

Considering user agreement in learning to predict the aesthetic quality

Suiyi Ling, Andreas Pastor, Junle Wang, Patrick Le Callet

▶ **To cite this version:**

Suiyi Ling, Andreas Pastor, Junle Wang, Patrick Le Callet. Considering user agreement in learning to predict the aesthetic quality. ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing, May 2022, Singapour, Singapore. 10.1109/ICASSP43922.2022.9746258 . hal-03632996

HAL Id: hal-03632996
https://hal.science/hal-03632996

Submitted on 13 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CONSIDERING USER AGREEMENT IN LEARNING TO PREDICT THE AESTHETIC QUALITY

Suiyi Ling[1], *Andréas Pastor*[1], *Junle Wang*[2], *Patrick Le Callet*[1]

[1] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, Nantes, France [2]Turing Lab, Tencent

ABSTRACT

How to robustly rank the aesthetic quality of given images has been a long-standing ill-posed topic. Such challenge stems mainly from the diverse subjective opinions of different observers about the varied types of content. There is a growing interest in estimating the user agreement by considering the standard deviation (σ) of the scores, instead of only predicting the mean aesthetic opinion score (μ). Nevertheless, when comparing a pair of contents, few studies consider how confident are we regarding the difference in the aesthetic scores. In this paper, we thus propose (1) a re-adapted multi-task attention network to predict both the mean opinion score and the standard deviation in an end-to-end manner; (2) a brand-new confidence interval ranking loss that encourages the model to focus on image-pairs that are less certain about the difference of their aesthetic scores. With such loss, the model is encouraged to learn the uncertainty of the content that is relevant to the diversity of observers' opinions, *i.e.*, user disagreement. Extensive experiments have demonstrated that the proposed multi-task aesthetic model achieves state-of-the-art performance on two different types of aesthetic datasets, *i.e.*, AVA and TMGA.

Index Terms— Aesthetic assessment, multi-task learning, observers' diversity of opinions, confidence interval loss

1. INTRODUCTION

Visual aesthetic assessment is important for many real-world use cases, including automatic image composition [1], image creation, editing [2], graphical design *etc.* and is intriguing more other piratical research topics. As people differ in how they respond to artworks [3], unlike general quality assessment, aesthetic assessment is associated more with high-level components of the contents in terms of emotions, composition, and beauty. Thus, it is more subjective compared to quality assessment of compressed/distorted contents [4]. State-of-The-Art (SoTA) approaches [5, 2, 6] concentrate on leveraging different deep neural networks with large-scale annotated aesthetic data, and achieve decent performance on benchmark datasets in recent year. Other than predicting the atheistic quality score in terms of mean opinion scores (μ) alone, more and more studies, *e.g.* NIMA [4], developed their models by taking the standard deviation (σ) of the scores, *i.e.*, how observer agree with each other, into account.

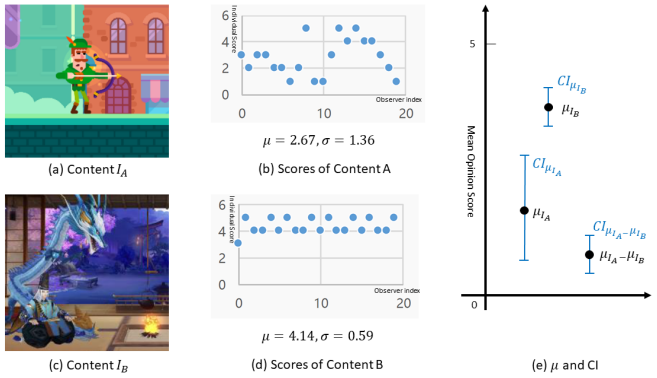


Fig. 1. Example of contents that have significantly different standard deviation σ, and demonstration of how Confidence Intervals (CI) could help to rank the aesthetic scores.

When ranking the aesthetic qualities of a pair of contents, it is important to not only compare their mean opinions scores, but also the corresponding confidence intervals. For instance, as shown in Fig 1, as the art-style of image I_A is simpler than I_B, its aesthetic votes are polarized. As depicted in Fig 1 (b), some of the observers voted very high score for I_A as they preferred the style, while most of the other observers gave very low scores. As a result, the standard deviation of its aesthetic score σ_{I_A} is much higher than the one of I_B. When comparing the aesthetic quality between I_A and I_B, we want to make a statistically solid decision based on the opinions of the majority. In other words, as presented in Fig 1 (e), we consider I_B is significantly better than I_A only if the confidence intervals of the two contents are not overlapped [7, 8]. In general, the Confidence Interval (CI) of the aesthetic score of a given image I_A is given by:

$$CI_{\mu_{I_A}} = \mu_{I_A} \pm z\frac{\sigma_{I_A}}{\sqrt{n_{obs}}} \tag{1}$$

where $z = 1.96$ is the critical value *w.r.t.* 95% of confidence level, and n is the number of observers. However, within modern deep learning framework, it is impractical to compute first the CIs and then compare them (*e.g.*, limited by the design of loss function). Therefore, as an alternative, we may prefer to know how confident we are regarding the difference of the aesthetic scores of a given pair. If $CI_{\mu_{I_A}-\mu_{I_B}}$ is small enough, we could then say confidently that they are significantly different, where $CI_{\mu_{I_A}-\mu_{I_B}}$ is the confidence interval

of $\mu_{I_A} - \mu_{I_B}$, defined as:

$$CI_{\mu_{I_A}-\mu_{I_B}} = |\mu_{I_A} - \mu_{I_B}| \pm z \cdot \sigma_{\mu_{I_A}-\mu_{I_B}} \quad (2)$$

It is not hard to see that the standard deviations of the observers' scores σ are in both aforementioned equations, and thus are important to be predicted along with the mean opinion score μ. Based on the discussion above, the contribution of this study is two-fold: (1) we propose to predict both μ and σ together via a novel multi-task attention network; (2) inspired by equation (2), we develop a new confidence interval loss so that the proposed model is capable of learning to rank the aesthetic quality regarding the confidence intervals.

2. RELATED WORK

In the past decade, the performances of aesthetic assessment models grow at a respectable pace. Li *et al.* [9] proposed one of the early efficient aesthetic metrics based on hand-crafted features. By formulating aesthetic quality assessment as a ranking problem, [10], Kao *et al.* developed a rank-based methodology for aesthetic assessment. Akin to [10], another ranking network was proposed in [11] with attributes and content adaptation. To facilitate heterogeneous input, a double-column deep network architecture was presented in [12], which was improved subsequently in [13] with a novel multiple columns architecture. Ma *et al.* developed a salient patch selection approach [14] that achieved significant improvements. Three individual convolutional neural networks (CNN) that capture different types of information were trained and integrated into one final aesthetic identifier in [15]. Global average pooled activations were utilized by Hii *et al.* in [16] to take the image distortions into account. Later, triplet loss was employed in a deep framework in [17] to further push the performances to the limits of most modern methods available at the time. The Neural IMage Assessment (NIMA) [4], developed by Talebi *et al.*, is commonly considered as the baseline model. It was the very first metric that evaluates the aesthetic score via predicting the distribution of the ground truth data. To assess UAV video aesthetically, a deep multi-modality model was proposed [18]. As global pooling is conducive to arbitrary high-resolution input, MLSP [6] was proposed, based on Multi-Level Spatially Pooled features (MLSP). Recently, an Adaptive Fractional Dilated Convolution (AFDC) [19] was proposed to incorporate the information of image aspect ratios.

3. THE PROPOSED MODEL

3.1. Deep Image Features for Arbitrary Input

It is verified in [6] that the wide MLSP feature architecture *Pool-3FC* achieves the SoTA performance. Thus, it is adopted in this study as the baseline model. Similar to MLSP, we extracted features from the Inception ResNet-v2 architecture. As the dimension of the features, *i.e.*, $5 \times 5 \times 16928$ is too high, the features are first divided into 16 sub-features with dimension of $5 \times 5 \times 1058$ (*i.e.*, the dark-grey blocks at left-upper corner of Fig. 2) to ease the computation of attention feature and reduce the number of parameters used across the architecture. After the division, each sub-feature is fed into a Lighter-weight Multi-Level Spatially Pooled feature block (LMLSP) as depicted in Fig. 2. The module of LMLSP is shown in Fig. 3. It is composed of three streams, including the 1×1, 3×3 convolutions streams, and the stream of average pooling followed by a 1×1 convolution. Through this design, the dimension of the multi-level spatial pooled feature was reduced to facilitate the latter fully connected layers.

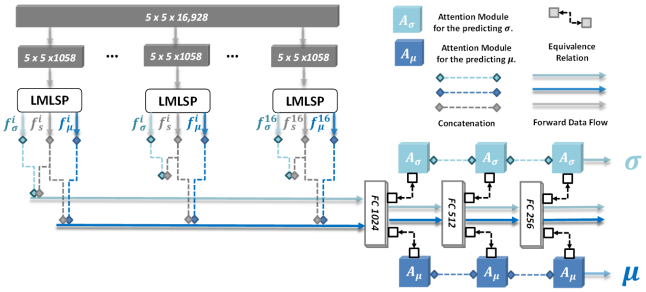


Fig. 2. Overall network architecture of the proposed model.

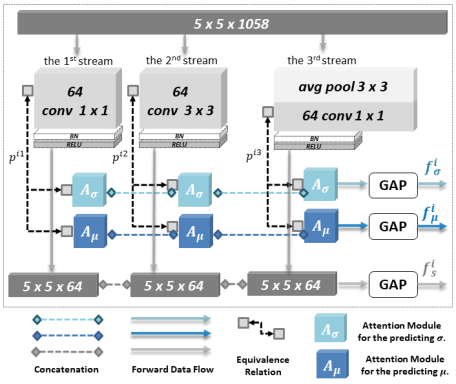


Fig. 3. Detailed visualization of each LMLSP block.

3.2. Predicting μ and σ at the Same Time with Multi-Task Attention Network

The goal of using multi-task learning is two-fold, *i.e.*, (1) predict both the mean opinion scores and the standard deviation at the same time; (2) via joint loss functions, the network learns to consider observers' diverse opinions (σ), the content uncertainty, *etc.*, when predicting the final aesthetic score.

In this study, the multi-task attention network proposed in [20] is adapted to predict both the mean opinion score μ, and the standard deviation of the subjects' scores σ. The proposed LMLSP blocks are utilized as task-shared features blocks. And the separate attention modules, which are linked to the shared model, are designed to learn task-specific features for μ and σ correspondingly. The attention modules of μ (in cyan color) and σ (in dark blue color) are shown in Fig. 2

and 3. As presented in Fig. 3, within each individual LMLSP block i, two attention masks are attached to each convolution stream, *i.e.*, one for task μ and one for task σ at each conv 1×1, conv 3×3, and the average pooling stream. The task specific features $\hat{a}_{\sigma}^{ij}$ and $\hat{a}_{\mu}^{ij}$ in the i_{th} LMLSP block, at the j_{th} stream are calculated via element-wise multiplication of the corresponding attention modules A_{σ}/A_{μ} with the shared stream features p^{ij}:

$$\hat{a}_{\sigma}^{ij} = A_{\sigma} \odot p^{ij}, i \in [1, 16], j \in [1, 3] \tag{3}$$

$$\hat{a}_{\mu}^{ij} = A_{\mu} \odot p^{ij}, i \in [1, 16], j \in [1, 3] \tag{4}$$

where $\odot$ indicates the element-wise multiplication operation.

Then, for each task μ/σ, the task-specific features obtained from each stream are concatenated into one dedicated feature, followed by a Global Average Pooling (GAP). Same for the shared features from the three streams. As such, for the i_{th} LMLSP block, three features are output, including f_{μ}^{i}, f_{σ}^{i} and f_{s}^{i}, which are the feature of task μ, the feature of task σ and the feature of the shared network respectively:

$$f_{\mu}^{i} = GAP([\hat{a}_{\mu}^{i1}, \hat{a}_{\mu}^{i2}, \hat{a}_{\mu}^{i3}]), i \in [1, 16] \tag{5}$$

$$f_{\sigma}^{i} = GAP([\hat{a}_{\sigma}^{i1}, \hat{a}_{\sigma}^{i2}, \hat{a}_{\sigma}^{i3}]), i \in [1, 16] \tag{6}$$

$$f_{s}^{i} = GAP([f_{\sigma}^{i1}, f_{\sigma}^{i2}, f_{\sigma}^{i3}]), i \in [1, 16], \tag{7}$$

where $GAP(\cdot)$ is the Global Average Pooling and $[X, Y]$ denotes the concatenation of tensor X and Y.

Afterward, as depicted in the lower part of Fig. 2, for each task, *i.e.*, μ, or σ, at each LMLSP block, the task-specific feature $f_{\mu/\sigma}^{i}$ is first concatenated with the shared feature f_{s}^{i}, and all the obtained features are further concatenated across all the 16 LMLSP blocks to generate the final feature for each task. The feature flow is highlighted in cyan color for task σ and dark blue color for task μ. Lastly, the features are forwarded to three continuous Fully-Connected (FC) layers with the same attention modules for each task after each FC layer to predict the final μ and σ. It is worth mentioning that the entire network is trained in an end-to-end manner. As such, the attention modules for μ, and σ serves as feature selectors that pick up the relevant dimensions in predicting μ or σ respectively, while the shared LMLSP blocks and the shared FC layers learn the general features across the two tasks.

3.3. Loss function

Under the multi-task learning setting, the joint loss function L_{MTL} of the tasks of predicting the mean μ and the standard deviation of subject' scores is defined as:

$$\mathcal{L}_{\mathcal{MTL}} = \alpha_{\mu}\mathcal{L}_{\mu} + \alpha_{\sigma}\mathcal{L}_{\sigma}, \tag{8}$$

where α_{μ} and α_{σ} are the parameters that balances the losses of predicting μ, *i.e.*, $\mathcal{L}_{\mu}$ and the one of predicting σ *i.e.*, $\mathcal{L}_{\sigma}$. The $\mathcal{L}_{\sigma}$ is simply defined as the Mean Absolute Error (MAE) between the ground truth σ_I of an image I, and the corresponding predicted $\hat{\sigma}_I$. N equals to the number of images:

$$L_{\sigma} = \frac{\sum_{I=1}^{N} |\sigma_I - \hat{\sigma}_I|}{N} \tag{9}$$

As emphasized in Section 1, when comparing the aesthetic scores of two images I_A and I_B, we want to know not only whether image I_A is significantly better than I_B in terms of their aesthetic quality, but also how confident we are regarding the difference. Specifically, we want to consider this 'certainty' when predicting μ. Thus, the L_{μ} is defined with a novel $\mathcal{L}_{CI}$, namely, the *Confidence Interval Loss*:

$$\mathcal{L}_{\mu} = (1 - \lambda)\frac{\sum_{I=1}^{N} |\mu_I - \hat{\mu}_I|}{N} + \lambda \cdot \mathcal{L}_{CI}, \tag{10}$$

where λ is a parameter that balances the two losses. Inspired by the confidence interval of the differences between two aesthetic scores $CI_{\mu_{I_A} - \mu_{I_B}}$, which is defined as $|\mu_{I_A} - \mu_{I_B}| \pm z \cdot \sigma_{\mu_{I_A} - \mu_{I_B}}$ (shown in Section 1), we further define $\mathcal{L}_{CI}$ as:

$$\begin{aligned}\mathcal{L}_{CI} = \sum_{(I_A, I_B) \subset S_{pair}} max\big(0, \\ l_{CI}(\mu_{I_A}, \mu_{I_B}) \cdot [\,|\,|\mu_{I_A} - \mu_{I_B}| - |\hat{\mu_{I_A}} - \hat{\mu_{I_B}}|\,|]\big),\end{aligned} \tag{11}$$

where μ_I and $\hat{\mu_I}$ are the ground truth mean and the predicted mean respectively, S_{pair} is the set of all possible pairs, and l_{CI} is defined as below:

$$l_{CI} = \begin{cases} 1, & z \cdot \sigma_{\mu_i - \mu_j} > \tau \\ 0, & otherwise \end{cases} \tag{12}$$

where τ is the margin. Without loss of generality, $\sigma_{y_i - y_j} \approx \sqrt{\frac{\sigma_i^2}{n_{obs}} + \frac{\sigma_j^2}{n_{obs}}}$ [21]. n_{obs} indicates the number of observers. l_{CI} serves as a gating function. In equation (11), the prediction error of a given image pair is accumulated only if l_{CI} equals to 1, under the condition of $z \cdot \sigma_{\mu_i - \mu_j} > \tau$. It can be noticed that, $z \cdot \sigma_{\mu_i - \mu_j}$ is from the (2), *i.e.*, the definition of $CI_{\mu_{I_A} - \mu_{I_B}}$. The proposed *Confidence Interval Loss* punishes only the pair that have $CI_{\mu_{I_A} - \mu_{I_B}}$ value higher than the margin τ. In other words, it focuses on pairs that have higher uncertainty (larger confidence interval of the aesthetic score difference). Furthermore, the loss for pair that triggers the gate function is $[|\mu_{I_A} - \mu_{I_B}| - |\hat{\mu_{I_A}} - \hat{\mu_{I_B}}|]$, which encourages the prediction of the difference between the predicted scores to be as close as the difference of the ground truths.

The intuition behind this selection of pairs is to concentrate on pairs that have larger uncertainties, which further enables the model to learn uncertainty from the ambiguous contents and improve the performance of the model on such pairs.

4. EXPERIMENT

4.1. Experimental setup

The performance of the proposed model was assessed on two different types of aesthetic quality datasets that were developed for different use cases/scenarios:

(1) The Tencent Mobile Gaming Aesthetic (TMGA) dataset [22], which is developed for improving gaming aesthetics. In this dataset, there are in total 1091 images collected from 100 mobile games, where each image was labeled with

four different dimensions including the 'Fineness', the 'Colorfulness', the 'Color harmony', and the 'Overall aesthetic quality'. The entire dataset is divided into 80%, 10%, and 10%, for training, validation, and testing correspondingly.

(2) The Aesthetic Visual Analysis (AVA) dataset [23], which was developed for general aesthetic purposes. It contains about 250 thousand images that were collected from a photography community. Each individual image was rated by an average of 210 observers. The same random split strategy was adapted as done in [19, 6] to obtain 235,528 images for training/validation (95%/5%) and 20,000 for test.

Since most of the metrics were evaluated on the AVA dataset regarding the binary classification performance, ACCuracy (ACC) was computed based on classifying images into two classes with a cut-off score equals to 5 [4, 6] when reporting the performances on the AVA dataset. However, as emphasized in [6], this predominant two-class accuracy suffers from several pitfalls. For instance, due to the unbalanced distribution of images in training, testing set (different numbers of images of different aesthetic quality score values), using 'accuracy' does not fully reveal/stress out the performances of under-test metrics regarding its capability in ranking the aesthetic score of the image. Thus, as done in [6, 22], we computed the Pearson correlation coefficient (PCC), and the Spearman's rank order correlation coefficient (SCC) between the ground truth and the predicted μ to evaluate the performance of different aesthetic quality assessment approaches. Similar to [4], as the range of σ differs across the dataset, the performances of models in terms of predicting σ are reported in PCC, and SCC. The proposed model is mainly compared to SoTA aesthetic models AFDC [19], MLSP [6] and one of the most popular baseline models NIMA [4].

During training, we applied ADAM optimizer with an initial learning rate of 10^{-4}, which was further divided by 10 every 20 epochs. As a preliminary study to learn both σ and μ with multi-task attention network by taking confidence interval into account, we still want to focus on predicting μ. Thus, the λ_μ and λ_σ in (8) were set as 0.6 and 0.4 accordingly, with slightly higher weight for task μ. λ in (10) was set as 0.5.

4.2. Experimental Results

Results on AVA dataset: The PCC, SCC, and ACC values of the considered models in predicting μ and σ on the AVA dataset are shown in Table 1. Regarding task μ, the proposed model accomplishes state-of-the-art performance by obtaining the best PCC, SCC values, and second best ACC value. Regarding task σ [1], the proposed model achieves the best performance. It can be noticed that the performances of objective models in predicting σ are lower than the ones of μ. One of the potential reasons is that the task of predicting σ is more challenging than μ, as it is not only relevant to the contents but also the subjects who participated in the test.

[1] As the code of 'AFDC' is not publicly available, its performance in predicting σ on AVA, and its performances on TMGA are not reported.

Results on TMGA dataset: For a fair comparison, NIMA and MLSP were first finetuned on the training set of TMGA dataset with the optimized hyper-parameters. As the performance of MLSP is higher than NIMA by a large margin, we only report the results of MLSP. The results are presented in Table 2. Similarly, the performances of our model are superior to the ones of the compared models.

Table 1. Performances of relevant models **AVA** dataset.

	PCC (μ)	SCC (μ)	ACC	PCC (σ)	SCC(σ)
NIMA [4]	0.636	0.612	81.51%	0.233	0.218
MLSP [6]	0.756	0.757	81.72%	0.568	0.549
AFDC [19]	0.671	0.648	**83.24%**	-	-
Proposed	**0.785**	**0.779**	82.97%	0.624	0.613

Table 2. Performances of relevant models on **TMGA** dataset.

	Fineness	Colorful	Harmony	Overall
PCC of models in predicting μ				
MLSP [6]	0.904	0.900	0.888	0.872
Proposed	**0.923**	**0.936**	**0.918**	**0.910**
SCC of models in predicting μ				
MLSP [6]	0.904	0.904	0.826	0.8652
Proposed	**0.917**	**0.923**	**0.889**	**0.897**
PCC of models in predicting σ				
MLSP [6]	0.607	0.413	0.289	0.5151
Proposed	**0.686**	**0.557**	**0.372**	**0.559**
SCC of models in predicting σ				
MLSP [6]	0.555	0.364	0.301	0.475
Proposed	**0.679**	**0.486**	**0.370**	**0.478**

4.3. Ablation Study

To validate the effeteness of each component within the proposed model, ablation studies (with 2 ablative models) were conducted. The results are presented in Table 3: (1) when comparing rows 1 and 2, it is demonstrated that by employing the proposed $\mathcal{L}_{CI}$ the performance of the baseline single-task MLSP model in predicting μ is improved; (2) By comparing row 1 and 3, it is showcased that the performance of the aesthetic model is boosted by learning both σ and μ together via multi-task attention network.

Table 3. Ablation studies on the **AVA** dataset.

	PCC	SCC	ACC
Single-Task μ (MLSP) [6]	0.756	0.757	81.72%
Single-Task μ with $\mathcal{L}_{CI}$	0.762	0.765	81.89%
Multi-Task μ & σ without $\mathcal{L}_{CI}$	0.775	0.772	82.24%
Multi-Task μ & σ with $\mathcal{L}_{CI}$ **(Proposed)**	**0.785**	**0.779**	**82.97%**

4.4. Conclusion

In this study, a novel aesthetic model is proposed to learn the aesthetic ranking by exploiting the confidence intervals based on (1) learning both the mean and the standard deviation of aesthetic scores through multi-task attention network; (2) a new confidence interval loss that facilitates the model to concentrate on the less confident pairs and learn the high-level ambiguous characteristics of the contents. Through experiments, it is showcased that our model achieves superior performances compared to the state-of-the-art models in predicting both the mean and the standard deviation of aesthetic scores.

5. REFERENCES

[1] Yu-Chuan Su, Raviteja Vemulapalli, Ben Weiss, Chun-Te Chu, Philip Andrew Mansfield, Lior Shapira, and Colvin Pitts, "Camera view adjustment prediction for improving image composition," *arXiv preprint arXiv:2104.07608*, 2021.

[2] Ying Xu, Yi Wang, Huaixuan Zhang, and Yong Jiang, "Spatial attentive image aesthetic assessment," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[3] Wolff Schlotz, Sebastian Wallot, Diana Omigie, Michael D Masucci, Sonja C Hoelzmann, and Edward A Vessel, "The aesthetic responsiveness assessment (area): A screening tool to assess individual differences in responsiveness to art in english and german.," *Psychology of Aesthetics, Creativity, and the Arts*, 2020.

[4] Hossein Talebi and Peyman Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[5] Jan Pfister, Konstantin Kobs, and Andreas Hotho, "Self-supervised multi-task pretraining improves image aesthetic assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 816–825.

[6] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 9375–9383.

[7] Lohic Fotio Tiotsop, Tomas Mizdos, Miroslav Uhrina, Marcus Barkowsky, Peter Pocta, and Enrico Masala, "Modeling and estimating the subjects' diversity of opinions in video quality assessment: a neural network based approach," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3469–3487, 2021.

[8] Suiyi Ling, Jing Li, Anne Flore Perrin, Zhi Li, Lukáš Krasula, and Patrick Le Callet, "Strategy for boosting pair comparison and improving quality assessment accuracy," *arXiv preprint arXiv:2010.00370*, 2020.

[9] Congcong Li and Tsuhan Chen, "Aesthetic visual quality assessment of paintings," *IEEE Journal of selected topics in Signal Processing*, vol. 3, no. 2, pp. 236–252, 2009.

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

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

[12] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

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

[14] Shuang Ma, Jing Liu, and Chang Wen Chen, "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*, 2017, pp. 4535–4544.

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

[16] Yong-Lian Hii, John See, Magzhan Kairanbay, and Lai-Kuan Wong, "Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1722–1726.

[17] Katharina Schwarz, Patrick Wieschollek, and Hendrik PA Lensch, "Will people like your image? learning the aesthetic space," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 2048–2057.

[18] Qi Kuang, Xin Jin, Qinping Zhao, and Bin Zhou, "Deep multimodality learning for uav video aesthetic quality assessment," *IEEE Transactions on Multimedia*, 2019.

[19] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14114–14123.

[20] Shikun Liu, Edward Johns, and Andrew J Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1871–1880.

[21] Deborah J Rumsey, *Statistics II for dummies*, John Wiley & Sons, 2009.

[22] Suiyi Ling, Junle Wang, Wenming Huang, Yundi Guo, Like Zhang, Yanqing Jing, and Patrick Le Callet, "A subjective study of multi-dimensional aesthetic assessment for mobile game image," in *Proceedings of the 1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications*, 2020, pp. 47–53.

[23] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.