Scaling Up Personalized Image Aesthetic Assessment via Task Vector Customization

A.1 Individual Personalization Results

The personalization of aesthetic assessment models is commonly evaluated based on the average personalization performance across multiple individuals, allowing us to test the approach on various individual preferences. While achieving a high average performance is important, it is equally crucial to ensure that personalization is effective for every individual, without exceptions. To address this, we present a visualization of the performance improvements observed after personalizing the aesthetic model for 37 individuals from the Flickr-AES database [17]. Specifically, we calculate the average zero-shot Spearman rank-order correlation coefficient (SROCC) across the six models (fine-tuned on six databases) as the baseline performance before personalization. We then plot the performance enhancements achieved after personalization with the 100 user-provided samples (*i.e.*, 100-shot). The average zero-shot SROCC can be understood as a simple logit-ensemble of the six models. As shown in Figure 1, it is clear that training the coefficients of the task vectors leads to a significant increase in personalization scores, demonstrating the effectiveness of our approach in adapting to the diverse personal preferences of individuals. We also add full details of individual persoanlization performances in Figures 2 to 7.

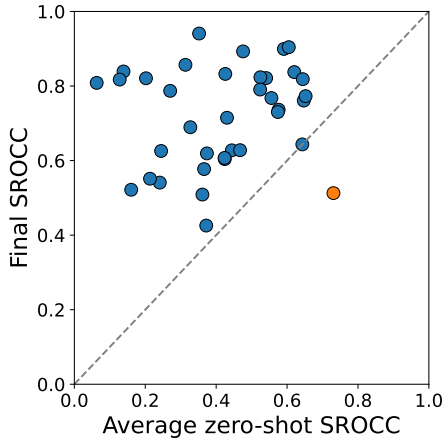

Fig. 1: Performance gains visualized for each individual in the Flickr-AES database [17]. Our approach generally improves the persoanlization performance for individuals.

Table 1: Comparison between the performance when deriving task vectors in a layer-agnostic and layer-wise manner.

Method	SROCC	
	10-shot	100-shot
PA-IAA [13]	0.443±0.004	0.562±0.013
BLG-PIAA [28]	0.448±0.007	0.578±0.015
PIAA-SOA [29]	0.487±0.003	0.589±0.015
TAPP-PIAA [14]	-	0.580
MTCL [24]	0.495±0.007	0.599±0.012
Ours (MSE loss)	0.553±0.002	0.600±0.003
Ours (layer-agnostic)	0.438±0.003	0.543±0.003
Ours	**0.572±0.005**	**0.621±0.007**

A.2 Importance of Layer-wise Task Vectors

We conduct an additional ablation study to assess the importance of deriving task vectors for each layer, as opposed to the method proposed by Ilharco *et al.* [7], which derives a single task vector for each task. When a single task vector is derived for each task, the total number of task vectors is reduced from $n \times L = 1776$ to just $n = 6$, where n represents the number of tasks, and L denotes the number of layers in the ViT-L/14 [2] model. The capacity of this approach, which we call as the layer-agnostic task vectors, is limited due to the extremely small number of trainable parameters. As seen in Table 1, training the coefficients of the layer-agnostic task vectors results in poor personalization performance, underlining the flexibility and importance of layer-wise task vectors.

A.3 Comparing Different Loss Functions

Given the distinct nature of our approach, which does not require training the entire model parameters, we explore various loss functions suitable for training the learnable coefficients. In the realm of general image aesthetic assessment (GIAA), two prevalent loss functions are the Earth Mover's Distance (EMD) loss and the Mean Squared Error (MSE) loss. However, EMD loss is exclusively applicable to databases like the AVA database [16], which collect multiple scores per image, making it less suited for personalization tasks involving individual preferences.

Considering an alternative, we turn to the MSE loss, known for its simplicity and effectiveness. Additionally, we delve into the Bradley-Terry model [1], extensively utilized in recent reinforcement learning with human feedback (RLHF) approaches for training reward models. Our findings indicate that the rank loss derived from the Bradley-Terry model excels in sample efficiency, leveraging a combination of samples for training. Additionally, we find that the convergence of the MSE loss is slower than the EMD loss, requiring twice as long training steps to achieve the accuracy reported in Table 1.

A.4 Detailed Architectural Designs

Task vector arithmetic, as described in [7], is performed by adding or subtracting the fine-tuned weights in an element-wise manner. Therefore, when combining multiple task vectors, they must be of the same size. However, architectural designs vary across tasks, as specialized strategies are employed for each task. Additionally, different databases have varying score ranges; some range from 1 to 10, while others range from 1 to 5.

To address these challenges, we present a model architecture that can accommodate all tasks and facilitate the derivation of task vectors. As mentioned in Section 5 of the manuscript, we select Vision Transformers [2] (ViT) as the suitable backbone architecture for deriving task vectors. Building upon the findings by Hosu *et al*. [6], we add two linear layers with the final output channel size fixed to 10. The final regression score is calculated as the dot product between the 10 output scores and a 10-dimensional template vector, which is a list ranging from 1 to 10. For tasks with score ranges from 1 to 5, we adjust the 10-dimensional template vector to include values from 0.5 to 5 in 0.5 increments. In this way, we are able to fix the number of parameters for all databases. The summary of these architectural details is as follows:

```
image_score_regression_model: {
    embed_dims: 768,
    backbone: {
        image_size: 224,
        layers: 24,
        width: 1024,
        patch_size: 14
    },
    head: {
        hidden_dim: 512,
        non_linearaity: GELU,
        dropout: 0.5,
        output_dim: 10,
        output_layer: Sigmoid
    }
}
```

A.5 Characteristics of the Training Database

In this section we provide descriptions on the data size, intentions, and collection schemes for each of the six training databases used to derive task vectors.

AVA. The AVA database [16] is the largest database for training models for general image aesthetic assessment (GIAA), containing over 200,000 images. These images, along with their corresponding aesthetic scores, are collected from a digital photography contest community, DPchallenge. Each image contains semantic tags and photographic style tags, along with a aesthetic score *vote* from multiple individuals. The number of votes per image ranges from 78 to

549, making the aesthetic score labels a robust representation of the community's average opinion. In our study, we exclusively utilize the aesthetic scores to train a single model. However, leveraging the rich semantic and photographic style tags to segment this vast database into multiple subsets represents a promising avenue for future research.

Flickr-AES. Utilizing images collected from the large photo collection Flickr, the Flickr-AES database [17] offers aesthetic score ratings for 40,000 images, with each image rated by 5 AMT annotators. Each image in this database has been rated by 5 AMT annotators. This diverse collection is suitable for both GIAA and PIAA tasks, as it includes tags for the annotator IDs with each image. When the dataset is restructured from the perspective of each individual annotator, it shows that 210 AMT annotators have each labeled between 105 to 171 images. However, we treat this database primarily for GIAA, as our scalable method for training PIAA models does not necessitate the use of annotator tags.

TAD66K. TADK66K [4] is one of the recently proposed databases for GIAA, specifically curated to ensure a balanced representation of various themes such as sunset, sea, winter, and street. Furthermore, different evaluation criteria were given to the annotators for different themes, as the aesthetic elements may differ for different themes. More than 60,000 images were collected and annotated by 1200 annotators.

PARA. The PARA database [23], meeting the needs of recent PIAA approaches, has gathered aesthetic scores of more than 30,000 images from 438 annotators, marking the largest involvement of annotators in any such database to date. The database also includes rich information about the annotators, such as personality traits, photographic and artistic experience, emotion, age and gender. Although these detailed attributes offer the potential to categorize annotators, our approach does not utilize these additional attributes for the derivation of task vectors.

KonIQ-10K. KonIQ-10K [5] stands as one of the pioneering large-scale image quality assessment (IQA) databases, distinct in its approach of collecting a variety of images and annotating their quality, rather than creating low-quality images by distorting high-quality ones. The 10,000 images in this database are labeled based on aspects of photographic quality such as brightness, colorfulness, and contrast, which is a criteria that differ from those used in GIAA databases. These alternative criteria provide unique task vectors, which are crucial for covering a wide range of personal preferences.

SPAQ. Specifically targeting the unique domain of smartphone photography, the SPAQ database [3] comprises a collection of 11,125 images captured using 66 different smartphones. Each image in this database is labeled according to its photographic quality. These collections offer domain-specific insights, particularly focusing on smartphone photography, thereby broadening the range of domains encompassed through the combination of task vectors.

A.6 Training Hyper-parameters

Our method comprises two key phases designed to optimize personalized model training. The initial phase focuses on generating task vectors for each database, laying the groundwork for our approach. This step, although time-intensive, is performed only once, enabling the model to be easily adapted for various individuals thereafter. The second phase is centered on fine-tuning the model for individual-specific requirements. Here, our objective is to minimize training time without compromising on performance. By investing in a robust first phase, we ensure the subsequent personalization process is both efficient and scalable, catering to multiple individuals with a single, comprehensive training cycle.

In the first stage, we initially train a regression head with the backbone parameters fixed, followed by fine-tuning the backbone. Note that this stage conducted *only once* and is not required again for personalizing to individual users. Furthermore, we will make available the model weights fine-tuned on each database to support future research efforts.

```
task_vector_acquisition_phase: {
    optimizer: AdamW,
    lr_schedule: Cosine annealing,
    train_head:{
        start_lr: 1.5e-5,
        end_lr: 1.5e-6,
        batch_size: 128,
        steps: 60,000
    }
    fine_tune_backbone:{
        start_lr: 1.5e-6,
        end_lr: 1.5e-7,
        batch_size: 32,
        steps: 5,000
    }
}
```

During the personalization phase, we keep all freeze parameters frozen, including the task vectors and pre-trained weights, and concentrate solely on training the personalization coefficients.

```
personalization_phase: {
    start_lr: 1.0e-2,
    end_lr: 1.0e-3,
    lr_schedule: Cosine annealing,
    batch_size: 32,
    steps: 500
}
```

Note that the presonalization phase only takes **500 iterations** due to the small number of trainable parameters.

Table 2: GIAA performance on the AVA database [16]

Method	PLCC	SROCC
NIMA [20]	0.636	0.612
Hosu *et al.* [6]	0.757	0.756
PA-IAA [13]	-	0.666
HLA-GCN [18]	0.687	0.665
MUSIQ-single [8]	0.731	0.719
MUSIQ [8]	0.738	0.726
VILA-R [9]	0.774	0.774
TANet [4]	0.765	0.758
TAVAR [12]	0.736	0.725
CSKD [22]	0.779	0.770
Ours (ViT-B/16)	0.780	0.781
Ours (ViT-L/14)	**0.804**	**0.808**

Table 3: GIAA performance on the TAD66K database [4]

Method	PLCC	SROCC
PAM [17]	0.440	0.422
NIMA [20]	0.405	0.390
HLA-GCN [18]	0.493	0.486
TANet [4]	**0.531**	**0.513**
Ours (ViT-B/16)	0.521	0.491
Ours (ViT-L/14)	0.523	0.492

Table 4: GIAA performance on the PARA database [23]

Method	PLCC	SROCC
PARA [23]	0.936	0.902
CSKD [22]	**0.951**	**0.926**
Ours (ViT-B/16)	0.945	0.921
Ours (ViT-L/14)	0.945	0.917

Table 5: IQA performance on the KonIQ-10K database [5]

Method	PLCC	SROCC
BRISQUE [15]	0.681	0.665
ILNIQE [25]	0.523	0.507
SFA [11]	0.872	0.856
DBCNN [26]	0.884	0.875
MetaIQA [27]	0.887	0.805
BIQA [19]	0.917	0.906
MUSIQ [8]	0.928	0.916
CLIP-IQA [21]	0.909	0.895
Ours (ViT-B/16)	**0.941**	**0.925**
Ours (ViT-L/14)	0.933	0.918

Table 6: IQA performance on the SPAQ database [3]

Method	PLCC	SROCC
BRISQUE [15]	0.817	0.809
ILNIQE [25]	0.721	0.713
DBCNN [26]	0.915	0.911
Fang *et al.* [3]	0.909	0.908
BIQA [19]	**0.928**	**0.925**
MUSIQ [8]	0.921	0.917
CLIP-IQA [21]	0.866	0.864
Ours (ViT-B/16)	0.882	0.914
Ours (ViT-L/14)	0.874	0.912

A.7 Performance on Each GIAA/IQA Database

In this section, we report the performance of the models trained for each image score regression task. While our unified architecture is not explicitly designed for specific tasks, such as image quality assessment (IQA) or general image aesthetic assessment (GIAA), the performance achieved for each task is comparable to existing approaches as demonstrated in Tables 2 to 6. This demonstrates the effectiveness of the unified architecture described in Section A.4. A high performance on each database indicates that the model has successfully learned the characteristics of the databases, including the distribution of scores and preference for certain types of images. Thus, we can assure that deriving task vectors from these models produces task vectors that accurately represents the characteristics of each database.

Furthermore, we observe that larger models do not always guarantee higher performance, aligning with some of the results reported in recent IQA studies [8, 21]. This underscores the importance of effectively harnessing the capabilities of pre-trained models, which is also the central idea of our approach in training the coefficients of task vectors.

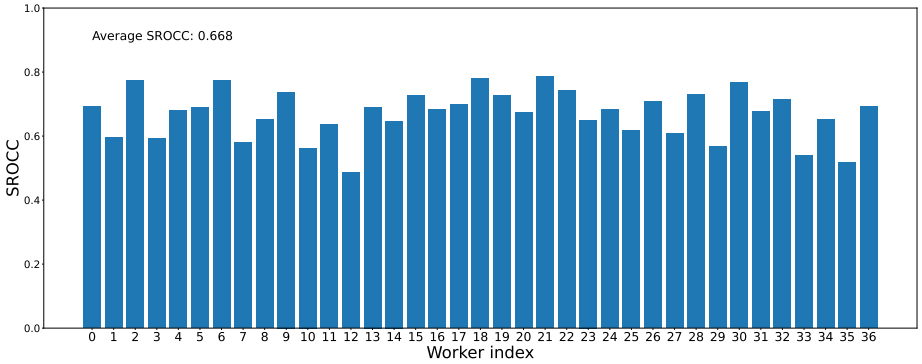

Fig. 2: Individual 10-shot personalization performance of the individuals in the Flickr-AES database [17].

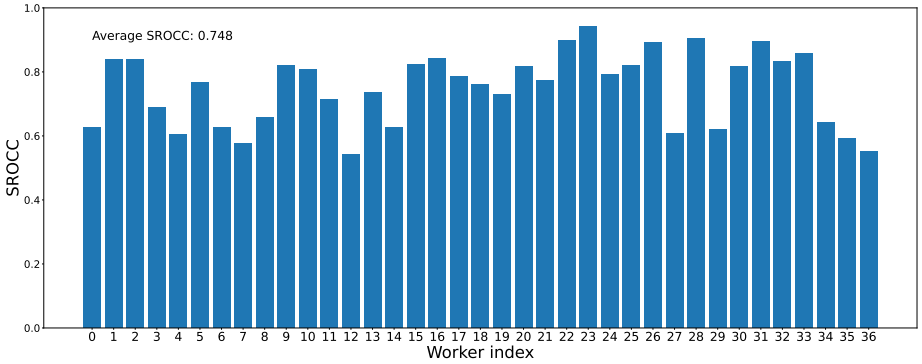

Fig. 3: Individual 100-shot personalization performance of the individuals in the Flickr-AES database [17].

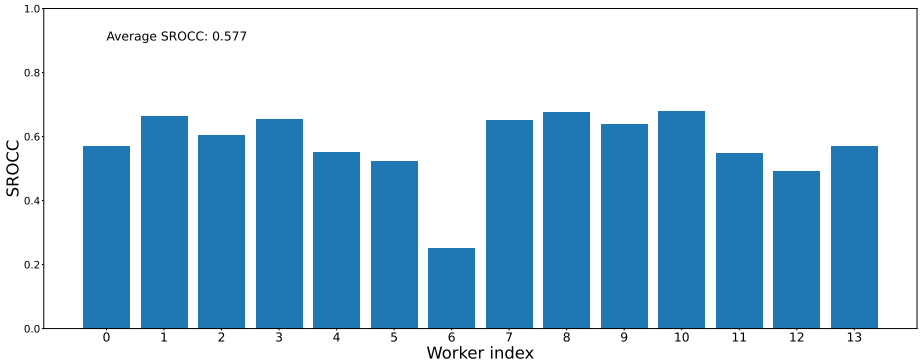

Fig. 4: Individual 10-shot personalization performance of the individuals in the REAL-CUR database [17]

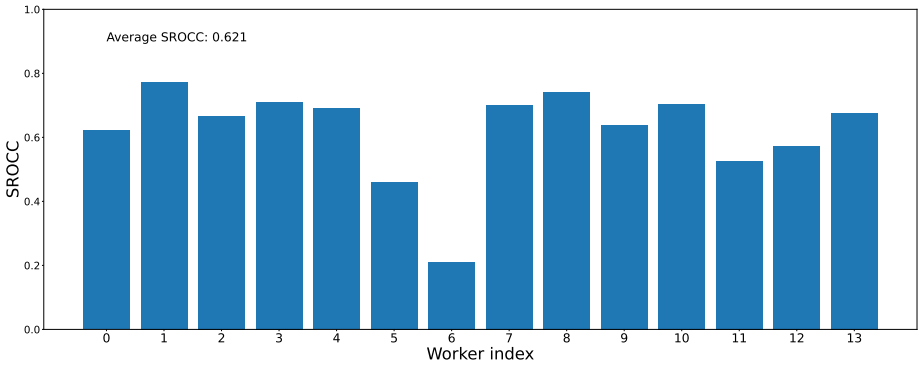

Fig. 5: Individual 100-shot personalization performance of the individuals in the REAL-CUR database [17]

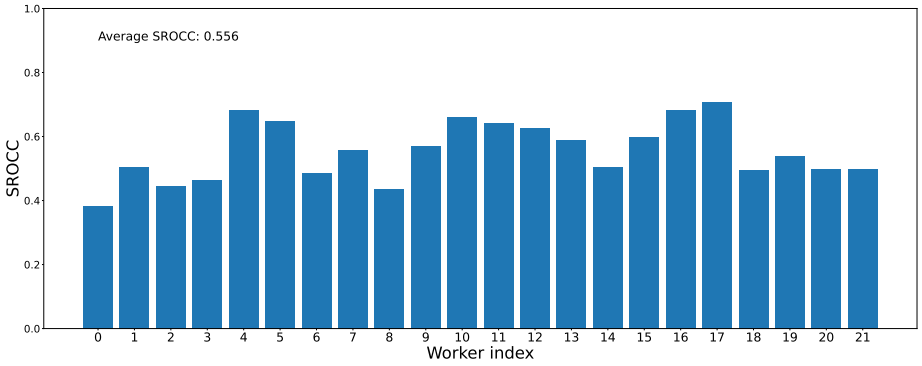

Fig. 6: Individual 10-shot personalization performance of the individuals in the AADB database [10]

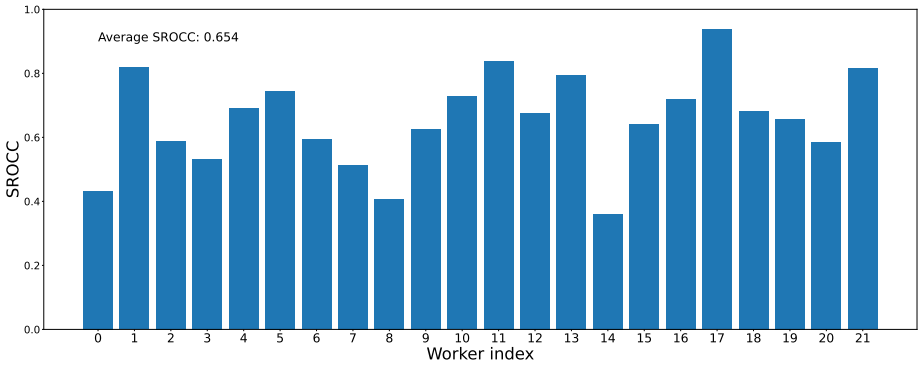

Fig. 7: Individual 100-shot personalization performance of the individuals in the AADB database [10]

References

1. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika (1952)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
3. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: CVPR (2020)
4. He, S., Zhang, Y., Xie, R., Jiang, D., Ming, A.: Rethinking image aesthetics assessment: Models, datasets and benchmarks. IJCAI (2022)
5. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing (2020)
6. Hosu, V., Goldlucke, B., Saupe, D.: Effective aesthetics prediction with multi-level spatially pooled features. In: CVPR (2019)
7. Ilharco, G., Ribeiro, M.T., Wortsman, M., Schmidt, L., Hajishirzi, H., Farhadi, A.: Editing models with task arithmetic. In: ICLR (2023)
8. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV (2021)
9. Ke, J., Ye, K., Yu, J., Wu, Y., Milanfar, P., Yang, F.: Vila: Learning image aesthetics from user comments with vision-language pretraining. In: CVPR (2023)
10. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: ECCV. Springer (2016)
11. Li, D., Jiang, T., Lin, W., Jiang, M.: Which has better visual quality: The clear blue sky or a blurry animal? IEEE Transactions on Multimedia (2018)
12. Li, L., Huang, Y., Wu, J., Yang, Y., Li, Y., Guo, Y., Shi, G.: Theme-aware visual attribute reasoning for image aesthetics assessment. IEEE Transactions on Circuits and Systems for Video Technology (2023)
13. Li, L., Zhu, H., Zhao, S., Ding, G., Lin, W.: Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. IEEE Transactions on Image Processing **29** (2020)
14. Li, Y., Yang, Y., Li, H., Chen, H., Xu, L., Li, L., Li, Y., Guo, Y.: Transductive aesthetic preference propagation for personalized image aesthetics assessment. In: ACM MM. ACM (2022)
15. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing (2012)
16. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: CVPR. IEEE (2012)
17. Ren, J., Shen, X., Lin, Z., Mech, R., Foran, D.J.: Personalized image aesthetics. In: ICCV (Oct 2017)
18. She, D., Lai, Y.K., Yi, G., Xu, K.: Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In: CVPR (2021)
19. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: CVPR (2020)
20. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE Transactions on Image Processing (2018)
21. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)

22. Xu, L., Xu, J., Yang, Y., Huang, Y., Xie, Y., Li, Y.: Clip brings better features to visual aesthetics learners. arXiv preprint arXiv:2307.15640 (2023)
23. Yang, Y., Xu, L., Li, L., Qie, N., Li, Y., Zhang, P., Guo, Y.: Personalized image aesthetics assessment with rich attributes. In: CVPR (2022)
24. Yang, Z., Li, L., Yang, Y., Li, Y., Lin, W.: Multi-level transitional contrast learning for personalized image aesthetics assessment. IEEE Transactions on Multimedia (2023)
25. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE Transactions on Image Processing (2015)
26. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology (2018)
27. Zhu, H., Li, L., Wu, J., Dong, W., Shi, G.: Metaiqa: Deep meta-learning for no-reference image quality assessment. In: CVPR (2020)
28. Zhu, H., Li, L., Wu, J., Zhao, S., Ding, G., Shi, G.: Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. IEEE Transactions on Cybernetics (2020)
29. Zhu, H., Zhou, Y., Li, L., Li, Y., Guo, Y.: Learning personalized image aesthetics from subjective and objective attributes. IEEE Transactions on Multimedia (2021)