Towards Artistic Image Aesthetics Assessment: a Large-scale Dataset and a New Method (Supplementary Material)

Ran Yi[1]*, Haoyuan Tian[1], Zhihao Gu[1], Yukun Lai[2], Paul L. Rosin[2]
[1]Shanghai Jiao Tong University, [2]Cardiff University
{ranyi,thy0210,ellery-holmes}@sjtu.edu.cn, {LaiY4,RosinPL}@cardiff.ac.uk

1. Overview

In this supplementary material, more discussion, visualization and experimental results are provided, which are organized as follows:

- Sec. 2 provides more details about the data collection (Sec. 2.1) and analysis (Sec. 2.2) of the proposed BAID dataset and the results of the MOS (Mean Opinion Score) test mentioned in the main paper (Sec. 2.3).
- Sec. 3 conducts an ablation study on each of our newly added operations.
- Sec. 4 provides an evaluation of the style-specific aesthetic branch in the proposed SAAN (Sec. 4.1), evaluates the performance of SAAN on the AVA dataset [9] (Sec. 4.2), and gives more prediction results on the test set of BAID (Sec. 4.3).

2. More analysis of the proposed BAID

2.1. User Interface of BoldBrush

In Section 3 of the main paper, we discuss the construction of our proposed BoldBrush Artistic Image Dataset (BAID). Fig. 1 shows the detail page of an entry on the BoldBrush (https://faso.com/boldbrush/popular) website:

The available information of an entry on BoldBrush includes: the title; the artist; the painting medium; the entry number and month of entry; the number of votes; the category of the entry. In this work, we utilize the number of votes and the month of entry to generate score annotations and form a large-scale artistic image aesthetic assessment dataset, BAID. Meanwhile, we collect and save all the above information. We believe the BAID dataset can effectively serve as a foundation for constructing artistic image datasets for other purposes, *e.g.*, developing automatic artist [2, 3] and style [1, 10] classification methods.

*Corresponding author.

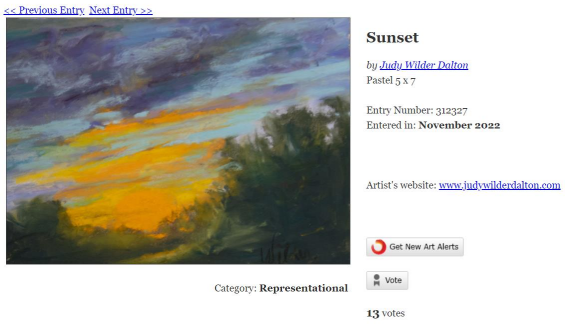

Figure 1. Interface of the BoldBrush website

Table 1. The most frequently used painting media in BAID and the average score of artworks created in these media.

Painting Medium	Number of Images	Average Score ↑
Oil	38,586	4.27
Acrylic	6,733	4.30
Watercolor	5,328	4.24
Pastel	5,156	4.22
Pencil	1,063	4.34

Table 2. Correlation between scores and hand-crafted features.

Features	SRCC ↑
Colorfulness	0.011
Contrast	0.049
Sharpness	0.029
Complexity	0.014

2.2. Further analysis

The generation of the score annotations in BAID is based on votes, which makes it hard to filter out unreasonable labels. To eliminate the concern about bias, we calcu-

late the most frequently used painting media and the average scores of the artworks created using the specific media. The results are shown in Tab. 1. Furthermore, following the benchmarks applied in [8], we calculate several hand-crafted features and measure their correlation (*i.e.*, the Spearman Rank-order Correlation Coefficient, SRCC) with the scores of the artworks in the BAID dataset. We randomly select 6,400 images from BAID and the results are shown in Tab. 2. The results indicate that the proposed BAID suffers little from art preference bias and is of high credibility.

There is a potential concern regarding the data imbalance mentioned in Section 3.3 of the main paper. The score distribution of BAID is imbalanced but it reflects the realistic distribution. We did consider reducing the imbalance. However, the most effective way would be to abandon most of the images with low votes, which would result in a significant drop in the size of BAID. Besides, the imbalance is related to the nature of the original data, and we believe that a well-developed IAA method should be able to deal with such an imbalance.

2.3. Results of the MOS test

As mentioned in Section 3.3 of the main paper, we sampled 100 artworks uniformly across the range of scores from the proposed BAID. We asked 10 college students majoring in art and design to score for these samples and calculated the mean opinion score (MOS) for each sample. We compared several designed functions we have experimented with during the construction of BAID. In the following equations, v_i denotes the number of votes of the image, $\bar{v}_{m_i}$ denotes the average number of votes of the month m_i, $\hat{v}_{m_i}$ denotes the maximum number of votes of the month m_i, and s_i denotes the generated score.

Choice A:

$$s_i = 5 \times \frac{v}{\bar{v}_{m_i}}, \tag{1}$$

Choice B:

$$\begin{aligned} s_i &= 5 - 5 \times \frac{\bar{v}_{m_i} - v}{\bar{v}_{m_i}}, (v \leq \bar{v}_{m_i}) \\ s_i &= 5 + 5 \times \frac{v}{\hat{v}_{m_i} - \bar{v}_{m_i}}, (v > \bar{v}_{m_i}) \end{aligned} \tag{2}$$

Choice C:

$$\begin{aligned} s_i &= 5 - v \times \frac{\bar{v}_{m_i} - v}{\bar{v}_{m_i}}, (v \leq \bar{v}_{m_i}) \\ s_i &= 5 + 5 \times \frac{v}{\hat{v}_{m_i} - \bar{v}_{m_i}}, (v > \bar{v}_{m_i}) \end{aligned} \tag{3}$$

Ours:

$$\begin{aligned} x_i &= \frac{\bar{v}_{m_i} - v_i}{\bar{v}_{m_i}}, \\ s_i &= 10 \times \frac{1}{1 + e^{x_i}}, \end{aligned} \tag{4}$$

Table 3. Comparison of different score-generating functions

Method	SRCC ↑	RMSE ↓
A	0.221	0.980
B	0.576	0.502
C	0.594	0.492
Ours	**0.734**	**0.305**

Note that, images with $\bar{v}_{m_i}$ votes are supposed to be given the score of 5, which leaves us few options when designing the score-generating function. We calculated the spearman rank-order correlation coefficient (SRCC) and root mean squared error (RMSE) between the scores generated by the above functions and the MOS results. The results are shown in Tab. 3, which indicates that our chosen method better reflects human aesthetics.

The designed method seems similar to and may be confused with psychometric scaling of human votes [7]. However, the votes in BAID are different from the ones commonly used in psychometric scaling tasks since a vote itself is not a personal opinion score or a binary variable.

3. More ablation study results

In Sections 4 and 5 of the main paper, we demonstrate the effectiveness of our proposed operation list compared to the one in [12]. Here we provide more results of the ablation study on each of the newly added operations:

Table 4. Ablation study results on the newly added operations.

Method	SRCC ↑	PCC ↑	Accuracy ↑
w/o cropping	0.471	0.463	76.59%
w/o stylization	0.471	0.462	76.58%
w/o convex	0.471	0.464	76.60%
w/o pencilsketch	0.472	0.465	76.63%
w/o cutmix [16]	0.470	0.462	76.65%
w/o new editing operations	0.460	0.445	76.14%
Ours	**0.473**	**0.467**	**76.80%**

We select one of the newly added operations at a time and discard it during the pretraining stage. The impact on the final assessment performance is shown in Tab. 4. Results demonstrate that all newly added operations improve the performance, where operations related to global aesthetic features (*e.g.* Cutmix [16], Cropping) are relatively more influential in learning aesthetic-aware features, while the PencilSketch operation is less powerful since it may generate low-level artifacts (*i.e.*, unnecessary lines) and can trick the network into learning trivial features. We also experiment with all new editing operations removed, and it leads to more significant performance drop.

4. More performance evaluation results

4.1. Evaluation of the Style-specific Aesthetic Branch

In Section 4.1 of the main paper, we propose a style-specific aesthetic branch, which adopts a VGG-19 [14] backbone to extract the style feature of the input image, and incorporate the style information into aesthetic features to obtain style-specific aesthetic features.

To better illustrate the effect of incorporating style feature into the assessing process, given an artwork, we randomly select several images with different styles from the artworks in the test set of the BAID dataset, make them the input of the style feature extractor (VGG-19 backbone) and compare the predicted aesthetic scores. The experimental setting is shown in Fig. 2. Since the goal of this branch is to extract style-related aesthetic features, if we extract a different style's feature, and incorporate the 'wrong' style into the aesthetic features, then the calculated style-specific aesthetic feature is not dedicated to the current style, and the predicted aesthetic score is expected to decrease.

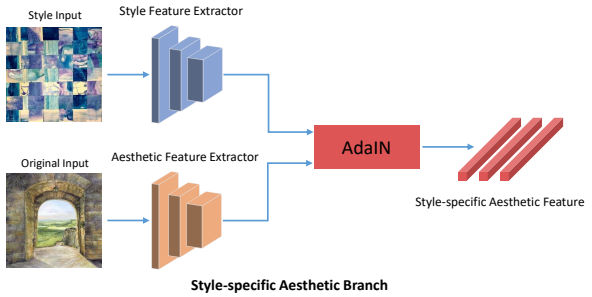

Figure 2. Validation of the style-specific aesthetic branch. We manually change the input to the style feature extractor (VGG-19) to be different from the original input in style. Note that the goal is still to predict the aesthetic score of the original input.

Fig. 3 shows the results of using different style inputs when evaluating artistic images. When the style feature is extracted from an artwork with a different style from the original input, the predicted aesthetic score will decrease and the prediction error will increase, which further validates our idea of utilizing style information in the AIAA task.

4.2. Performance on AVA dataset

We modified the output layer of SAAN and trained it on AVA dataset [9] using EMD (Earth Mover's Distance) loss [15]. Tab. 5 shows the performance of the state-of-the-art methods and SAAN on AVA dataset. The results of the state-of-the-art (SOTA) IAA methods come from the original papers and [4]. Our model gives competitive results compared with the SOTA methods. We believe SAAN works better on BAID since the distortions we used in the pretraining stage and the style-specific aesthetic branch are designed for and better suited to artistic images.

Table 5. Comparison with the SOTA IAA methods on AVA.

Methods	SRCC ↑	LCC ↑	Accuracy ↑	EMD ↓
NIMA [15]	0.612	0.636	81.5%	0.050
MP_{ada} [13]	0.727	0.731	83.0%	-
MLSP [5]	0.756	0.757	81.7%	
BIAA [17]	0.651	0.668	-	-
PA_IAA [6]	0.677	-	83.7%	0.049
HLA-GCN [11]	0.665	0.687	**84.6%**	**0.043**
TANet [4]	**0.758**	**0.765**	-	0.047
Ours	0.742	0.748	80.6%	0.048

4.3. Visualization of the prediction results

Fig. 4 shows the aesthetic score prediction results on some randomly picked artistic images from the test set of the proposed BAID dataset.

References

[1] Yaniv Bar, Noga Levy, and Lior Wolf. Classification of artistic styles using binarized features derived from a deep neural network. In *Proceedings of the European Conference on Computer Vision*, pages 71–84. Springer, 2014. 1

[2] Eva Cetinic and Sonja Grgic. Automated painter recognition based on image feature extraction. In *Proceedings ELMAR-2013*, pages 19–22. IEEE, 2013. 1

[3] Omid E David and Nathan S Netanyahu. DeepPainter: Painter classification using deep convolutional autoencoders. In *International Conference on Artificial Neural Networks*, pages 20–28. Springer, 2016. 1

[4] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 942–948. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track. 3

[5] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9375–9383, 2019. 3

[6] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing*, 29:3898–3910, 2020. 3

[7] Aliaksei Mikhailiuk, María Pérez-Ortiz, and Rafal Mantiuk. Psychometric scaling of TID2013 dataset. In *Proceedings of International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018. 2

[8] David Mould and Paul L Rosin. Developing and applying a benchmark for evaluating image stylization. *Computers & Graphics*, 67:58–76, 2017. 2

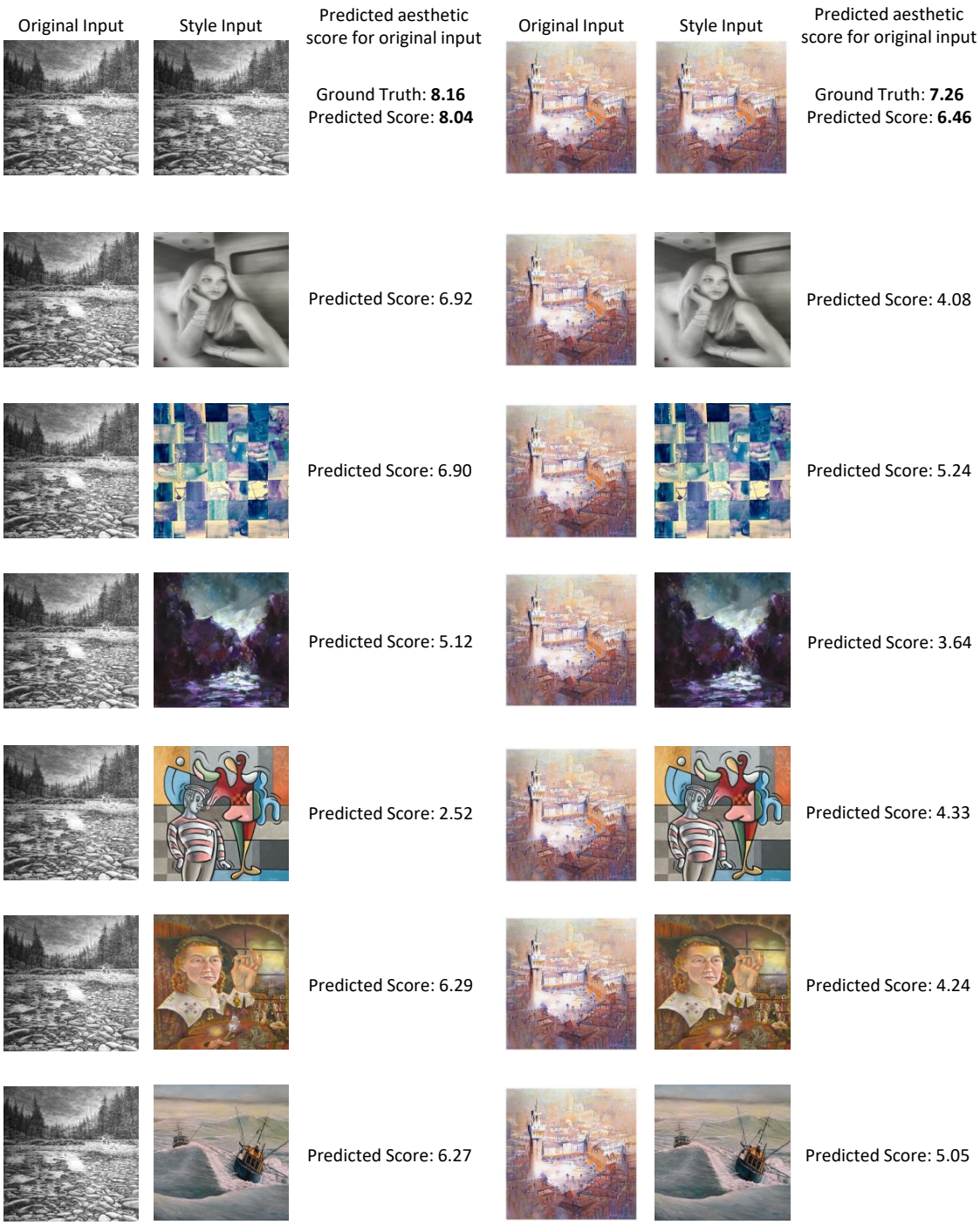

Figure 3. Prediction results of changing the input to the style feature extractor to be different from the original input in style (Fig. 2).

[9] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012. 1, 3

[10] Lior Shamir, Tomasz Macura, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception (TAP)*, 7(2):1–17, 2010. 1

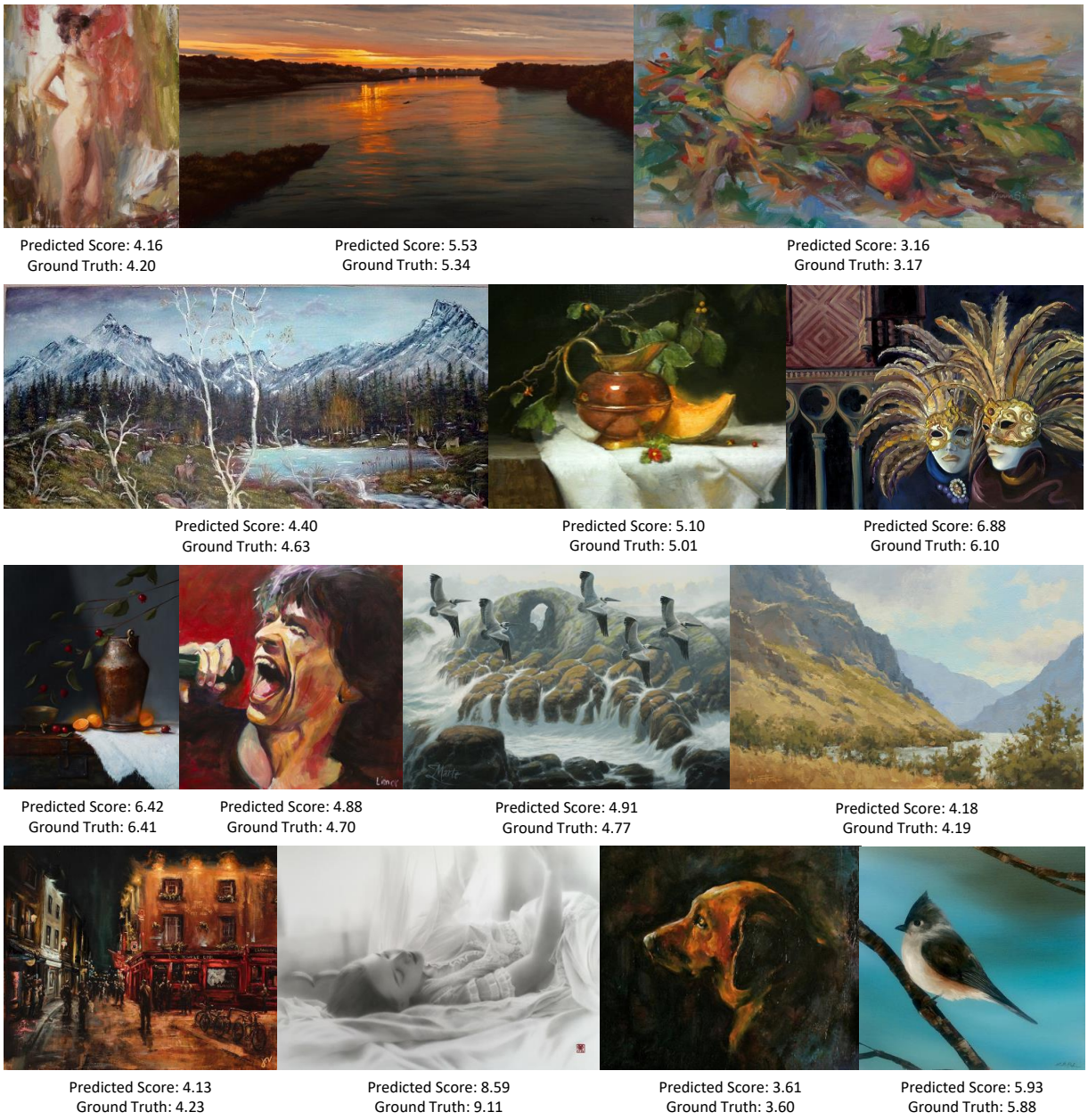

Figure 4. Some results on the test set of BAID, showing both the predicted scores by our method and ground truth scores.

[11] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8475–8484, 2021. 3

[12] Kekai Sheng, Weiming Dong, Menglei Chai, Guohui Wang, Peng Zhou, Feiyue Huang, Bao-Gang Hu, Rongrong Ji, and Chongyang Ma. Revisiting image aesthetic assessment via self-supervised feature learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5709–5716, 2020. 2

[13] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 879–886, 2018. 3

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[15] Hossein Talebi and Peyman Milanfar. NIMA: Neural im-

age assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. 3

[16] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 2

[17] Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics*, 2020. 3