Product Review Image Ranking for Fashion E-commerce

Sangeet Jaiswal
sangeet.jaiswal@myntra.com
Myntra Design Pvt Ltd
Bangalore, India

Dhruv Patel
dhruv.patel@myntra.com
Myntra Design Pvt Ltd
Bangalore, India

Sreekanth Vempati
sreekanth.vempati@myntra.com
Myntra Design Pvt Ltd
Bangalore, India

Konduru Saiswaroop
konduru.saiswaroop@myntra.com
Myntra Design Pvt Ltd
Bangalore, India

ABSTRACT

In a fashion e-commerce platform where customers can't physically examine the products on their own, being able to see other customer's text and image reviews of the product is critical while making purchase decisions. Given the high reliance on these reviews, over the years we have observed customers proactively sharing their reviews. With an increase in the coverage of User Generated Content (UGC), there has been a corresponding increase in the number of customer images. It is thus imperative to display the most relevant images on top as it may influence user's online shopping choices and behavior. In this paper, we propose a simple yet effective training procedure for ranking customer images.

We created a dataset consisting of Myntra (A Major Indian Fashion e-commerce company) studio posts and highly engaged (upvotes/downvotes) UGC images as our starting point, and used selected distortion techniques on the images of the above dataset to bring their quality at par with those of bad UGC images. We train our network to rank bad quality images lower than the high-quality ones. Our proposed method outperforms the baseline models on two metrics, namely correlation coefficient and accuracy, by substantial margins.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Computer vision tasks**; **Semi-supervised learning settings**.

KEYWORDS

Image Aesthetics, Image ranking, Deep Learning, Neural Networks, Pre-trained Models

ACM Reference Format:
Sangeet Jaiswal, Dhruv Patel, Sreekanth Vempati, and Konduru Saiswaroop. 2022. Product Review Image Ranking for Fashion E-commerce. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'22).* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Online purchases have seen tremendous growth over the last few years [8]. In 2021, there were approximately 190 million customers annually, compared to 135 million customers in 2019 in India. This increase can be attributed to the growth of the e-commerce industry and to the COVID-19 pandemic, which led to a change in the shopping behavior of the customers. A similar trend is expected in 2022 too.

Accurate, non-misleading visual representation is essential for fashion products sold via e-commerce platforms. Catalog images, at Myntra, are taken under well exposed environments. However, users tend to trust images generated by other users more than generated by the brands[1]. At Myntra, we allow our users to optionally post their photos along with the review they write. However, user generated images sometimes, do not comply with required standards of accuracy of visual representation. For instance, some photos are mirror selfies taken at awkward angles, some are under-exposed, while others are over-exposed, many are cropped representations of the products. In this paper, we train a machine learning model that can differentiate “good” UGC images from “bad” UGC images. With the help of this model, we can sort available UGC images by their quality so that the users do not have to browse too many images to get the feel of the product they are examining.

How can we tell if the image is good or bad? In an ideal world, we can either explicitly ask users to rate the images on some scale, and based on the collective wisdom, sort the available images, or we can implicitly log average time spent by the users on each image, and then sort the images that are viewed for longer duration first. The explicit approach is not feasible as we have to decouple UGC images from the UGC reviews, and ask users to rate both, making the experience inconvenient. The implicit approach requires us to augment current infrastructure to log the average time spend metric.

An alternative approach would be to sort the images based on "likes to total ratings ratio" based on the votes on associated text review. This approach does not work well when we have new reviews or relatively lesser explored products. At Myntra, 50% of the reviews containing at least one photo do not have any votes at all, while 75% of such reviews have less than five votes. One can use bandit-based approaches[27] to solve for this cold start, but those too will require some augmentation to the existing system to capture how users are reacting to the ranking.

SIGIR eCom'22, July 15, 2022, Madrid, Spain

Our model is trained on the image pairs generated in such a way that one image will almost always be superior to another one in quality. We train a multi layer perceptron to score good images higher than bad ones using pairwise hinge loss. To generate such a dataset, we make certain assumptions. One such assumption is that a professionally taken image will be better than a user-generated image. Another assumption is that for highly engaged reviews, users also consider the quality of the associated image while deciding whether that review is helpful or not.

Our key contributions are,

(1) We propose an effective learning scheme by leveraging pre-trained models to extract features for image aesthetic assessment in fashion e-commerce without manual annotation.
(2) To the best of our knowledge, this is the first attempt of ranking fashion UGC images.

The rest of the paper is organized in the following way. Section 2 gives the overview of the related work. Section 3 explains how we generate synthetic ranked dataset and our approach to learn the ranking. In Section 4, we explain the experimental setup. Results are given in Section 5. Section 6 conclude the work.

2 RELATED LITERATURE

The recent trends towards approaching the Image Aesthetics Assessment (IAA) problem have been based on either regression or classification. Most of these models use the AVA[19] or AADB[14] datasets to benchmark their performance.

Yeqing et al.[30] consider the IAA as a binary classification task where they segregate images based on their Mean Opinion Score (MOS). Images with MOS less than 5 will be treated as bad images, and those whose MOS is greater than equal to 5 are considered as good images. They finetune Convolution Neural Network (CNN) models pre-trained on ImageNet such as AlexNet and VggNet to report their accuracies. There are other approaches which deal with the problem of fixed-size image constraints of CNN [11, 17, 18, 29] but eventually, they also solve IAA as binary classification.

Neural Image Assessment (NIMA) [28] introduced a simple strategy. While most of the then existing approaches were based on predicting the MOS, they predicted the aesthetic rating distribution using a CNN that is trained using Earth Mover's Distance Loss (EMD) on human sourced rating distribution from AVA dataset. Despite its simple architecture, it achieves a result which is comparable to State of the art results. We have adopted NIMA for our experiments. We have used MobileNet[12] architecture based CNN as our backbone network to generate image features which are trained on AVA and TID2013[22] datasets.

A task related to IAA is No Reference - Image Quality Assessment (NR-IQA) which assesses the technical quality of an image. Many recent approaches[2, 13, 28] make use of labeled data such as TID2013, LIVE[24] and CSIQ[3] to predict the quality score. Another set of approaches treat this task as a ranking problem and try to minimize the ranking loss using ground truth labels[4, 23]. One of the drawbacks of using deep learning based NR-IQA methods is the need of large labeled dataset which is not available for NR-IQA tasks. Annotation process for IQA image datasets requires multiple human annotations for every image. This process of collecting annotation is very time-consuming and expensive, due to which all the above approaches train shallow networks directly on the dataset. To address this problem, Liu et al. in RankIQA[16] paper have taken large unlabeled high quality images and applied image distortions to generate a ranking image dataset. For example, given an image, upon the addition of gaussian blurs of increasing intensities on it, we end up with a set of images which can be ranked easily as gaussian blur will decrease the image quality. In such datasets we don't have the absolute aesthetic score of an image, but we certainly know for a pair of images which one is of the higher quality. This synthetic data generation allowed them to better train a deep network. Subsequently, they trained a siamese network[6] using efficient ranking loss and further fine tuned the network on labeled dataset to achieve better performance in NR-IQA task.

Our approach is inspired from RankIQA, where they distort the technical aspect of high-quality images by adding gaussian noise, gaussian blur etc. to generate ranking dataset, we are using image manipulation techniques which not just degrade the technical aspect but other aspects of image qualities as well which we generally encounter in our "bad" UGC images. We have also created a pair-sampling strategy which is suited to our use case of ranking images, this strategy narrows the scope of learning and provides a more consistent training to our network.

3 METHODOLOGY

Recent IAA methods rely on training CNN that receive an image as an input and generate a score that is higher for an aesthetically superior image. These networks are generally pretrained on ImageNet dataset and further trained end-to-end on AVA or TID datasets for image aesthetics or technical assessment. However, such a trained network performs suboptimally in ranking domain-specific images directly because of the high diversity in the image content of these datasets.

3.1 Data Collection

In the RankIQA[16] paper, high-quality images were subject to different kinds of image manipulation techniques with different control parameters. Applying such techniques on high-quality images ensures that introducing any type of distortion will certainly degrade the aesthetics. Using this approach as a reference point for creating a synthetic dataset, prior to introducing pertinent distortions, we started off by compiling 19k highly aesthetic images from Myntra Studio and to prevent the network from overfitting to a specific type of image creation style we sampled 16k highly up voted UGC images drawn from customer reviews. After introducing distortions to the compiled images, to the resulting dataset, we added another 3.5k highly down voted UGC images.

Myntra Studio Myntra studio is a platform where fashion influencers post their images/videos wearing products which one can buy from Myntra.

UGC Images In Myntra, a verified buyer can write reviews and can upload images to support their opinion. Any customer can up vote or down vote a review.

A sample representative images from the training set is shown in figure 1.

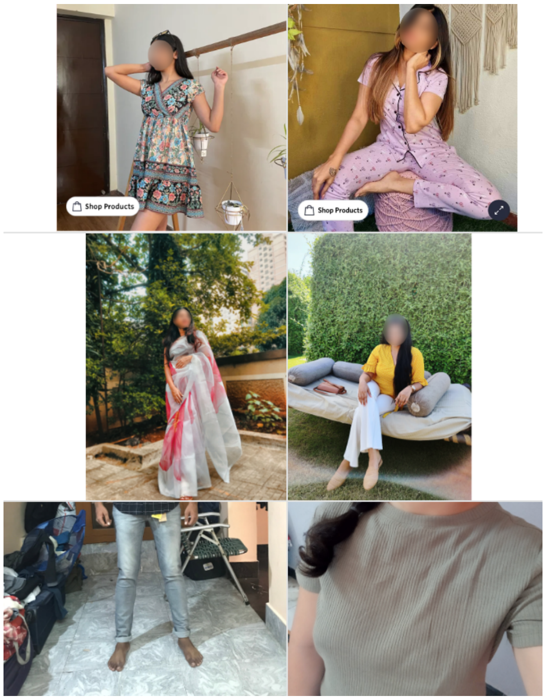

Figure 1: Sample training set images. First row represents Studio images, second and third row represents good UGC and bad UGC images, respectively.

3.2 Image Manipulations Techniques

As described in the papers [16, 25] different image manipulation techniques have diverse effects on the manipulated image. For example, in grayscale image conversion it is difficult to compare the input with output in terms of aesthetics. But in our case we want to rank such images lower than the colored counterpart because it will be hard for the customer to make sense of the color of the product in grayscale image. Likewise, we have identified certain image manipulation techniques that are guaranteed to render a degraded effect on the image quality, as listed in Table 1.

We have adopted a variety of techniques to generate synthetic training instances which emulate the low-quality images that we get in our product reviews, including (1) Vertical and horizontal crop - Location of subject and object in an image do play an important role in defining the aesthetics of an image. Partial subject in an image do affect the aesthetics of an image, for instance, shirt buyers excludes the portion below the abdominal area while uploading pictures of them wearing the shirts they purchased. To mimic the same, good quality images were subjected to cropping (Vertical and Horizontal). (2) Addition of color jittering by changing the Brightness, contrast and hue based on a scaling factor to mimic the poor lighting conditions. (3) Gaussian blur and Gaussian noise to add fuzziness and graininess in an image. (4) Grayscale to mimic one simple basic filter which customers apply. Customers sometimes do use advance filters, but we are not taking them into account for now. (5) Random Rotation and Rotation with Mixup is applied to achieve camera shake effect.

3.3 Ranked Images Generation and Sampling Strategy

In addition to what has already described in section 3.1, It would be worthwhile to mention that UGC images were further divided into two classes, one in which the customer is actually wearing a product and the other being flat shot images. We achieved this using YOLOv5[10] model to classify images as with/without human. We segregated them because in case of apparels, images with usability have precedence and relevance when it comes to standalone images of product purchased. With all the images, we make the following pairs -

$$\{(x, y) : \exists x \in D_{+}, y \in D_{-}\}$$

where D_{+} and D_{-} comes from the Table 2.

We sample pairs as defined in the table uniformly. For pair 1, we considered studio images as positive samples and used image distortion techniques as described in section 3.2 in random order on positive samples to generate corresponding negative samples. An analogous description holds for Pair 2. In pair 3,4 we considered studio images as positive samples and UGC images as negative samples. In pair 5, we considered "good" UGC images of users wearing the product as a positive sample and "bad" UGC images of users wearing the product as negative samples. In pair 6 we ranked "good" flat shot(non human) images higher than "bad" flat shot images.

3.4 Neural Network Architecture

In our experiments, we have used NIMA[28] as our reference to extract image features. As described in the NIMA paper, they have trained deep CNNs on two datasets. One network tries to capture the style, content, composition etc. and another network tries to capture the technical quality of an image. We aggregate the features generated from the penultimate layer, i.e. the average pooling layer, which is 1024 dimension in our case. We have also aggregated the probability distribution of predicted rating as a feature, as described in the Figure 3.

NIMA - Aesthetics This CNN model is based on MobileNet architecture whose weights are initialized by training on ImageNet dataset and then end-to-end training is performed on AVA[1] dataset. The AVA dataset contains 2,55,000 images, rated based on image aesthetics such as style, content, composition etc. by photographers. Each image is roughly rated by 200 people in response to a photography context on a scale of 1-10. This model tries to predict the normalized distribution of rating for an image.

NIMA - Technical This CNN model is based on MobileNet architecture. It is trained on the TID2013[22] dataset. This contains 3000 images which are generated from 25 reference images, 24 types of distortion with 5 levels of each distortion. Ratings are collected by showing a pair of distorted images for each reference image, and the observer has to select the better one in the pair. Unlike AVA dataset, TID2013 provides just the mean opinion score and standard deviation. NIMA paper requires training on score probability distribution.

[1] AVA images are obtained from www.dpchallenge.com, which is an on-line community for amateur photographers.

Table 1: Image Manipulation Techniques used in our approach

Operations	Parameters	Rationale
Random Crop, Vertical Crop, Horizontal Crop	[0.4,0.6]	Partial Subject
Color Jitter - Brightness, Contrast and Hue	$[0.3,0.6] \cup [1.2,1.4]$	Poor Lighting
Gaussian Blur	[0.8,1.2]	Soft Images
Gaussian Noise	[0.2,0.8]	Grainy Images
Grayscale	-	Image Filter
Random Rotation, Random Rotation + Mixup	{5,10,15,20} alpha - [0.2,0.4]	Camera Shake

Table 2: Image Pair Sampling Strategy

S. No	D_+	D_-
1	D_{studio}	$D_{studio\_distorted}$
2	D_{ugc_good}	$D_{ugc\_good\_distorted}$
3	D_{studio}	D_{ugc_good}
4	D_{studio}	D_{ugc_bad}
5	$D_{ugc\_good\_human}$	$D_{ugc\_bad\_human}$
6	$D_{ugc\_good\_non\_human}$	$D_{ugc\_bad\_non\_human}$

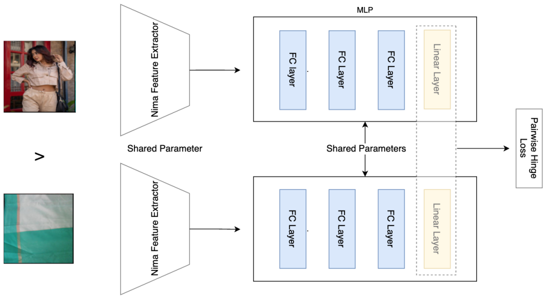

Figure 2: Network Architecture

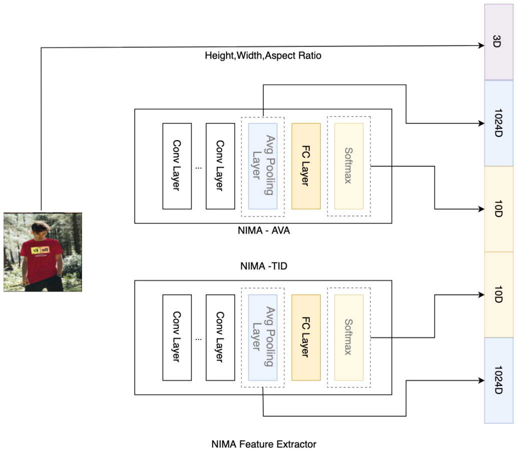

Figure 3: NIMA Aesthetic and Technical Model

The score distribution is approximated through maximum entropy optimization[7].

We have also taken the height, the width, and the aspect ratio of an image as features. The reason to incorporate them as features is that in NIMA we have to rescale all the images to a fixed size regardless of their original image aspect ratios. The lack of information about the original image size during the training of CNN in NIMA can affect its prediction, as the human rater may not give the same rating to the resized version of the image.

Given a pair of images I_1 and I_2 as an input to the NIMA feature extractor, the output feature representation is denoted by x_1 and x_2 respectively. Now these features will be given as an input to our Siamese network shown in Figure 2. The output is represented by $f(x;\Theta)$ which is obtained by capturing the output of the last layer. Here Θ are the network parameters, here we will use y to denote the ground truth value for the image. The network output of the final layer is a single scalar. The network is supposed to output higher score for high-quality images and smaller score for low-quality images. For a pair of images, ranking loss is defined as

$$L_{rank} = \sum_{i,j} max(0, m - \delta(y_i \geq y_j)(f(x_i;\Theta) - f(x_j;\Theta)))$$

where

$$\delta(y_i \geq y_j) = \begin{cases} 1, & \text{if } y_i \geq y_j \\ -1, & \text{if } y_i < y_j. \end{cases}$$

where m is the minimal margin denoting the desired difference between the scores generated by the ranking network for a pair of images.

4 EXPERIMENTS

4.1 Implementation Details

We train our Siamese network with the image pairs that we generated as described in the section 3.3. We have set the hyperparameter m, which describes the minimal margin between the positive and negative image pair, to 1.

We have collected 19K images from Myntra studio, 16K highly up voted UGC images and 3.5K highly down voted UGC images for generating training image pairs.

For validation, we have kept 1000 images each from the studio, "good" UGC images and "bad" UGC images for 1000 different styles[2]. Therefore, we have 3 images for each style. We have used accuracy (which is defined in section 5) as the metric on validation set to select the best model.

We have used pretrained NIMA feature extractor[15] which is implemented in Keras[5] and we have converted them into ONNX[20], as we do not train these networks. We run this ONNX models using ONNX Runtime[9].

Our MLP network contain 3 hidden layers and an output layer. The first hidden layer transforms the output from NIMA feature extractor to 512 dimensions. Subsequent layers transform it to 256 and 128 dimensions, respectively. The network output of the final layer is a scalar value representing the score.

During training, input images are rescaled to 224 x 224. We train the network using ADAM optimizer. We have used default learning rate(10^{-3}) for fully connected layers and weight decay regularization of $5 * 10^{-4}$. We have also used learning rate scheduler, which will halve the learning rate if validation accuracy doesn't improve in five consecutive epochs. We have experimented with 16 as a batch size. All our implementation is done in PyTorch[21].

5 RESULTS

To compare our model, we use NIMA models as our baselines. Both NIMA-Aesthetics and NIMA-Technical predict the probability distribution of the score in range 1-10, inclusive. We take the expected value of the NIMA-X's output on an image as the score of that image by model x. That is,

$$f^X(I) = \sum_{i=1}^{i=10} iPr_X(i;I)$$

where X is either A (for Aesthetics) or T (for Technical), and $Pr_X(i;I)$ denotes the probability for score i by NIMA-X. Images with higher predicted scores are ranked higher.

Since we do not have the ground truth rankings, we take users' ratings as a proxy to the quality of an image. That is, if a particular image I has u up votes and d down votes, we assume that the ground truth quality score for that image is $\frac{u}{u+d}$. To create such a test set, we gathered around 850 images associated with highly engaged review from 20 popular styles. None of the images from these 20 styles (highly engaged or not) were kept in our training set or validation set. As mentioned in the introduction, the ratings are not given to the images, but the reviews. However, since these reviews are highly engaged, we assume that raters would have considered accompanying image while rating the review. To validate this hypothesis, we analyzed our reviews for their engagement, and found that reviews with associated images had on average 6.5x engagement in terms of upvotes/downvotes as compared to reviews without images. We also did a paired t-test and found that this difference was statistically significant. A subset of the images with their computed scores is presented in Figure 4.

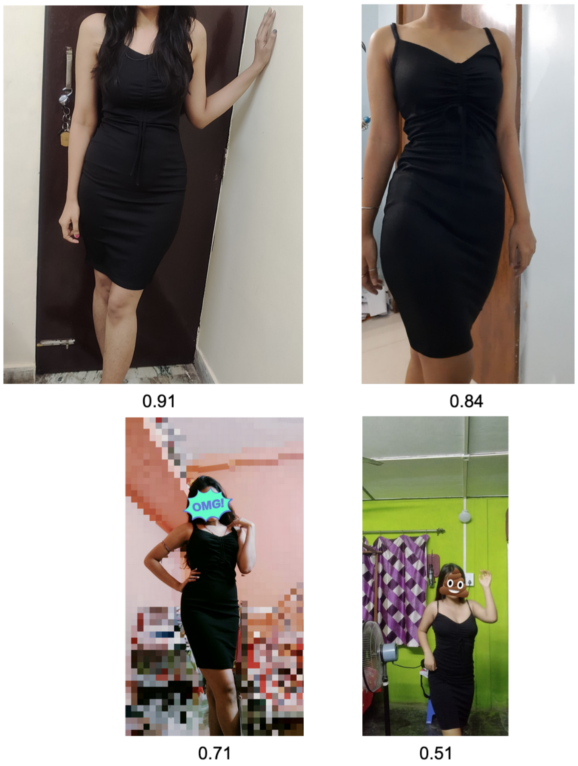

Figure 4: Sample test set images for a particular style. For each image, the number in the bottom center is the ground truth score for that image.

We use two metrics to quantify our results. The first one is Pearson correlation coefficient, which is a common metric to compare the performance of image ranking models on labeled datasets. It is computed as,

$$\rho(f,S) = \frac{\sum_{I\in S}(f(I)-\bar{f}(S))(g(I)-\bar{g}(S))}{\sqrt{\sum_{I\in S}(f(I)-\bar{f}(S))^2(g(I)-\bar{g}(S))^2}}$$

where f is either our baseline or our model, g is the ground truth score function, S is the set of images belonging to a particular style(i.e. all images in S belong to the same product), $\bar{h}(S)$ is the mean of the scores of the images in S computed by h. 0 correlation implies that the model gives scores that are unrelated to the ground truth scores (i.e. likes), positive correlation implies that the model scores highly liked images higher than highly disliked images.

Another metric we use is accuracy. To compute accuracy for a particular style, we randomly (without replacement) pick 50 pairs of images for that style. We compute scores for these pairs using our model or the baselines. If we pick a pair (I_1, I_2), and $g(I_1) > g(I_2)$, then models should generate $f(I_1) > f(I_2)$, otherwise that pair is considered as misclassified.

Table 3 compares our approach with our baselines. We report the average of the metrics for all 20 styles. As can be seen in the first two rows of the table, NIMA models without finetuning, even though trained for aesthetics, outputs results that are completely uncorrelated with the proxy ground truth. Note that we did not

[2] At Myntra we use the term style to mean a specific product.

Table 3: Quantitative results for our approach.

model	correlation coefficient	accuracy
NIMA-Aesthetics	-0.05	0.48
NIMA-Technical	-0.02	0.50
Ours	**0.19**	**0.58**

train our model to predict the proxy scores. We are using a pairwise loss function to differentiate between positive and negative image, still our model has positive correlation with the proxy ground truth. The accuracies of the NIMA models are no good than the accuracy one would get by guessing the binary prediction randomly.

6 CONCLUSION AND FUTURE WORK

This paper presents an effective scheme of leveraging existing Deep learning models for Image Aesthetic Assessment as feature extractor to fine tune a Siamese network trained on synthetic data generated to rank UGC images. We created the dataset by systematically degrading the quality of studio and "good" UGC images. This was done to emulate the kind of low-quality imagery that we encounter on a routine basis. We have seen that this approach helps in improving the ranking as compared to baseline NIMA-Aesthetics and NIMA-Technical. Our technique is not limited to NIMA feature extractor, as we can replace NIMA feature extractor with any other feature extractor (e.g. [17, 26]) trained on image aesthetics assessment datasets [22, 28]. Extending our existing approach to pretrain a CNN and fine tune on our labeled ranked dataset for fashion will be an interesting experiment to perform. We can leverage this model for Thumbnail generation, ranking catalog & studio images. This model can be leveraged for providing the customer feedback through a prompt about the quality of the image they have taken before they submit the images.

REFERENCES

[1] bkmediagroup. 2022. ugc.

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* 27, 1 (2017), 206–219.

[3] Damon M. Chandler. 2013. CSIQ dataset. https://www.hindawi.com/journals/isrn/2013/905685/#abstract.

[4] Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhi-Ming Ma, and Hang Li. 2009. Ranking measures and loss functions in learning to rank. *Advances in Neural Information Processing Systems* 22 (2009).

[5] François Chollet et al. 2015. Keras. https://keras.io.

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 539–546.

[7] Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

[8] Statista Research Department. 2022. Online shoppers in India.

[9] ONNX Runtime developers. 2022. ONNX Runtime. https://onnxruntime.ai/.

[10] YOLOv5 developers. 2022. YOLOv5. https://pytorch.org/hub/ultralytics_yolov5/.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 37, 9 (2015), 1904–1916.

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[13] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1733–1740.

[14] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *ECCV*.

[15] Christopher Lennan, Hao Nguyen, and Dat Tran. 2018. Image Quality Assessment. https://github.com/idealo/image-quality-assessment.

[16] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*. 1040–1049.

[17] Shuang Ma, Jing Liu, and Chang Wen Chen. 2017. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4535–4544.

[18] Long Mai, Hailin Jin, and Feng Liu. 2016. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 497–506.

[19] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2408–2415.

[20] ONNX. 2022. Open Neural Network Exchange. https://onnx.ai.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[22] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication* 30 (2015), 57–77.

[23] D Sculley. 2009. Large scale learning to rank. (2009).

[24] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing* 15, 11 (2006), 3440–3451.

[25] Kekai Sheng, Weiming Dong, Menglei Chai, Guohui Wang, Peng Zhou, Feiyue Huang, Bao-Gang Hu, Rongrong Ji, and Chongyang Ma. 2020. Revisiting image aesthetic assessment via self-supervised feature learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5709–5716.

[26] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. 2018. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia*. 879–886.

[27] Aleksandrs Slivkins. 2019. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272* (2019).

[28] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing* 27, 8 (2018), 3998–4011.

[29] Lijie Wang, Xueting Wang, and Toshihiko Yamasaki. 2020. Image aesthetics prediction using multiple patches preserving the original aspect ratio of contents. *arXiv preprint arXiv:2007.02268* (2020).

[30] Yeqing Wang, Yi Li, and Fatih Porikli. 2016. Finetuning convolutional neural networks for visual aesthetics. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 3554–3559.