Neural Aesthetic Portrait Image Assessment

Xiyuan Wu
Stanford University
450 Jane Stanford Way
xiyuan26@stanford.edu

Cici Hou
Stanford University
450 Jane Stanford Way
xhou@stanford.edu

Abstract

Portraits form a significant portion of personal photo albums, making the development of neural aesthetic assessment specialized for portraits beneficial to everyone's everyday life. To improve upon a general aesthetic model to perform neural aesthetic assessment on portraits, we integrated facial embeddings from the VGG-Face model into the Neural Image Assessment (NIMA) framework. This proved effective, with the enhanced model achieving an accuracy of 0.8341, an LCC of 0.7786 and an SRCC of 0.7849, surpassing the baseline results. Additionally, we conducted interpretability experiments using gradient-weighted class activation mapping (Grad-CAM) and guided backpropagation to visualize the regions of images that contribute most to the model's predictions. These visualizations help validate the model's decision-making process and highlight its focus on relevant facial features.

1. Introduction

With the proliferation of digital photography, organizing and curating personal photo albums has become an essential task for many individuals. Following our original idea of creating an intelligent album organizer, we have identified that portrait photos constitute the majority of images in most people's albums. This observation has led us to focus on refining neural aesthetic models to better assess the aesthetic quality of portrait photographs.

Existing aesthetic models tend to generalize across all image types, which may not adequately capture the unique features and nuances specific to portraits. By developing a specialized model for portraits, we aim to improve the accuracy and relevance of neural aesthetic assessments in personal photo album organising.

The input to our algorithm is a portrait image. We then use a Convolutional Neural Network (CNN) architecture based on MobileNet, augmented with facial recognition embeddings derived from VGG-Face[8], to output a predicted aesthetic score probability distribution on a scale of 1-10. This score is intended to reflect the human perception of aesthetic quality, providing a valuable tool for intelligent photo album organization. We have proven the viability and effectiveness of this approach by our experiments.

We have also conducted interpretability experiments, such as using gradient-weighted class activation mapping (Grad-CAM) and guided backpropagation, to visualize which regions of the images contribute most to the model's predictions. These experiments help in understanding the decision-making process of our model.

2. Problem Statement

We use 4000+ portraits from Aesthetic Visual Analysis Dataset [7] as our primary dataset. We trained a model to perform a neural aesthetic assessment on a portrait which takes in a photo and outputs a probability function of how humans may score it on a scale of 1-10.

Running the NIMA model with MobileNet as a base model with provided weights gives an accuracy of 75.2% when tasked with categorizing images into 2 classes (below and above 5.5). It also achieves an LCC(linear correlation coefficient) of 0.645 and an SRCC(Spearman's rank correlation coefficient) of 0.636. We will compare our neural portrait aesthetic model against these results where accuracy shows we can correctly identify good and bad portrait photos and where LCC and SRCC make sure we can correctly rank the quality of portraits.

3. Related Work

The field of aesthetic quality assessment of images has advanced significantly with deep learning models, leveraging convolutional neural networks (CNNs) and attention mechanisms.

The Aesthetic Visual Analysis (AVA) dataset [7] serves as a cornerstone for many aesthetic assessment studies, comprising over 250,000 images with extensive metadata, including aesthetic scores, semantic labels for over 60 categories, and photographic style labels. This comprehensive dataset provides a solid foundation for training and evaluat-

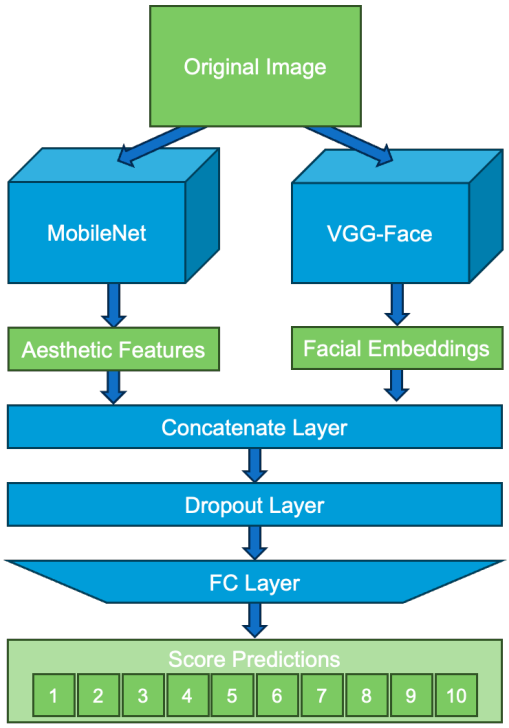


Figure 1. Model Structure

ing models, enabling a nuanced understanding of aesthetic quality across diverse images.

Pioneering work in this domain includes the Adaptive Layout-Aware Multi-Patch Deep CNN (A-Lamp) model by Ma et al. [6], which uses a multi-patch approach to capture both global and local features, providing a comprehensive aesthetic evaluation. Similarly, the Attention-based Multi-Patch Aggregation (MP_ada) model [13] incorporates an attention mechanism to focus on difficult-to-assess features, enhancing prediction accuracy.

Kao et al. [3] introduced a deep convolutional neural network specifically for aesthetic quality assessment, capturing both global and local features for robust evaluation. Lu et al. [5] proposed a deep multi-patch aggregation network, leveraging multiple image patches to improve performance by capturing detailed visual information. Wang et al. [17] presented a brain-inspired deep network, mimicking hierarchical visual processing to incorporate both low-level and high-level features for aesthetic evaluation.

However, models like A-Lamp and MP_ada do not specifically target the unique characteristics of portrait photography, where facial features play a critical role in aesthetic evaluation.

Building on the Neural Image Assessment (NIMA) model by Talebi and Milanfar [15], our work aims to enhance NIMA's performance for portrait photography by incorporating facial recognition and feature embedding. We generate rich facial embeddings, combining them with NIMA's output to focus on critical facial features, thus improving aesthetic assessment of portraits. Inspired by advancements in facial recognition pipelines, Serengil and Ozpinar [12] benchmarked various facial recognition pipelines, highlighting effective modules for integration into aesthetic assessment models.

In visual interpretability, guided backpropagation [14] and Grad-CAM [11] are prominent methods providing visual explanations for model predictions, highlighting important regions in images that contribute to decisions.

Overall, our work extends these approaches by addressing the specific challenges of portrait photography, using facial recognition, feature embedding, and interpretability techniques to create a robust aesthetic assessment model tailored for portraits.

4. Dataset

We used the Aesthetic Visual Analysis (AVA) dataset [7]. The AVA dataset is a comprehensive collection of over 250,000 images, each accompanied by a rich variety of metadata. This metadata includes:

- **Distribution of Ratings:** Each image in the dataset has a distribution of aesthetic ratings on a scale from 1 to 10, provided by multiple human raters. This allows for a nuanced understanding of the perceived aesthetic quality of each image.
- **Semantic Labels:** The dataset includes over 60 categories of semantic labels, offering detailed annotations about the content of the images. These labels help in identifying and categorizing the images based on various visual and contextual elements.
- **Photographic Style Labels:** The AVA dataset also provides labels related to different photographic styles, such as portrait, landscape, macro, and more. These labels are crucial for tasks that involve style-specific analysis or enhancement.

Given our task to examine and optimize the model's performance on portrait photos, we focused specifically on this subset of the AVA dataset. The extraction process involved the following steps. First, we filtered the dataset to identify images labeled with portrait-related tags. This ensured that the selected images were relevant to our specific focus on portrait photography. The filtered set contained over 4000 portrait images. Then, this dataset was further divided into training and testing subsets to facilitate robust model training and evaluation. Each subset had 2000+ images.

By leveraging the AVA dataset and carefully extracting and preparing a high-quality subset of portrait images, we aimed to build and refine a model capable of delivering accurate and interpretable aesthetic assessments specifically for portrait photography.

We used the same preprocessing procedures used by the original NIMA model[15], whereby input images are firstly scaled to 256 × 256, before randomly cropped into 224 × 224. The initial scaling was performed to alleviate the issue of changing composition by cropping directly and the cropping is to avoid over-fitting. Random horizontal flipping is also performed for data augmentation purpose.

5. Methods

Now we will discuss how we generated the validation test set accuracy on the specific portrait category of the AVA dataset and made explorations on our model with reference to the NIMA model.

5.1. Portrait Validation Accuracy

We extracted images from the "portrait" category within the AVA dataset and processed them using the NIMA model to generate the "mean_score_prediction" for each image, identified by their unique "image_id." We then computed the ground truth mean score for each image. We performed a binary comparison to assess the predictive accuracy, where a mean score above 5.5 was categorized as having high aesthetic quality. In contrast, a score below 5.5 was categorized as having low aesthetic quality.

5.2. Facial Embedding

We experimented with facial feature embedding using wrapped facial recognition algorithms called DeepFace, which provides different 4096-dimensional embeddings generated from various facial recognition base models. We primarily used VGG-Face [8] for our experiments. VGG-Face is trained on a large dataset of face images and has demonstrated high accuracy in recognizing and distinguishing facial features. This makes it well-suited for capturing the detailed and nuanced features necessary for assessing the aesthetic quality of portrait photographs. We believe the result could be generalised to other embedding methods, and will experiment and compare performance in the future.

In our model, we concatenate the facial representation vector with the output from NIMA (with the last fully-connected layer removed) and pass them through a fully-connected layer. By doing this, we hope to incorporate detailed facial features directly into the aesthetic assessment process, allowing the model to focus on critical elements such as facial expressions and skin texture.

To further understand the extent of impact of a richer facial embedding on the performance of the model, we used PCA to reduce the dimension of the embeddings and did comparative experiments to show how embeddings of 4096, 2048, 1024, 512, 128 dimensions differently affect portrait aesthetic evaluation.

5.3. Face and Body Cropping

Given that portrait images predominantly feature live subjects, photographers often emphasize these subjects, and viewers naturally focus on them as well. To investigate the impact of this emphasis, we explored the cropping of human faces and bodies and conducted predictions based on these cropped images. Utilizing the "haarcascades" algorithm in OpenCV, we extracted facial regions from the portrait images and input them into our prediction model for evaluation. This approach allowed us to assess the influence of subject-centric cropping on our prediction accuracy.

Figure 2 shows an example of a cropped face and its original image.

(a) Cropped face of a girl. (b) Original image.

Figure 2. Illustration of face croppping on a example image.

Then, we attempted to optimize our model output with face rating. Based on our analysis of the face rating results, we observed a trend where the majority of the data points fall below the line defined by cropped face rating $=$ $1.15 \times$ ground truth. Consequently, we postulate that the face rating can serve as a lower bound for the prediction outcomes. To refine our predictions, we apply an affine transformation to the face rating. By comparing the maximum value between the transformed face rating and the original model's prediction score, we establish a new, adjusted prediction score.

6. Experiments and Results

6.1. Portrait Validation Accuracy

Table 1 provides a detailed summary of the prediction statistics and the corresponding test accuracy metrics. Figure 3 is a visualization of the data where each data point represents an image with its ground truth rating as x-value and portrait rating as y-value. We found a test accuracy rate of 75.15%.

6.2. Facial Embedding

In this section, we will discuss our finding which shows an improved performance of NIMA model with a richer fa-

Mean	5.509
Median	5.542
Standard Deviation	0.544
Minimum Score	3.597
Maximum Score	6.791
Test Accuracy	75.15%

Table 1. Prediction statistics on portrait dataset.

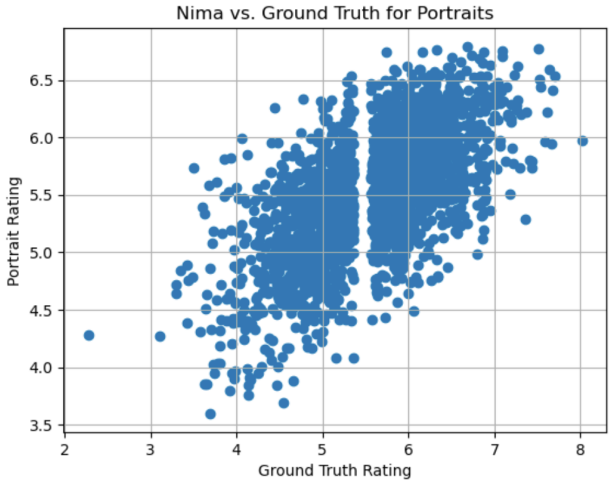


Figure 3. Plot of Nima vs. Ground Truth for Portraits

cial embedding.

6.2.1 Training detail

Training of the model is performed in two stages. In the first stage, only the top layer, namely the concatenated fully-connected layer which takes in MobileNet-embeddings and facial embeddings are trained. After experimenting with different hyper-perimeters, we decide to compare the performance of the improved model on different embedding size with a FC layer training rate of 0.001, dropout rate of 0.75 and train it for 10 epochs. This first stage of training is then followed by the training of the entire network, which includes the weights of the CNN. In this stage, we train the model for 15 epochs with a learning rate of 0.00003. The epoch choice is decided to be such because we realise not capping the epochs of the first training stage tend to over train the dense layer and result in little or no learning of the full model.

We trained the improved NIMA with different embedding size. We have also re-trained the basic NIMA model using the same hyper-parameters for a fairer comparison.

6.2.2 Quantitative Comparison

To evaluate the results of the experiment, we calculate the binary classification accuracy. This is determined by comparing the mean score of the predicted probability distribution with a threshold of 5.5. If the predicted mean score is greater than 5.5, it is classified as positive; otherwise, it is classified as negative. This binary classification accuracy is then compared to the ground truth mean score.

No. Of Embedding	Accuracy	LCC	SRCC
0	0.7718	0.6814	0.6872
128	0.7789	0.6992	0.7026
512	0.7908	0.7229	0.7304
1024	0.8032	0.7404	0.7479
2048	0.8124	0.7518	0.7626
4096	0.8341	0.7786	0.7849

Table 2. Performance metrics for different numbers of embeddings

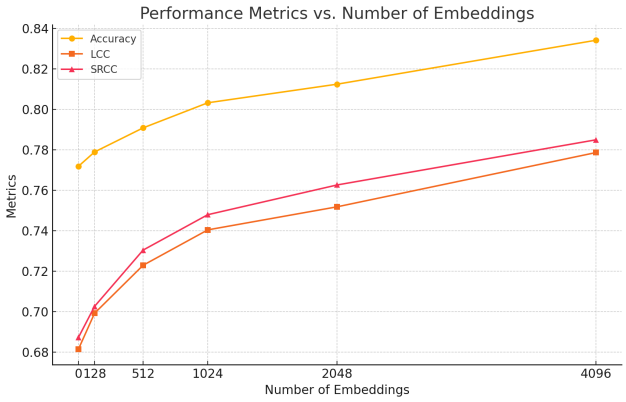


Figure 4. Performance metrics vs. number of embeddings

In addition, we also calculate the Linear Correlation Coefficient (LCC) and the Spearman Rank Correlation Coefficient (SRCC). The LCC measures the linear relationship between the predicted scores and the ground truth scores, indicating how well the predicted scores match the actual scores in a linear sense. A higher LCC value indicates a stronger linear relationship. The SRCC assesses the monotonic relationship between the predicted scores and the ground truth scores. It evaluates how well the predicted rankings of the scores correspond to the actual rankings, irrespective of the linearity. A higher SRCC value indicates a stronger monotonic relationship.

The results, as presented in the table and graph, show that as the number of embeddings increases, all three metrics (accuracy, LCC, and SRCC) significantly improve. This suggests that increasing the number of embeddings enhances the model's performance in terms of both binary classification accuracy and the correlation with ground truth scores.

This trend indicates that a higher number of embeddings allows the model to capture more complex features of the facial features in the portraits, leading to better performance across all metrics. Therefore, our improvements of the model by concatenating with facial features is effective.

6.2.3 Qualitative Comparison

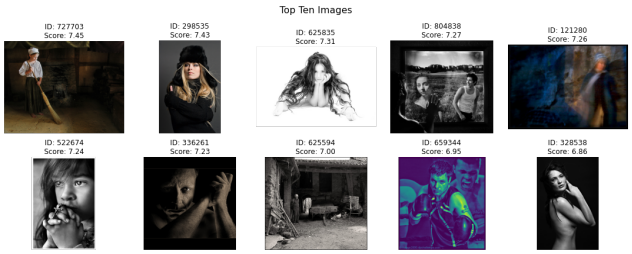


Figure 5. Top 10 images rated by human

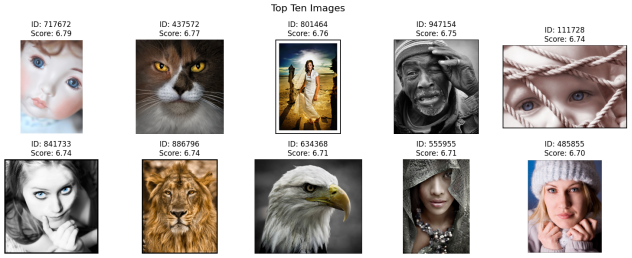


Figure 6. Top 10 images rated by NIMA

Figure 7. Top 10 images with 4096D embeddings

Figure 8. Comparison of Top 10 Images

By observing Figure 8 the top 10 rated images by human, NIMA, and facial-embedding-enhanced NIMA, we can effectively discern the differences in rating criteria, despite all images being valid aesthetic representations.

Both NIMA and enhanced-NIMA tend to prefer portraits with a clear and prominent subject, often framed in close-up shots. In contrast, human selections show more appreciation for a balanced environment and the subject. This may be because our models are built on top of MobileNet and VGG-Face, which were primarily trained for classification and recognition tasks and therefore tend to favor images with well-defined, central subjects. This preference highlights the models' limitation in appreciating a holistic aesthetic. It is also evident that human-preferred images are much more subdued in terms of color compared to those favored by the models, indicating that the models struggle to understand more nuanced aesthetic qualities and prefer stronger technical features.

Figure 9. Worse 10 images with 4096D embeddings

Mean	4.407
Median	4.277
Standard Deviation	0.884
Minimum Score	2.986
Maximum Score	6.739
Test Accuracy	46.57%
Normalized test accuracy	53.54%

Table 3. Prediction statistics of cropped face and body on portrait dataset.

Comparing enhanced NIMA and NIMA, we observe that enhanced NIMA filters out images of animals, likely because the facial features in these images cannot be recognized effectively. On the other hand, when examining Figure 9 the worst 10 images rated by enhanced NIMA, it becomes clear that most have obstructed faces. This observation highlights that, due to the embedding features used in the model, there is a strong preference for clear facial features, which may be a significant limitation in assessing portrait aesthetic holistically.

6.3. Face and Body Cropping

Table 3 provides a detailed summary of the prediction statistics and the corresponding test accuracy metrics for cropped face and body images. Figure 4 is a visualization of the data where each data point represents an image with its ground truth rating as x-value and face rating as y-value.

The initial accuracy of 46.57% was suboptimal. We hypothesize that this is attributable to the down-rating of cropped images, which suffer from insufficient contextual information, thus impairing aesthetic evaluation performance. To address this, we normalized the statistics from both the ground truth and the cropped face images. Upon recalculating the accuracy, we observed an improvement, with the accuracy increasing to 53.54%.

However, visualizing the data distribution in Figure 2, we found that, interestingly, most of the data fit under the cropped face rating $= 1.15 \times$ ground truth line, which means there is a potential threshold of face and body predictions.

Based on our result of face cropping, we explored optimizations with Face Rating. Specifically, we set a minimum

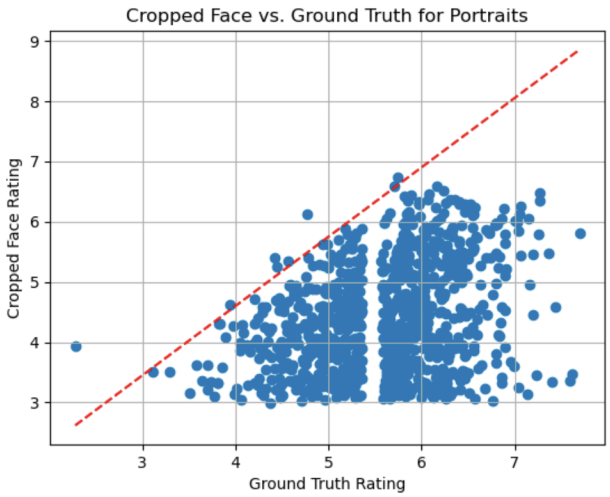


Figure 10. Plot of Cropped Face vs. Ground Truth for Portraits

bar for prediction scores based on the face ratings, defined by the linear equation $threshold = m \times face\ rating + b$. The modified face rating is calculated as $(face\ rating - b)/threshold$. We tried a range of bs and $threshold$s and calculated the resulting accuracy, though we did not see an improvement in the accuracy rate.

7. Visual Interpretability

On top of model enhancement, we also explored providing visual explanations for the decisions made by the NIMA model using the Grad-CAM method, helping to understand which regions of an image contribute most to the model's predictions. We also experimented with guided backpropagation on images to enhance the score of an image.

7.1. Method

7.1.1 Gradient Mapping

We produced 'visual explanations' for decisions from the NIMA model using the Grad-CAM method [11]. This method utilizes the gradients of any target concept flowing into the final convolutional layer to create a coarse localization map. This map highlights important regions in the image that contribute to predicting the concept. We produced maps on our portrait images based on the PyTorch library for CAM methods [2]. We computed the gradients of the rating classes from 1 to 10 concerning the image pixels. These gradients are then averaged with each layer's weight corresponding to their normalized prediction distributions.

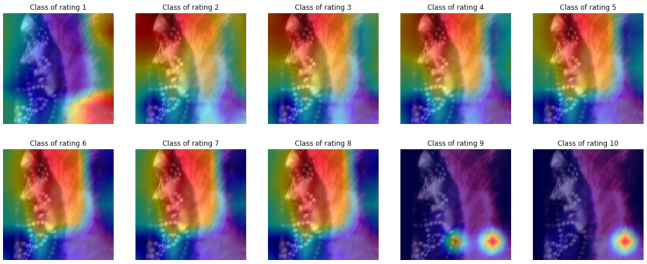
7.1.2 Guided Backpropagation

We employed guided backpropagation to identify the regions of an image that most significantly influence the prediction score, thereby enhancing the overall aesthetics rating of the image [14][18]. The process began with an initial calculation of the gradient for a single step to determine how alterations in the image affect the NIMA score. Following this, the image underwent 100 iterations of guided backpropagation, specifically focusing on higher rating classes (7-10) to maximize the aesthetic score. By selectively amplifying or modifying these critical regions, the overall aesthetic quality of the image was improved.

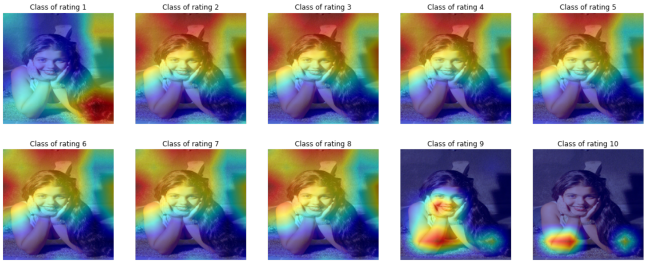
7.2. Result

7.2.1 Gradient Mapping

We generated 'visual explanations' for the NIMA model's decisions using the Grad-CAM method [11]. We created these maps for our portrait images. The explainability varied across images, while some mapping clearly highlighted the subject of the portrait, some generated mappings that did not correlate to the position of the subject.



(a) A case where the mapping appears explainable



(b) A case where the mapping appears unexplainable

Figure 11. GradCAM visualizations by rating class 1-10 for two examples.

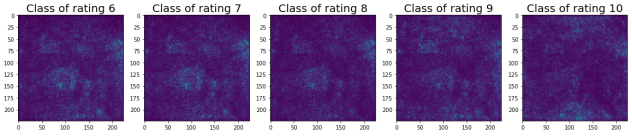
Figure 12. Average GradCAM

Here we provide some visualization examples, Figure 11 shows two groups of gradient maps by classes of ratings ranging from 1 to 10. Figure 12 shows the average mapping based on the score distribution of the rating classes. As shown in Figure 11(a), this set of images shows an instance where the Grad-CAM mapping appears explainable. The gradient maps clearly highlight the important regions in the image that correspond to the face of the subject. In cases such as that depicted in Figure 11(b), where the mapping appears unexplainable, it highlights a significant challenge in the interpretability of deep learning models. When the gradient maps fail to show clear and distinct regions of importance, it becomes difficult to trust the model's predictions, especially in critical applications where understanding the model's decision-making process is essential.

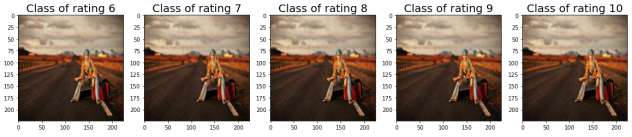
While the explainable mappings prove the strength of NIMA in understanding portrait photos, the unexplainable mappings underscore the need for further development of interpretability methods that can provide more consistent and reliable insights into model behavior. Ensuring that models not only perform accurately but also do so in a manner that is comprehensible to human users is crucial for fostering trust and facilitating the responsible deployment of AI systems. Additionally, the contrast between explainable and unexplainable mappings in these examples emphasizes the importance of continuous evaluation and improvement of interpretability tools like Grad-CAM to achieve better transparency and accountability in AI models.

7.2.2 Guided Backpropagation

We explored improving the mean score of the portraits via guided backpropagation. We picked channels of distribution of scores from 6 to 10 and calculated the gradients of those neuron values with respect to image pixels.



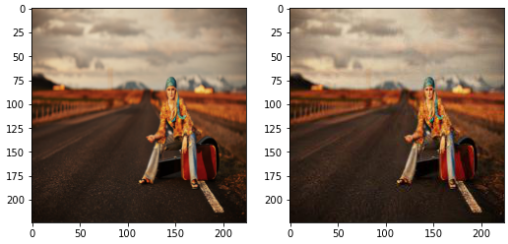
(a) Gradient Map Visualization



(b) Enhanced Images Visualization

Figure 13. Guided Backpropagation Visualization by Rating Class

Figure 14 illustrates the process and results of guided backpropagation used to improve the mean score of portraits. Figure 13(a) shows the gradient maps for classes of ratings 6, 7, 8, 9, and 10. These maps display the distribution of gradients with respect to the image pixels, highlighting the areas that contributed most to the class predictions. However, upon closer inspection, the gradient maps do not emphasize significant shapes. The variations between the maps are subtle, suggesting that the gradients do not provide clear distinctions or improvements across different pixels. Figure 13(b) shows the enhanced images generated using guided backpropagation for the same classes of ratings. Despite the application of guided backpropagation, the enhanced images appear largely unchanged from the original ones.



(a) Original Image: 4.44 (b) Enhanced Image: 6.24

Figure 14. We used 100 steps of guided backpropagation on rating classes 7-10 to increase the mean score of the image passes into the NIMA model. The original image (a) has a prediction score of 4.44, while the enhanced image (b) achieves a score of 6.24.

Based on the calculation for a single step described above, we enhanced the image through 100 iterations of backpropagation. Figure 14 shows our result. While we can not visually recognize significant changes from 14(a) to 14(b), the original image gets a prediction score of 4.44, whereas the enhanced image gets a prediction score of 6.24. Since the evaluation of the image is binary - whether the image rating is above or below 5.5 - the enhanced image is classified as an image with high aesthetics just by making imperceptible modifications.

The enhancement process using guided backpropagation effectively raised the aesthetic rating of an image as measured by the NIMA model while only subtly modifying its appearance. This raises significant questions about the unexplainability inherent in such models. Aesthetic judgment is inherently subjective, varying widely among individuals based on personal preferences, cultural backgrounds, and contextual factors. The NIMA model's training data, which consists of human ratings, may not capture the full diversity of aesthetic preferences. Moreover, the enhancements made by the model reflect an averaged or generalized aesthetic judgment, which might not align with specific individual tastes.

8. Conclusion

In this paper, we presented a specialized neural aesthetic assessment model for portrait photography, building upon

the NIMA model and integrating facial recognition embeddings to enhance its performance. Our approach, leveraging a convolutional neural network architecture based on MobileNet and augmented with facial recognition embeddings from VGG-Face, demonstrated significant improvements in assessing the aesthetic quality of portrait images.

Through extensive experiments, we showed that our model, enriched with detailed facial embeddings, outperformed the baseline NIMA model across several key metrics, including binary classification accuracy, Linear Correlation Coefficient (LCC), and Spearman Rank Correlation Coefficient (SRCC). The results indicated that a higher number of embeddings enhanced the model's ability to capture complex features in portrait photographs, thereby improving overall performance.Additionally, we explored the impact of subject-centric cropping and incorporated visual interpretability techniques such as Grad-CAM and guided backpropagation. While Grad-CAM provided insights into the decision-making process of our model by highlighting important regions in the images, guided backpropagation was used to subtly enhance the aesthetic quality of images. These methods underscored the strengths and limitations of our model, particularly in terms of transparency and explainability.

Our work underscores the importance of tailored aesthetic models for specific image categories, such as portraits, where unique features and nuances play a critical role in overall aesthetic judgment. By incorporating facial recognition and feature embedding, we have demonstrated a promising direction for improving the accuracy and relevance of neural aesthetic assessments. Future work will focus on further refining our model by experimenting with different embedding methods and expanding our dataset to include a more diverse range of portrait images. Additionally, enhancing the interpretability of our model remains a priority, aiming to provide more consistent and reliable visual explanations that align with human aesthetic preferences.

9. Contributions & Acknowledgements

This project was a collaborative effort between Xiyuan Wu and Cici Hou, with both team members contributing equally to the research and development process.

Xiyuan Wu focused on training the model with facial embeddings. This involved integrating facial recognition embeddings from the VGG-Face model into the NIMA framework, conducting experiments with various embedding sizes, and analyzing the impact on model performance. Additionally, Xiyuan contributed to the overall project design, implementation, and evaluation.

Cici Hou concentrated on the visual interpretability aspects of the project. This included using gradient-weighted class activation mapping (Grad-CAM) and guided backpropagation to visualize the regions of images that contribute most to the model's predictions. Cici also played a significant role in the project's design, implementation, and evaluation.

Both team members collaborated on the remaining parts of the project, including data preprocessing, model training, performance evaluation, and the writing of this paper.

We would like to thank the developers of the PyTorch library for CAM methods [2], the Image Quality Assessment [4] Github repositories, the Deepface library [12] and Python libraries Tensorflow [1], PyTorch [9], SciPy [16], and scikit-learn [10], which were instrumental in our experiments.

Finally, we are grateful for the resources and help provided by Stanford University's CS231N faculties, which played a critical role in inspiring and facilitating our research and experiments.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] J. Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[3] Y. Kao, R. He, and K. Huang. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, PP:1–1, 01 2017.

[4] C. Lennan, H. Nguyen, and D. Tran. Image quality assessment. `https://github.com/idealo/image-quality-assessment`, 2018.

[5] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, Nov. 2015. Publisher Copyright: © 2015 IEEE.

[6] S. Ma, J. Liu, and C. W. Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2017.

[7] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis.

[8] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and

S. Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019.

[12] S. Serengil and A. Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilisim Teknolojileri Dergisi*, 17(2):95–107, 2024.

[13] K. Sheng, W. Dong, X. Mei, F. Huang, and B.-G. Hu. Attention-based multi-patch aggregation for image aesthetic assessment. pages 879–886, 10 2018.

[14] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net, 2015.

[15] H. Talebi and P. Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, Aug. 2018.

[16] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[17] Z. Wang, S. Chang, F. Dolcos, D. Beck, D. Liu, and T. S. Huang. Brain-inspired deep networks for image aesthetics assessment, 2016.

[18] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks, 2013.