Multi-Modal Aesthetic Assessment for Mobile Gaming Image

Zhenyu Lei, Yejing Xie, Suiyi Ling, Andreas Pastor, Junle Wang, Junyu Dong, Patrick Le Callet

MULTI-MODAL AESTHETIC ASSESSMENT FOR MOBILE GAMING IMAGE

Zhenyu Lei[1], Yejing Xie[1], Suiyi Ling[1], Andréas Pastor[1], Junle Wang[2], Junyu Dong[3], Patrick Le Callet[1]

[1]LS2N, University of Nantes [2]Turing Lab, Tencent [3]Ocean University of China

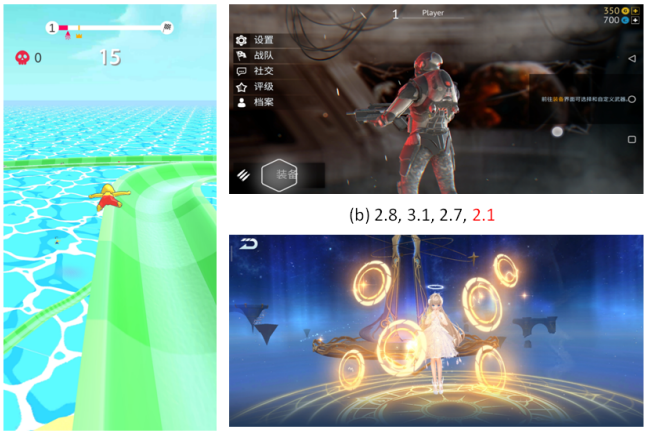
ABSTRACT

With the proliferation of various gaming technology, services, game styles, and platforms, multi-dimensional aesthetic assessment of the gaming contents is becoming more and more important for the gaming industry. Depending on the diverse needs of diversified game players, game designers, graphical developers, *etc.* in particular conditions, multi-modal aesthetic assessment is required to consider different aesthetic dimensions/perspectives. Since there are different underlying relationships between different aesthetic dimensions, *e.g.,* between the 'Colorfulness' and 'Color Harmony', it could be advantageous to leverage effective information attached in multiple relevant dimensions. To this end, we solve this problem via multi-task learning. Our inclination is to seek and learn the correlations between different aesthetic relevant dimensions to further boost the generalization performance in predicting all the aesthetic dimensions. Therefore, the 'bottleneck' of obtaining good predictions with limited labeled data for one individual dimension could be unplugged by harnessing complementary sources of other dimensions, *i.e.,* augment the training data indirectly by sharing training information across dimensions. According to experimental results, the proposed model outperforms state-of-the-art aesthetic metrics significantly in predicting four gaming aesthetic dimensions.

Index Terms— Image aesthetic assessment, multi-task learning, multi-modal image Aesthetic assessment, mobile game image, aesthetic assessment of graphical content

1. INTRODUCTION

The last decade has witnessed an exceeding boost of mobile games with diverse gaming styles, and growing expectations of higher gaming quality regarding divergent aesthetic aspects. In catching up with the increasingly diverse needs, multi-modal aesthetic evaluation models that take into account the entire game design, development, quality-control, pipeline is essential [1]. Robust objective multi-dimensional aesthetic assessment metrics are in need to offer specific guidance for game designers and developers concerning different game styles [2]; leverage trade-off between the gaming-graphic complexity and the resource consumed for different gaming contents based on players' preferences (setting); ensure gaming streaming quality, *e.g.,* game video streaming platforms, online Cloud Gaming [3] *etc.*



(a) 1.7, 1.7, 1.6, 3.4 (b) 2.8, 3.1, 2.7, 2.1 (c) 3.5, 3.4, 3.5, 3.5

Fig. 1: Examples of four-dimensional aesthetic scores defined in [1], from left to right: the 'Overall Aesthetic', the 'Colorfulness', the 'Fineness', and the 'Color Harmony'.

Heaps of different factors could be considered for aesthetic assessment of gaming images/videos in the modern gaming industry. Among the varied potential dimensions, including game style, equipment-related viewing conditions, illuminance simulation, shadow generation, gaming sound effect, rendered quality, and so on. In [1], four aesthetic dimensions were considered, as they are less subjective. More specifically, these four dimensions include (1) the 'Overall aesthetic quality', which evaluate the quality of an image in the sense of visual aesthetics, and rate it regarding whether it is beautiful or not in human's eyes; (2) the 'Colorfulness', which assesses the amount, intensity, and saturation of colors in an image; (3) the 'Fineness', which quantifies the details, granularity of an image; and (4) the 'Color Harmony', which evaluates the property that certain aesthetically pleasing color combinations possess. It is also stated in [1] that there are different relationships between the four aesthetic dimensions, and they are content-dependent. Examples are depicted in Fig. 1. For sub-figure (a), it has a high 'Color Harmony' score (pleasing color combination), but its 'Fineness' (the graphic content is coarse), and 'Colorfulness' (contains only cold colors) scores are low. Therefore, its corresponding 'Overall Aesthetic' is low. Oppositely, although sub-figure (b) has a

Zhenyu Lei, Yejing Xie, and Suiyi Ling make equal contributions.

considerably low 'Color Harmony' score due to overall gray color, its 'Overall Aesthetic' is still higher as it contains finer details. Obviously, different dimensions correlate with each other differently.

In the literature, most of the existing studies about aesthetic assessment are restricted to natural content, *i.e.*, considering only photography images. Furthermore, none of them was developed to predict multiple aesthetic dimensions due to limited labels on other aesthetic dimensions. Not to mention developing multi-modal metrics by exploring the underlying correlations among aesthetic dimensions. In this study, we thus aim to develop a gaming-specific aesthetic metric that is in light of the peculiarities of the gaming contents via multi-task learning.

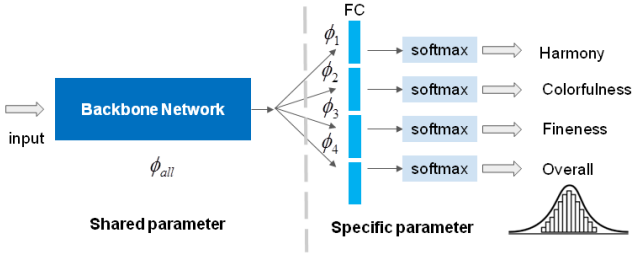


Fig. 2: The diagram of the proposed multi-modality model.

2. RELATED WORK

Recently, the performances of aesthetic assessment models grow at a respectable pace. Li *et al.* [4] proposed a one of early efficient aesthetic metric based on hand-crafted feature. Follow a similar recipe, another aesthetic approach was presented in [5] by combing faces, technical, perceptual, and social relationship features. By formulating aesthetic quality assessment as a ranking problem, [6], Kao *et al.* developed a rank-based methodology for aethetic assessment. Akin to [6], another ranking network was proposed in [7] with attributes and content adaptation. To facilitate heterogeneous input, a double-column deep network architecture was presented in [8], which was improved subsequently in [9] with a novel multiple columns architecture. Ma *et al.* developed a salient patch selection approach [10] that achieved significant improvements. By introducing a five spatial pooling sizes method [11], state-of-the-art models by that time were enhanced with appreciable margins. Three individual convolutional neural network (CNN) that capture different types of information were trained and integrated into one final aesthetic identifier in [12]. Global average pooled activations were utilize by Hii *et al.* in [13] to take the image distortions into account. Later, triplet loss was employed in deep framework in [14] to further push the performances to the limits of most modern methods available at the time. The Neural IMage Assessment (NIMA) [15], 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 [16]. As global pooling is conducive to arbitrary high-resolution input, MLSP [17] was proposed in based on Multi-Level Spatially Pooled features. Even though some of the state-of-the-art models achieve appealing performances, none of them were designed dedicated for mobile gaming images considering multiple aesthetic dimensions.

3. THE PROPOSED FRAMEWORK

Details of the proposed multi-modality aesthetic model are given in this section. The overall framework of the model is summarized in Fig. 2.

Multi-Task Learning is a multi-modality learning paradigm that tends to leverage useful information contained in multiple relevant tasks so that the overall performances of all the related tasks could be improved by sharing generalization information [18]. As mentioned in Section 1, the four aesthetic dimensions studied in [1], correlates differently with each other. For instance, the 'Fineness', 'Colorfulness' dimensions both correlates well with the 'Overall aesthetic score', while the correlations between 'Color Harmony' and other dimensions are low. Based on these observations, we propose to train a multi-modal aesthetic assessment metric by considering the intricate correlations among different dimensions using multi-task learning techniques.

Given T task, with N_t samples per task. In general, the objective of common multi-task learning models was designed as the linear combination of the losses across T task, with weights w_t corresponds to each individual task:

$$\underset{\phi^{all},\phi^1,\ldots,\phi^T}{\arg\min} \sum_{t=1}^{T} w_t \mathcal{L}^t(x_i^t, y_i^t; \phi^t, \phi^{all}), \tag{1}$$

where $\mathcal{L}^t(\cdot)$ indicates the loss, *i.e.*, the empirical risk, of the t_{th} task. In this study, the r-norm Earth Mover's Distance loss [19, 15] was employed for $\mathcal{L}^t\left(\boldsymbol{\phi}^{all}, \boldsymbol{\phi}^t\right)$ to better capture the complex inter-class relationships:

$$\mathrm{EMD}(\mathbf{y}, \hat{\mathbf{y}}) = \left(\frac{1}{N_c} \sum_{c=1}^{N_c} |\mathrm{CDF}_{\mathbf{y}}(c) - \mathrm{CDF}_{\hat{\mathbf{y}}}(c)|^r \right)^{1/r}, \tag{2}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the probability distributions of the ground-truth and the prediction respectively. $\mathrm{CDF}_{\mathbf{y}}(\cdot)$ is the cumulative distribution function of a distribution y. N denotes the number of aesthetic levels, and c indicates the c_{th} aesthetic level. Different from the influential AVA dataset [20] (10 aesthetic level), the aesthetic scores collected from [1] fall in the range of $[1, 5]$. In this study, $N_c = 5$. The EMD loss is defined as the minimum cost of transporting values from one distribution to another. Intuitively, it penalizes the inaccurate prediction via accumulating the distances between the distribution, *i.e.*, the likelihoods of each aesthetic level, of the ground-truths and the ones of the predictions.

In (1), x_i^t and y_i^t denote the i_{th} sample and its corresponding label for the t_{th} task. ϕ^{all} is the model parameters (network) that shared by all the tasks, when ϕ^t is the set of specific parameters (network) for the t_{th} task. According to [21], most of the existing deep MTL models could be covered by a 'Encoder-Decoder-like' network architecture, where the relationship between the shared representation function and the task-specific decision functions was given by:

$$f^t\left(\mathbf{x}; \boldsymbol{\phi}^{sh}, \boldsymbol{\phi}^t\right) = f^t\left(f^{all}\left(\mathbf{x}; \boldsymbol{\phi}^{sh}\right); \boldsymbol{\phi}^t\right), \tag{3}$$

f^{all} is the shared network with the shared parameters of all the tasks, and f^t is the individual network branch with the specific parameters for each t task. As depicted in Fig. 2, in this work, a backbone network was utilized as the shared network f^{all} to extract the latent representation, followed by four Fully Connected (FC) layers as the task-specific network for the four aesthetic dimensions. It is worth noting that different, especially light-weight, network architectures could be utilized as the shared network, *e.g.,* the ResNet [22], ShuffletNet [23], GoogLeNet [24], *etc.*

However, as mentioned earlier, the aesthetic dimension of 'Color Harmony' correlates poorly with the other dimensions, indicating higher conflicts against other dimensions. These conflicting relations among tasks is hard for a simple linear weighted combination of losses across task to tackle. Another alternative is to form the objective function that aims to search for the Pareto optimal solutions. Particularly, this type of MTL algorithms seek to find solutions that are not suppressing any others, and solve the problem via gradient-based Multi-Objective Optimization (MOO):

$$\begin{aligned} &\arg\min_{\delta^1,\ldots,\delta^T} \left\| \sum_{t=1}^{T} \delta^t \nabla_{\boldsymbol{\phi}^{\mathbf{all}}} \mathcal{L}^t\left(\boldsymbol{\phi}^{all}, \boldsymbol{\phi}^t\right) \right\|_2^2 \\ &s.t. \sum_{t=1}^{T} \delta^t = 1, \forall t\ \delta^t \geq 0 \end{aligned} \tag{4}$$

where $\nabla_{\boldsymbol{\phi}^{\mathbf{all}}}$ is the gradient of the shared parameters, and $\delta^1, ..., \delta^t$ are the solutions, *i.e.,* corresponding weights of tasks. In brief, gradient descent was employed on the specific parameters regarding (4), and utilizes $\sum_{t=1}^{T} \delta^t \nabla_{\boldsymbol{\phi}^{\mathbf{all}}}$ as a gradient update for shared parameters. For instance, Multiple-gradient descent algorithm (MGDA) [18] is one of the most efficient MOO based approaches. Nonetheless, it is not suitable to be applied directly for high-dimensional problems, especially with limited amount of training samples, and it suffers from a high computation complexity per-task. To further overcome these disadvantages, Multiple Gradient Descent Algorithm-Upper Bound (MGDA-UB) [21] was proposed by optimizing an upper bound of the MOO objective with only one pass backward propagation. More concretely, for each sample x_i^t of the t_{th} task, its corresponding shared representations could be obtained via feeding it to f^{all}, *i.e.,* $g_i = f^{all}(x_i^t; \phi^{all})$. For N_t samples of the t_{th} task, their representations can be gathered as $\mathcal{G} = (g_1, ..., g_{N^t})$. By employing the Frank-Wolf solver, it was shown in [21] that an upper bound of the objective could be obtained after applying the chain rule:

$$\begin{aligned} &\left\| \sum_{t=1}^{T} \delta^t \nabla_{\boldsymbol{\phi}^{\mathbf{all}}} \mathcal{L}^t\left(\boldsymbol{\phi}^{all}, \boldsymbol{\phi}^t\right) \right\|_2^2 \\ &\leq \left\| \frac{\partial\, \mathcal{G}}{\partial \phi^{all}} \right\|_2^2 \left\| \sum_{t=1}^{T} \delta^t \nabla_{\mathcal{G}} \mathcal{L}^t\left(\boldsymbol{\phi}^{all}, \boldsymbol{\phi}^t\right) \right\|_2^2, \end{aligned} \tag{5}$$

where the matrix norm of the Jacobian of $\mathcal{G}$, *i.e.,* $\left\| \frac{\partial\, \mathcal{G}}{\partial \phi^{all}} \right\|_2^2$ can be omitted as it does not contain δ. As such, equation (4) can be re-written by simply replacing $\nabla_{\boldsymbol{\phi}^{\mathbf{all}}}$ with $\nabla_{\mathcal{G}}$.

4. EXPERIMENT

4.1. Experimental Setup

The performance of the proposed model is evaluated on the Tencent Mobile Gaming Aesthetic (TMGA) dataset [1], which is the only existing public multi-modality aesthetic dataset. In this dataset, there are totally 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. All images were rescaled and padded into a size of 454×984, without changing the aspect-ratio, of the input image for training efficiency. Different network architectures have been explored, including the ResNet-18 and ResNet-50, as the backbone network of the encoder within our multi-task framework. During training, the momentum SGD optimizer was utilized with a momentum of equals to 0.9. The learning rate was set as 10^{-4} at the beginning of training and was halve every 30 epochs. Experiments were conducted with a machine equipped with an Nvidia GeForce RTX 2080 Ti GPU. All models were implemented using PyTorch.

For fair comparisons, when reporting the performances of the deep-learning based models, like NIMA and MLSP, we first finetuned their models on the training set of TMGA dataset with the best configurations, *e.g.,* best hyper-parameters, network architectures, *etc.* As highlighted in [17], the predominant performance evaluation measure, *i.e.,* the binary classification accuracy, suffers from several drawbacks. For example, due to the unbalanced distribution of images in training, testing set (unbalanced in terms of different aesthetic quality ranges), using 'accuracy' does not necessarily reveal/stress out the performances of under-test metrics regarding its capability in ranking the aesthetic score of the image. Therefore, similar to [17, 1], we calculated the Pearson correlation coefficient (PCC), Spearman's rank order correlation coefficient (SCC), and Root mean squared error

(RMSE) between the ground truth and the predicted scores to benchmark different objective aesthetic metrics.

4.2. Experimental Results

The overall results are shown by Table 1. On the whole, the proposed multi-task model outperforms all the compared state-of-the-art no reference aesthetic metrics in terms of predicting the four aesthetic scores. Affirmatively, our model surpasses the traditional non-deep-leaning models significantly with large margins. To further confirm whether the difference of performances between the proposed model and the two other deep-learning based models are significant, the F-test based significant analysis as presented in [25] was utilized. It is shown that our model outperforms NIMA and MLSP significantly in predicting the scores of all four dimensions. As we used a similar loss function and an analogous backbone network, with similar FC layers as used in the NIMA framework, it was thus considered as a baseline model without using the multi-task learning. The boosted results compared to single-task model NIMA also demonstrate the effeteness of applying multi-task learning. It is proven that, by fully mining and leveraging the internal correlations between different aesthetic dimensions, the overall performances can be improved significantly.

Table 1: Performances of no reference image metrics.

	Fineness	**Colorful**	**Harmony**	**Overall**
Pearson correlation coefficient (PCC)				
Color [26]	0.3353	0.3624	0.6563	0.3679
CPBD [27]	0.5545	0.6007	0.3171	0.4868
Blur [28]	0.1412	0.1293	0.1783	0.1408
NIMA [15]	0.8414	0.8330	0.8397	0.8255
MLSP [17]	0.9046	0.9004	0.8885	0.8724
Proposed	**0.9266**	**0.9330**	**0.8982**	**0.9113**
Spearman's rank order correlation coefficient (SCC)				
Color [26]	0.3376	0.3651	0.5992	0.3632
CPBD [27]	0.4297	0.4322	0.2799	0.3874
Blur [28]	0.1121	0.0966	0.1400	0.1171
NIMA [15]	0.8392	0.8428	0.7661	0.8209
MLSP [17]	0.9047	0.9045	0.8262	0.8652
Proposed	**0.9260**	**0.9276**	**0.8592**	**0.9030**
Root mean squared error (RMSE)				
Color [26]	0.6590	0.7440	0.4500	0.6013
CPBD [27]	0.5818	0.6381	0.5655	0.5648
Blur [28]	0.6921	0.7915	0.5867	0.6401
NIMA [15]	0.3998	0.4669	0.3232	0.4381
MLSP [17]	0.3622	0.4143	0.3067	0.3137
Proposed	**0.2813**	**0.3093**	**0.2599**	**0.2944**

4.2.1. Ablation Study

Extensive ablation studies have been conducted to explore the impact of different settings on the performances.

- **Impact of different network architecture:** In this work, we delved into different network architectures backbone network for the shared network f^{all}, including ResNet-18, ResNet-50, VGG16, *etc.* Due to limited space, only the top two architectures are reported.
- **Impact of different multi-patch strategies:** As demonstrated in [29, 30] that the performances of random patch-selection strategy based aesthetic assessment models can be improved by applying the aspect-ratio-preserving Multi-Patch (MP) approach. Be that as it may, as also pointed out in [31] that predicting quality/aesthetic scores of an image based on patches may be less accurate due to the loss of global information. Hence, in contemplation of the common MP method, an adapted 'MP with Global Patch' strategy, namely the 'MP with GP', was explored in this study to take the global information into account. Notably, a set of global patches was added to the whole patch set by randomly cropping and resizing the original input into new patches with similar size to the local patches without changing the original aspect-ratio.

The ablation results are presented in Table 2. It is evident that there is no significant difference between the framework using the multi-patch strategy with and without the global patch. Surprisingly, the framework with the multi-patch strategy (1-2 row in the Table) does not outperform the ones with *'padding + re-scaling'* (3-4 row in the Table). It is showcased that, for the aesthetic evaluation of mobile gaming images, *'padding + re-scaling'* strategy is more suitable. A few of these factors could be that, unlike natural images with diverse patches, gaming images are normally generated graphical content. The first overall aesthetic impression matters more than local details. As a result, a strategy that preserves the overall structure of the image fits better the scenario. Regarding different backbone architecture (Due to limited space, only the results of the two top network architectures were presented), it is observed that some improvement could be obtained by utilizing ResNet-50 instead of ResNet-18.

Table 2: Results of Ablation Study, where 'Multi-Patch' is denoted as 'MP', and 'Global Patch' as 'GP'.

SCC	**Fineness**	**Colorful**	**Harmony**	**Overall**
MP with GP	0.9167	0.9113	0.8268	0.8819
MP without GP	0.9139	0.9197	0.8207	0.8782
ResNet-18	0.9220	0.9242	0.8535	0.9020
ResNet-50	0.9260	0.9276	0.8592	0.9030

5. CONCLUSION

In this work, by observing the correlations between different aesthetic dimensions, a multi-task learning based model is developed for mobile gaming images. Extensive experiments have demonstrated that the proposed model is superior to the compared state-of-the-art aesthetic models. It was also found that, when dealing with different images with different resolutions, the re-scaling plus padding strategy is more suitable for gaming contents compared to the multi-patch approach.

6. REFERENCES

[1] 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.

[2] Simon Niedenthal, "What we talk about when we talk about game aesthetics," 2009.

[3] Asif Ali Laghari, Hui He, Kamran Ali Memon, Rashid Ali Laghari, Imtiaz Ali Halepoto, and Asiya Khan, "Quality of experience (qoe) in cloud gaming architectures: A review," 2019.

[4] 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.

[5] Congcong Li, Andrew Gallagher, Alexander C Loui, and Tsuhan Chen, "Aesthetic quality assessment of consumer photos with faces," in *2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 3221–3224.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] Jean-Antoine Désidéri, "Multiple-gradient descent algorithm (mgda) for multi-objective optimization," *Comptes Rendus Mathematique*, vol. 350, no. 5-6, pp. 313–318, 2012.

[19] Le Hou, Chen-Ping Yu, and Dimitris Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv preprint arXiv:1611.05916*, 2016.

[20] 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.

[21] Ozan Sener and Vladlen Koltun, "Multi-task learning as multi-objective optimization," *Advances in Neural Information Processing Systems*, vol. 31, pp. 527–538, 2018.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[25] Suiyi Ling, Jesús Gutiérrez, Ke Gu, and Patrick Le Callet, "Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 204–216, 2019.

[26] David Hasler and Sabine E Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*. International Society for Optics and Photonics, 2003, vol. 5007, pp. 87–95.

[27] Niranjan D Narvekar and Lina J Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (cpbd)," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2678–2683, 2011.

[28] Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," in *Human vision and electronic imaging XII*. International Society for Optics and Photonics, 2007, vol. 6492, p. 64920I.

[29] Lijie Wang, Xueting Wang, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Aspect-ratio-preserving multi-patch image aesthetics score prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

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

[31] Suiyi Ling, Jing Li, Zhaohui Che, Junle Wang, Wei Zhou, and Patrick Le Callet, "Re-visiting discriminator for blind free-viewpoint image quality assessment," *IEEE Transactions on Multimedia*, 2020.