Learning Photography Aesthetics with Deep CNNs

Gautam Malu*
gautam.malu@research.iiit.ac.in
International Institute of Information Technology
Hyderabad, India

Raju S. Bapi
raju.bapi@iiit.ac.in
International Institute of Information Technology
& University of Hyderabad
Hyderabad, India

Bipin Indurkhya
bipin@iiit.ac.in
International Institute of Information Technology
Hyderabad, India

ABSTRACT

Automatic photo aesthetic assessment is a challenging artificial intelligence task. Existing computational approaches have focused on modeling a single aesthetic score or class (good or bad photo), however these do not provide any details on why the photograph is good or bad; or which attributes contribute to the quality of the photograph. To obtain both accuracy and human-interpretability, we advocate learning the aesthetic attributes along with the prediction of the overall score. For this purpose, We propose a novel multi-task deep convolution neural network (DCNN), which jointly learns eight aesthetic attributes along with the overall aesthetic score. We report near-human performance in the prediction of the overall aesthetic score. To understand the internal representation of these attributes in the learned model, we also develop the visualization technique using back propagation of gradients. These visualizations highlight the important image regions for the corresponding attributes, thus providing insights about model's understanding of these attributes. We showcase the diversity and complexity associated with different attributes through a qualitative analysis of the activation maps.

KEYWORDS

Photography, Aesthetics, Aesthetic Attributes, Deep Convolution Neural Network, Residual Networks

1 INTRODUCTION

Aesthetics is the study of science behind the concept and perception of beauty. Although aesthetics of photograph is subjective, some aspect of its depends on the standard photography practices and general visual design rules. With the ever increasing volume of digital photographs, automatic aesthetic assessment is becoming increasingly useful for various applications, such as a personal photo assistant, photo manager, photo enhancement, image retrieval etc. Conventionally, automatic aesthetic assessment tasks have been modeled as either a regression problem (single aesthetic score) [7, 8, 27] or as a classification problem (aesthetically good or bad photograph) [2, 13, 14].

Intensive data driven approaches have made substantial progress in this task, although it is a very subjective and context dependent task. Earlier approaches used custom designed features based on photography rules (e.g., focus, color harmony, contrast, lighting, rule of thirds) and semantic information (e.g., human profile, scene category) from low level image descriptors (*e.g.* color histograms, wavelet analysis) [1, 2, 4, 9, 12, 15, 16, 21, 24] and generic image descriptors [19]. With the evolution of deep learning based techniques, recent approaches have introduced deep convolution neural networks (DCNN) in aesthetic assessment tasks [8, 10, 13, 14, 26].

Although these approaches give near-human performance in classifying whether a photograph is “good" or “bad", they do not give detailed insights or explanation for such claims. For example, if a photograph received a bad rating, one would not get any insights about the attributes (e.g., poor lighting, dull colors etc.) that led to that rating. We propose an approach in which we identify (eight) such attributes (such as Color Harmony, Depth of Field etc.) and report those along with the overall score. For this purpose, we propose a multi-task deep convolution network (*DCNN*) which simultaneously learns the eight aesthetic attributes along with the overall aesthetic score. We train and test our model on the recently released aesthetics and attribute database (*AADB*) [10]. Following are the eight attributes as mentioned in [10] (Figure 1):

(1) *Balancing Element* - Whether the image contains balanced elements.
(2) *Content* - Whether the image has good/interesting content.
(3) *Color Harmony* - Whether the overall color composition is harmonious.
(4) *Depth of Field* - Whether the image has shallow depth of field.
(5) *Light* - Whether the image has good/interesting lighting.
(6) *Object Emphasis* - Whether the image emphasizes foreground objects.
(7) *Rule of Thirds* - Whether the image follows rule of thirds principle. The rule of thirds involves dividing the photo into 9 parts with 2 vertical and 2 horizontal lines. The important elements and leading lines are placed on/near these these lines and intersections of these lines.
(8) *Vivid Color* - Whether the image has vivid colors, not necessarily harmonious colors.

We also develop attribute activation maps (Figure 3) for visualization of these attributes. These maps highlight the salient regions for the corresponding attribute, thus providing us insights about the representation of these attributes in our trained model.
In summary, followings are the main contributions of our paper:

[*] Corresponding author

MAICS'17, Purdue University Fort Wayne
© 2017 Copyright held by the owner/author(s). .
DOI:

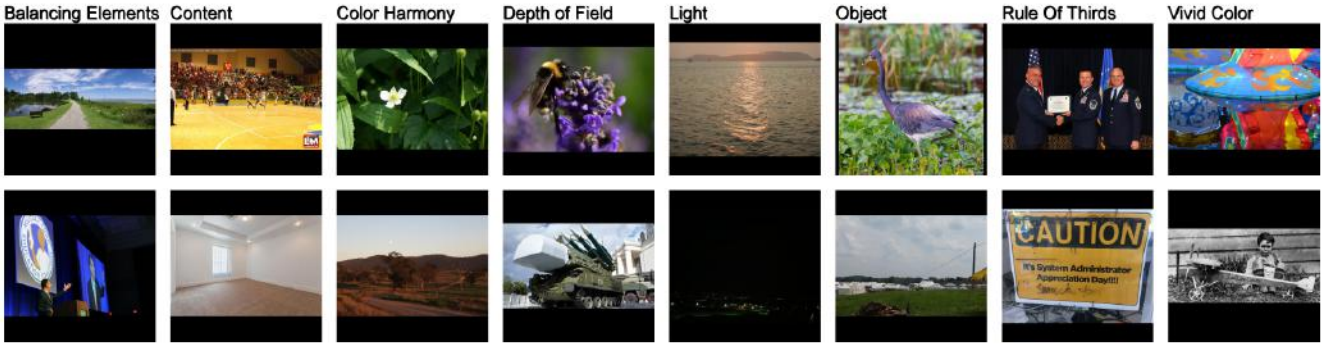


Figure 1: Sample images taken from AADB for each attribute. top row: Highest rated images, bottom row: Lowest Rated Images. All images were padded to maintain aspect ratio for illustration purposes.

(1) We propose a novel deep learning based approach which simultaneously learns eight aesthetic attributes along with the overall score. These attributes enable us to provide more detailed feedback on automatic aesthetic assessment.

(2) We also develop localized representation of these attributes from our learned model. We call these *attribute activation maps* (Figure 3). These maps provide us more insights about model's interpretability of the attributes.

2 RELATED WORK

Most of the earlier works have used low-level image features to design high level aesthetic attributes as mid-level features and trained aesthetic classifier over these features. Datta *et al.* [2] proposed 56 visual features based on standard photography and visual design rules to encapsulate aesthetic attributes from low-level image features. Dhar *et al.* divided aesthetic attributes into three categories Compositional (*e.g.* Depth of field, Rule of thirds), Content(*e.g.* faces, animals, scene types), Sky-Illumination (*e.g.* clear sky, sunset sky). They trained individual classifiers for these attributes from low-level features (*e.g.* color histograms, center surrounding wavelets, haar features) and used outputs of these classifiers as input features for the aesthetic classifier.

Marchesotti *et al.* [18], proposed to learn aesthetic attributes from textual comments on the photographs using generic image features. Despite increased performance, many of these textual attributes (good, looks great, nice try) do not map to well-defined visual characteristics. Lu *et al.* [13] proposed to learn several meaningful style attributes, and used these to fine-tune the training of aesthetics classification network. Kong *et al.* [10] proposed attribute and content adaptive DCNN for aesthetic score prediction.

However, none of the previous works report the aesthetic attributes themselves. These attributes are used as features to predict the overall aesthetic score/class. In this paper, we learn aesthetic attributes along with the overall score, not just as intermediate features but as auxiliary information. Aesthetic assessment is relatively easier in images with evident high and low aesthetics than in ordinary images with marginal aesthetics (Figure 2). For these images, attributes information would greatly supplement the quality of feedback from an automatic aesthetic assessment system.

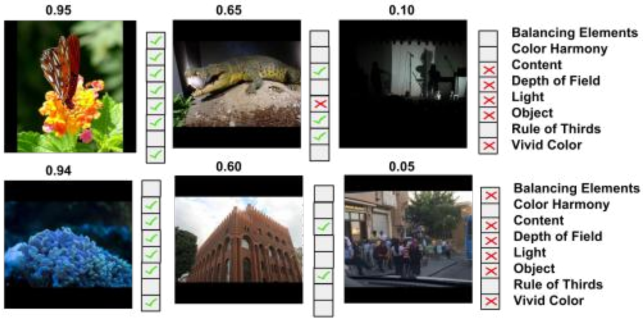


Figure 2: Sample images from AADB testing data. first column: Images rated high on aesthetic score, second Column: Images rated at mid-level, third Column: Images rated low.

Recently, deep learning techniques have shown significant performance gains in various computer vision tasks such as object classification, localization [11, 23, 25]. In deep learning, non-linear features are learned in a hierarchical fashion in increasing complexity (e.g. colors, edges, objects). The aesthetic attributes can be learned as combinations of these features. Deep learning techniques have shown significant performance gains in comparison with traditional machine learning approaches for aesthetic assessment tasks [8, 10, 13, 14, 26]. Unlike traditional machine learning techniques, features are also learned during training in deep learning techniques. However these internal representations of DCNNs are still opaque. Various visualization techniques [5, 17, 22, 28–30] have been proposed to visualize the internal representations of DCNNs in an attempt to have a better understanding of their working. However, these visualization techniques have not been applied in aesthetic assessment tasks. In this article, we apply the gradient based visualization technique proposed by Zhou *et al.* [30] to obtain attribute activation maps. These maps provide localized representation of these attributes. Additionally we also apply similar visualization technique [22] to the model provided by Kong *et al.* [10] to obtain similar maps for qualitative comparison of our results with the earlier approach.

3 METHOD

3.1 Architecture

We use the deep residual network (ResNet50) [6] to train all the attributes along with the overall aesthetic score. ResNet50 has 50 layers which can be divided into 16 successive residual blocks. Each residual block contains 3 convolution layers followed by the batch normalization layer (Figure 3). Each residual block is followed by a rectified linear activation layer (ReLU) [20]. We take these rectified convolution maps from the ReLU output of all these 16 residual blocks, and pool features from each of these 16 blocks with a global average pooling (GAP) layer. GAP layer gives the spatial average of these rectified convolution maps. Then we concatenate all these pooled features and use this as a feature for a fully connected layer which produces the desired outputs (aesthetic attributes and the overall score) as shown in Figure 3. We model the attribute and score prediction as a regression problem with mean squared error as loss function. Due to this simple connectivity structure, we are able to identify the importance of image regions by projecting the weights of the output layer on to the rectified convolution maps, a technique we call *attribute activation mapping*. This technique was first introduced by Zhou *et al.* [30] to get class activation maps for different semantic classes in image classification task.

3.2 Attribute Activation Mapping

For a given image, let $f_k(x, y)$ represent the activation of unit k in the rectified convolution map at spatial location *(x, y)*. Then, for unit k, the result of performing global average pooling is $F^{\mathrm{k}} = \sum_{x,y} f_k(x, y)$. Thus, for a given attribute a, the input to the regression layer, R_a, is $\sum_x w_k^a F_k$ where w_k^a is the weight corresponding to attribute a for unit k. Essentially, w_k^a indicates the importance of F_k for attribute a as shown in Figure 3.

We also synthesized similar attribute maps from the model proposed by Kong *et al.* [10]. We did not have the final attribute and content adapted model from [10] due to patent rights but Kong *et al.* shared the attribute adapted model with us. That model is based on alexnet architecture [11] consisting of fully connected layers along with convolution layers. In this architecture, outputs of convolution layers are separated from desired outputs by three stacked fully connected layers. The outputs from last FC layer are regression scores of attributes. In this architecture we compute weight of layer k for attribute a as summation of gradients (g_k^a) of outputs with respect to k^{th} convolution layer $w_k^a = \sum_{x,y} g_k^a(x, y)$. This technique was first introduced by Selvaraju *et al.* [22] to get class activation maps for different semantic classes and visual explanation (answers for questions).

3.3 Implementation Details

Out of 10000 samples present in the AADB dataset, we have trained our model on 8500 training samples. 500 and 1000 images have been set aside for validation and testing purposes, respectively. As the number of training samples (8500) is not adequate for training of such a deep network (23,715,852 parameters) from scratch, we used a pre-trained ResNet50. It was trained on 1000-class Imagenet classification dataset [3] with approximately 1.2 million images. We fixed the input image size to 299 × 299. We used *horizontal flip* of the input images as a data augmentation technique. The last residual block gives convolution maps of size 10 × 10, so we reduce the sizes of the convolution maps from the previous Res-Blocks to the same size with appropriate sized average pooling. As ResNet50 has batch normalization layers, it is very sensitive to batch size. We fixed the batch size to 16 and trained it for 16 epochs. We report our model's performance on test set (1000 images) provided in AADB. We have made our implementation publicly available [1].

3.4 Dataset

As mentioned earlier, we have used the aesthetics and attribute database (AADB) provided by Kong *et al.* [10]. AADB provides overall ratings for the photographs along with the ratings on the eleven aesthetic attributes as mentioned in [10] (Figure 1). Users were asked to provide information about the effectiveness of these attributes on the overall aesthetic score. For example, if *object emphasis* is positively contributing towards the overall aesthetics of a photograph, user will give a score of *+1* for the attribute, if object is not emphasized adequately and this is contributing negatively towards the overall aesthetic score of the photograph, user will give a score of *-1* for the attribute (See Fig 5). The users also rated the overall aesthetic score on a scale of 1 to 5, with 5 being the most aesthetically pleasing score. Each image was rated by at least *5* persons. The mean score was taken as the ground truth score for all attributes and the overall score.

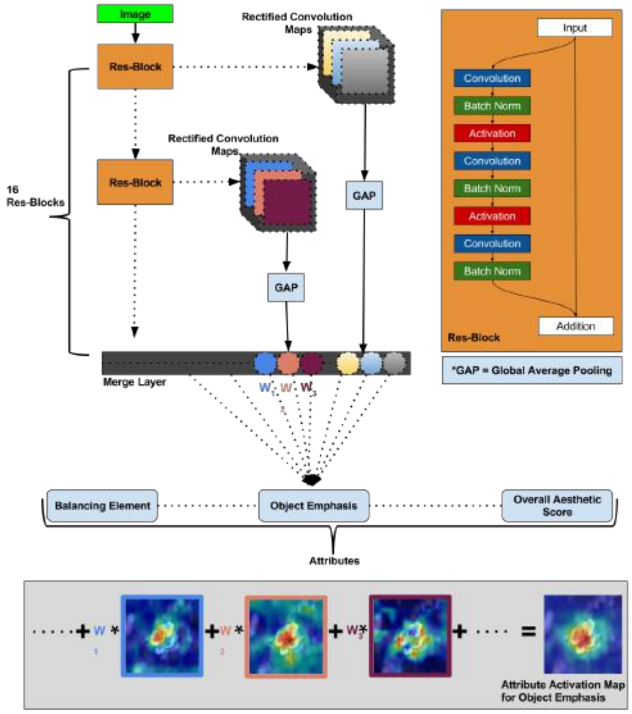


Figure 3: Our approach for generating attribute activation maps. The predicted score for a given attribute (object emphasis in the figure) is mapped back to the rectified convolution layers to generate the attribute activation maps. These maps highlight the attribute-specific discriminative regions as shown in the bottom section.

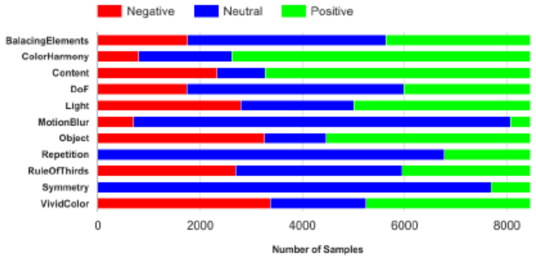
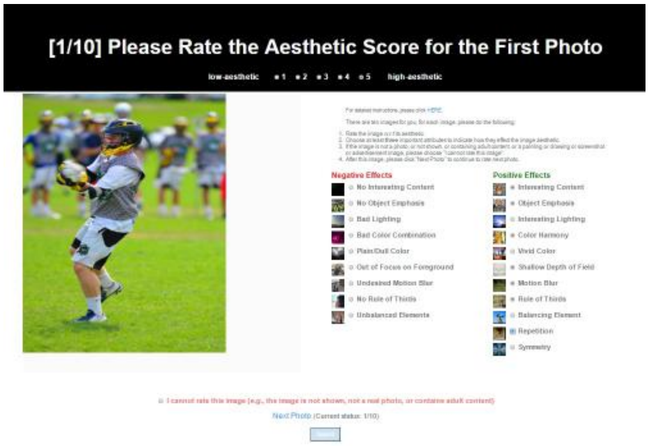


Figure 4: The distribution of all the eleven attributes in the training data of AADB. Most of the images are rated neutral for motion blur, repetition and symmetry.

[1] https://github.com/gautamMalu/Aesthetic_attributes_maps



[1/10] Please Rate the Aesthetic Score for the First Photo

low-aesthetic ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 high-aesthetic

For detailed instructions, please click HERE.

There are ten images for you; for each image, please do the following:

1. Rate the image w.r.t its aesthetic.
2. Choose at least three important attributes to indicate how they effect the image aesthetic.
3. If the image is not a photo, or not shown, or containing adult content or a painting or drawing or screenshot or advertisement image, please choose "I cannot rate this image".
4. After this image, please click "Next Photo" to continue to rate next photo.

Negative Effects	Positive Effects
○ No Interesting Content	○ Interesting Content
○ No Object Emphasis	○ Object Emphasis
○ Bad Lighting	○ Interesting Lighting
○ Bad Color Combination	○ Color Harmony
○ Plain/Dull Color	○ Vivid Color
○ Out of Focus on Foreground	○ Shallow Depth of Field
○ Undesired Motion Blur	○ Motion Blur
○ No Rule of Thirds	○ Rule of Thirds
○ Unbalanced Elements	○ Balancing Element
	☑ Repetition
	○ Symmetry

□ I cannot rate this image (e.g., the image is not shown, not a real photo, or contains adult content)

Next Photo (Current status: 1/10)

Figure 5: Interface of data collection adopted by Kong *et al.*[10].

If an attribute has enhanced the image quality, it was rated positively and if the attribute has degraded the image aesthetics it was rated negatively. The default zero (null) means the attribute does not affect the image aesthetics. For example, positive *vivid color* means the vividness of the color presented in an image has a positive effect on the image aesthetics; while the negative *vivid color* means the image has dull color composition. All the attributes except for *Repetition* and *Symmetry* are normalized to the range of [-1, 1] *Repetition* and *Symmetry* are normalized to the range of [0, 1], as negative values are not justifiable for these two attributes. The overall score is normalized to the range of [0, 1].Out of these eleven attributes, we omit *Symmetry*, *Repetition* and *Motion blur* attributes from our experiment as most of the images rated null for these attributes (Figure 4). We model the other eight attributes along with the overall aesthetic score as a regression problem.

4 RESULTS & DISCUSSION

To evaluate the aesthetic attribute scores predicted by our model, we report the Spearman's ranking correlation coefficient (ρ) between the estimated aesthetic attribute score and the corresponding ground truth score for the testing data. The ranking correlation coefficient (ρ) evaluates the monotonic relationship between estimated scores and ground truth scores, hence there is no need of explicit calibration between them. The correlation coefficient lies in the range of [-1, 1], with greater values corresponding to higher correlation and vice-versa. For baseline comparison, we also train a model by fine tuning a pre-trained ResNet50 and label it as ResNet50-FT. Fine-tuning here refers to modifying the last layer of the pre-trained ResNet50 [6] and training it for our aesthetic attribute prediction task. Table 1 lists the performance on AADB using the two approaches. We also report the performance of the model shared by Kong *et al.*[10].

It should be noted that the spearman's coefficient between the estimated overall aesthetic score and the corresponding ground truth reported by Kong *et al.*[10] was 0.678. They did not report any metrics for the other aesthetic attributes. They used ranking loss along with mean squared error as loss functions. Their final approach was also content adaptive. As can be seen from the results reported in Table 1, our model managed to outperform their approach in overall aesthetic score in-spite of only being trained with mean square error and without any content adaptive framework . Our model significantly underperformed for *Rule of Thirds* and *Balancing elements* attributes. These attributes are location sensitive attributes. *Rule of thirds* deals with positioning of the salient elements, *Balancing Elements* deals with relative positioning of objects with each other and the frame. In our model, due to use of global average pooling (GAP) layers after activation layers we are losing location specificity. We selected GAP layer to reduce the number of parameters. The number of training samples (8500) allows learning of only small parameter space. We also warp the input images to the fixed size input (299x299), thus destroying the aspect ratio. These could be possible reasons for the under-performance of the model for these compositional and location sensitive attributes. Across all the attributes, our proposed method reports better results than ResNet50 fine-tuned model. Our model performs better than the model provided by Kong et al. [10] for five-out-of-eight attributes.

Table 1: Spearman's rank correlations for all the attributes. All correlation coefficients (ρ) are significant at $p < 0.0001$. The coefficients marked with a * are best results for respective attributes.

Attribute	ResNet50-FT	Kong *et al.*[10]	Our method
Balancing Elements	0.184	**0.220***	0.186
Content	0.572	0.508	**0.584***
Color Harmony	0.452	0.471	**0.475***
Depth of Field	0.450	0.479	**0.495***
Light	0.379	**0.443***	0.399
Object Emphasis	0.658	0.602	**0.666***
Rule of Thirds	0.175	**0.225***	0.178
Vivid Colors	0.661	0.648	**0.681***
Overall Aesthetic Score	0.665	0.654[2]/0.678	**0.689***

Table 2: Human performance on AADB. Our model actually outperforms the human consistently (as measured by ρ, last row) averaged across all raters (first row). However, when considering only the "power raters" who have annotated more images, human consistently outperform our model (second and third row).

Number of Images rated	Number of Raters	ρ
> 0	195	0.6738
> 100	65	0.7013
> 200	42	0.7112
Our Approach	–	0.689

Aspects of aesthetic judgments are very subjective in nature. To quantify this subjectivity. In AADB the ground-truth score is the mean score of ratings given by different individuals. To quantify the agreement between ratings, ρ between each individual's ratings and the ground-truth scores was calculated. The average of ρ is reported in Table 2. Our model actually outperforms the human consistently (as measured by ρ) averaged across all raters. However, when considering only the "power raters" who have annotated more images, human evaluators consistently outperform model's results.

5 VISUALIZATION

As mentioned above we generate attribute activation maps for different attributes, to get their localized representations. Here we omit the following attributes, namely, emphbalancing element and the *rule of thirds*, as our model's performance is very low for these attributes as shown in Table 1. For each attribute, we have analyzed the activation maps and present the

[2] The ρ reported by Kong *et al.* [10] for their final content and attribute adaptive model is 0.678, here we are reporting the performance of the model shared by them.

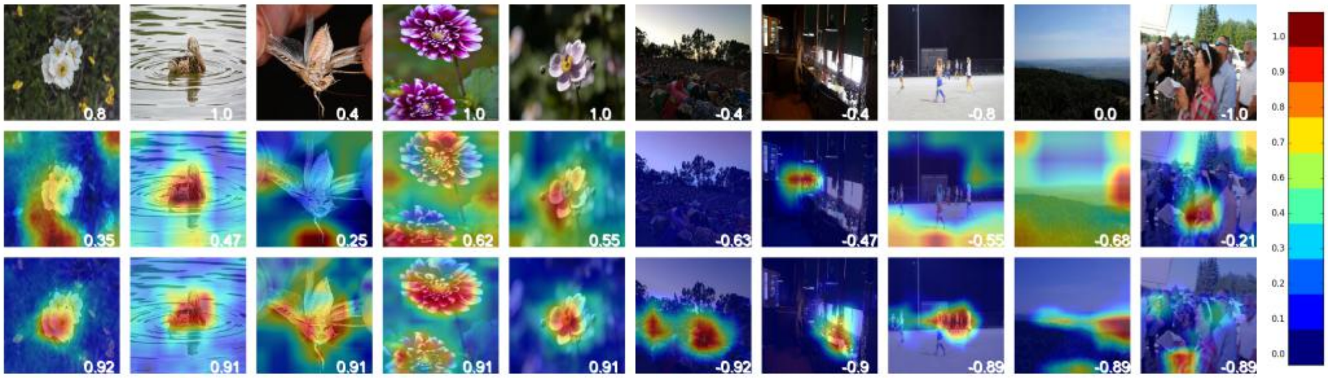


Figure 6: Object Emphasis activation maps. First row: Original Images (marked with ground truth score at the bottom right), second row: Activation Maps from Kong *et al.* [10] model (marked with predicted score from their given model), third row: Activation Maps from our method (marked with our predicted score). Color-bar indicates the color encoding of activation.

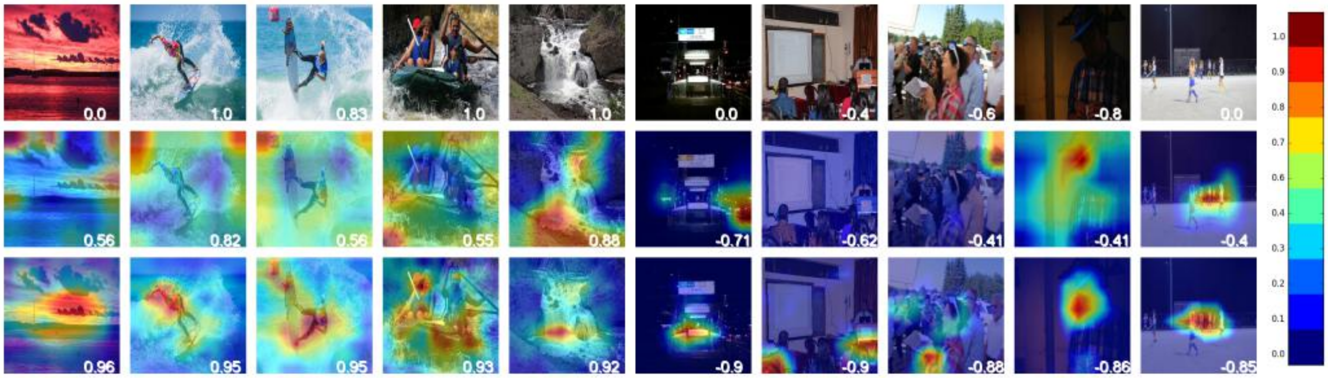


Figure 7: Content activation maps. First row: Original Images (marked with ground truth score at the bottom right), second row: Activation Maps from Kong *et al.* [10] model (marked with predicted score from their given model), third row: Activation Maps from our method (marked with our predicted score). Color-bar indicates the color encoding of activation.

insights in this section. For illustration purposes, We have selected ten samples for each attribute. Out of these ten samples, first five are the highest rated by our model, and the next five are the lowest rated. We have selected these samples from test samples (1000) and not from the train samples. We also have included the activation maps from model given by Kong *et al.* [10] (Kong's model). These activation maps highlight the most important regions for the given attributes. We define these activation maps as "gaze" of the model.

5.1 Object Emphasis

By qualitative analysis of activation maps of object emphasis, it was observed that model gazes at the main object on the image. Even when the model predicts negative rating, i.e. object is not emphasized, the model searches for regions which contain objects Figure 6. In comparison, activation maps from Kong's model are not always consistent as can be seen in the second row of activation maps in Figure 6. It showcases that our model has learned the object emphasis attribute as an attribute which is indeed related to objects.

5.2 Content

Interestingness of content is significantly subjective and is a context-dependent attribute. However, if a model is trained on this attribute, one would expect the model would have maximum activation at the content of the image while making this judgment. If there exists a well-defined object in an image, then that object is considered as the content of the image, for e.g., 2^{nd} and 3^{rd} columns of Figure 7. Further, it can be observed in these columns that our proposed approach is better at identifying the content than Kong's model. Without the presence of explicit objects, the content of the image is difficult to localize, for e.g. 1^{st} and 5^{th} columns of Figure 7. As shown in Figure 7, our model's activation maps are maximally active at the content of the image. In comparison activation maps from Kong's models are not consistent.

5.3 Depth of Field

On analyzing the representations of shallow depth of field, it was observed that model looks for blurry regions near the main object of the image while making the judgment as showcased in Figure 8. Shallow depth of field technique is used to make the subject of the photograph stand out from its background. The model's interpretation of it is in that direction. The images for which model has predicted the negative score on this attribute,

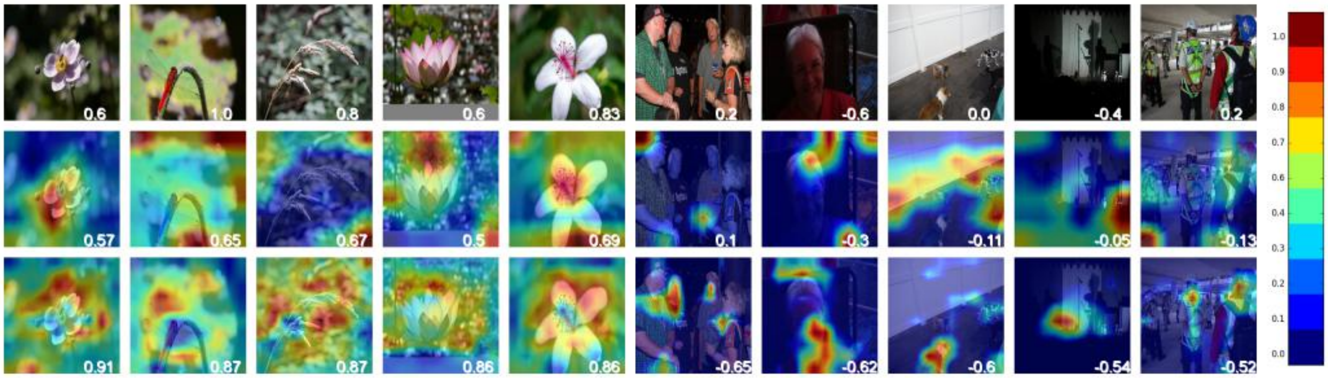


Figure 8: Depth of Field activation maps. First row: Original Images (marked with ground truth score at the bottom right), second row: Activation Maps from Kong *et al.* [10] model (marked with predicted score from their given model), third row: Activation Maps from our method (marked with our predicted score). Color-bar indicates the color encoding of activation.

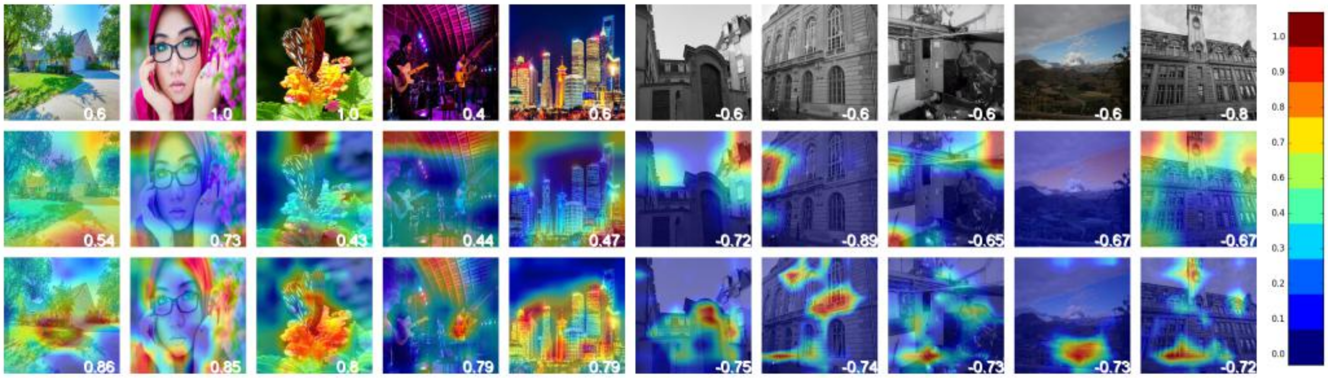


Figure 9: Vivid Color activation maps. First row: Original Images (marked with ground truth score at the bottom right), second row: Activation Maps from Kong *et al.* [10] model (marked with predicted score from their given model), third row: Activation Maps from our method (marked with our predicted score). Color-bar indicates the color encoding of activation.

the activation maps are random. Activation maps from Kong's model also showcase a similar behavior, these maps are more active at the corner of the images.

5.4 Vivid Color

Vivid Color means the presence of bright and bold colors. The model's interpretation of this attribute seems to be along these lines. As shown in Figure 9, model gazes at vivid color areas while making the judgment about this attribute. For example, in 2^{nd} column of the Figure 9 pink color of flowers and scarf, and in 3^{rd} column butterfly and flower were the most activated regions. Authors couldn't find any pattern in activation maps from Kong's model.

5.5 Light

Good Lighting is quite a challenging concept to grasp. It does not merely depend on the light in the photograph, but rather how that light complements the whole composition. As shown in Figure 10, most of the time model seems to look at bright light, or source of the light in the photograph. Although model's behavior is consistent, its understanding of this attribute is incomplete. This was also evident in the low correlation ratings of our proposed model for this attribute, as reported in Table 1.

5.6 Color Harmony

Although model's performance is significant for this attribute, we could not find any consistent pattern in its activation maps. As color harmony is of many types, e.g., analogues, complementary, triadic; it is difficult to get a single representation pattern. For example, in the first example shown in Figure 11, the green color of hills is in analogous harmony with blue color of water and sky; in the 3^{rd} example, brown sand color is in split complementary harmony with blue and green. The attribute activation maps for Color Harmony are shown in Figure 11.

6 CONCLUSION

In this paper, we have proposed deep convolution neural network (DCNN) architecture to learn aesthetic attributes. Results show that estimated scores of five aesthetic attributes (Interestingness of Content, Object emphasis, shallow Depth of Field, Vivid Color, and Color Harmony) correlate significantly with their respective ground truth scores. Whereas in the case of attributes such as Balancing Elements, Light and Rule of Thirds, the correlation is inferior. The activation maps corresponding to the learned aesthetic attributes such as object emphasis, content, depth of field and vivid color indicate that the model has acquired internal representation suitable to

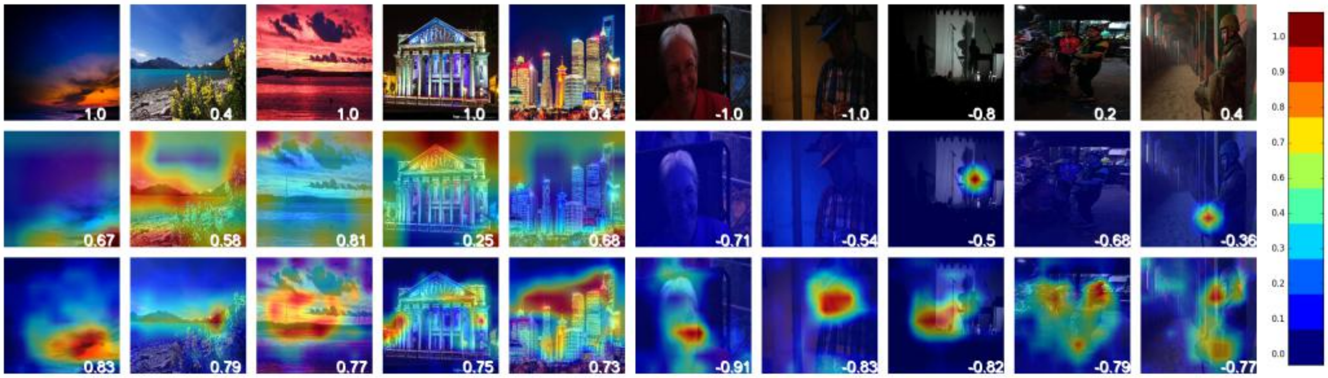


Figure 10: Light activation maps. First row: Original Images (marked with ground truth score at the bottom right), second row: Activation Maps from Kong *et al.* [10] model (marked with predicted score from their given model), third row: Activation Maps from our method (marked with our predicted score). Color-bar indicates the color encoding of activation.

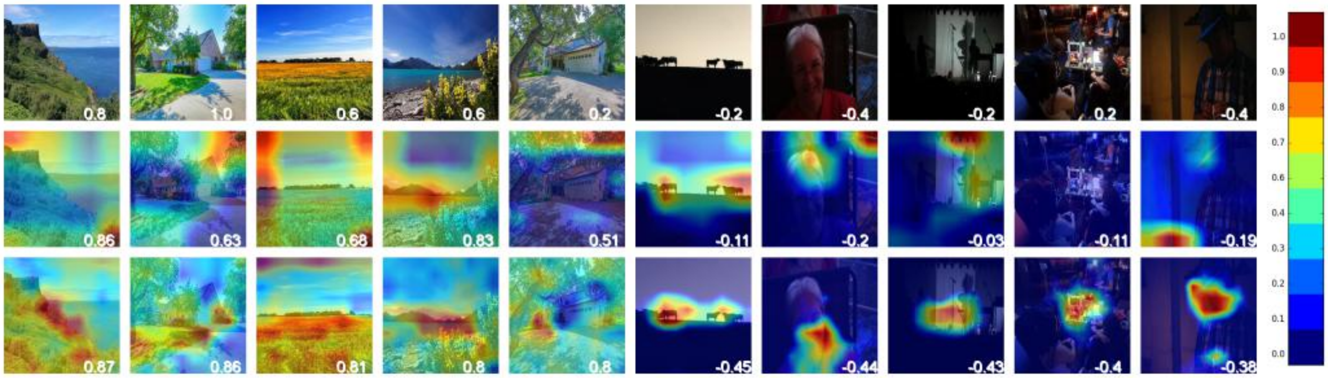


Figure 11: Color Harmony activation maps. First row: Original Images (marked with ground truth score at the bottom right), second row: Activation Maps from Kong *et al.* [10] model (marked with predicted score from their given model), third row: Activation Maps from our method (marked with our predicted score). Color-bar indicates the color encoding of activation.

highlight these attributes automatically. However, for color harmony and light, the visualization maps were not consistent.

Aesthetic judgment involves a degree of subjectivity. For example, in AADB the average correlation between the mean score and an individual's score for the overall aesthetic score is 0.67 2. Moreover, as reported by Kong *et al.* [10], the model learned on a particular dataset might not work on a different dataset. Considering all these factors, empirical validity of aesthetic judgment models is still a challenge. We suggest that the visualization techniques presented in the current work is a step forward in that direction. Empirical validation could proceed by asking subjects to annotate the images (identifying the regions that correspond to different aesthetic attributes) and these empirical maps could in turn be compared with the predicted maps of the model. Such experiments need to be conducted in future to validate the current approach.

ACKNOWLEDGMENTS

The authors would like to thank Ms. Shruti Naik, and Mr. Yashaswi Verma of International Institute of Information Technology, Hyderabad, India for their help in manuscript preparation.

REFERENCES

[1] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 271–280.

[2] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*. Springer, 288–301.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[4] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1657–1664.

[5] Alexey Dosovitskiy and Thomas Brox. 2015. Inverting convolutional networks with convolutional networks. *CoRR abs/1506.02753* (2015).

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[7] Yueying Kao, Kaiqi Huang, and Steve Maybank. 2016. Hierarchical aesthetic quality assessment using deep convolutional neural networks. *Signal Processing: Image Communication* 47 (2016), 500–510.

[8] Yueying Kao, Chong Wang, and Kaiqi Huang. 2015. Visual aesthetic quality assessment with a regression model. In *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 1583–1587.

[9] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE*

Computer Society Conference on, Vol. 1. IEEE, 419–426.

[10] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*. Springer, 662–679.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[12] Li-Yun Lo and Ju-Chin Chen. 2012. A statistic approach for photo quality assessment. In *Information Security and Intelligence Control (ISIC), 2012 International Conference on*. IEEE, 107–110.

[13] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 457–466.

[14] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang. 2015. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 990–998.

[15] Wei Luo, Xiaogang Wang, and Xiaoou Tang. 2011. Content-based photo quality assessment. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2206–2213.

[16] Yiwen Luo and Xiaoou Tang. 2008. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*. Springer, 386–399.

[17] Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5188–5196.

[18] Luca Marchesotti, Naila Murray, and Florent Perronnin. 2015. Discovering beautiful attributes for aesthetic image analysis. *International journal of computer vision* 113, 3 (2015), 246–266.

[19] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 1784–1791.

[20] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[21] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. 2011. Aesthetic quality classification of photographs based on color harmony. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 33–40.

[22] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391* (2016).

[23] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[24] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Shaohui Liu. 2009. Photo assessment based on computational visual attention model. In *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 541–544.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[26] Xinmei Tian, Zhe Dong, Kuiyuan Yang, and Tao Mei. 2015. Query-dependent aesthetic model with deep learning for photo quality assessment. *IEEE Transactions on Multimedia* 17, 11 (2015), 2035–2048.

[27] Ou Wu, Weiming Hu, and Jun Gao. 2011. Learning to predict the perceived visual quality of photos. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 225–232.

[28] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[29] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2014. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856* (2014).

[30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.