foerschner, T. Co, S. Calhoun, E. George, M. Hatay, E. Dinsdale, S. A. Sandin, J. E. Smith, M. Vermeij, B. Felts, P. Dustan, P. Salamon, and F. Rohwer, "Can we measure beauty? computational evaluation of coral reef aesthetics," *PeerJ 3:e1390 https://doi.org/10.7717/peerj.1390*, 2015.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *In Proc. Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893.

[11] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *In Proc. ACM International Conference on Multimedia*, 2010, pp. 271–280.

[12] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *In Proc. European Conference on Computer Vision*, 2006, pp. 288–301.

[13] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *In Proc. International Conference on Computer Vision*, 2011, pp. 1784–1791.

[14] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *In Proc. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[15] J. S. F. Perronnin and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *In Proc. European Conference on Computer Vision*, 2010, pp. 143–156.

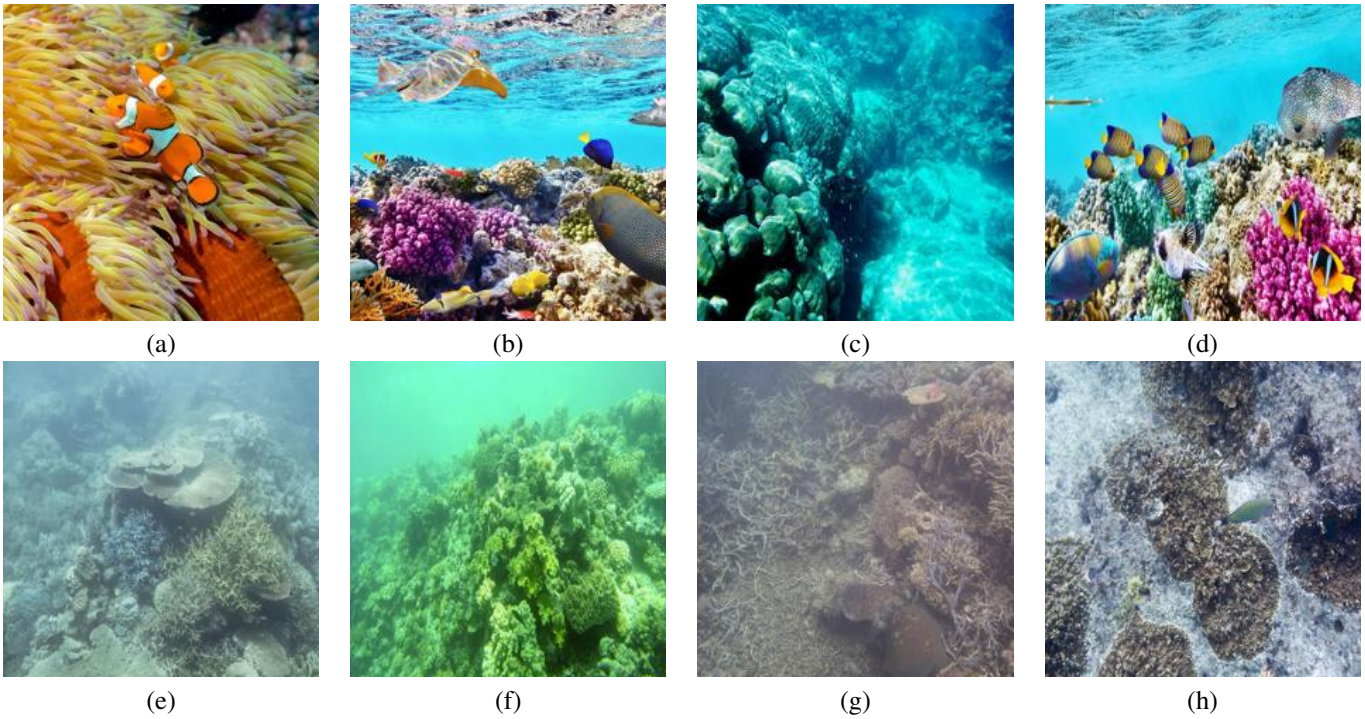

Fig. 6. Ranking some sampled images from GBR dataset with high and low aesthetic values using our proposed aesthetic assessment model using NIMA on Xception. Predicted (and ground truth) scores are shown together for every images (a) 9.583 (9.43), (b) 9.462 (9.22), (c) 5.916 (5.09), (d) 5.916 (5.60), (e) 2.273 (2.55), (f) 3.0 (3.5), (g) 3.454 (3.23), (h) 4.1 (4.22)

[16] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *In Proc. Computer Vision and Pattern Recognition*, 2013, pp. 995–1002.

[17] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *In Proc. Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[18] S. Bosse, D. Maniry, T.Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *In Proc. Computer Vision and Pattern Recognition*, 2016, pp. 3773–3777.

[19] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," in *arXiv preprint arXiv:1602.05531*, 2016.

[20] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *In Proc. Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

[21] N. Ponomarenko, O. Ieremeiev, V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. C. Kuo, "A new color image database tid2013: Innovations and results," in *15th International Conference on Advanced Concepts for Intelligent Vision Systems - Volume 8192*, 2013, pp. 402–413.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, 2015.

[23] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *In Proc. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450–6458.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV*, 2014.

[26] R. B. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[27] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[30] H. T. Esfandarani and P. Milanfar, "NIMA: neural image assessment," *CoRR*, vol. abs/1709.05424, 2017. [Online]. Available: http://arxiv.org/abs/1709.05424

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *CoRR*, vol. abs/1603.05027, 2016. [Online]. Available: http://arxiv.org/abs/1603.05027

[32] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *In Proc. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5987–5995.

[33] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[35] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," *CoRR*, vol. abs/1505.04366, 2015. [Online]. Available: http://arxiv.org/abs/1505.04366

[36] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *CoRR*, vol. abs/1505.07293, 2015. [Online]. Available: http://arxiv.org/abs/1505.07293

[37] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *CoRR*, vol. abs/1507.06228, 2015. [Online]. Available: http://arxiv.org/abs/1507.06228