Scaling Up Personalized Image Aesthetic Assessment via Task Vector Customization

Jooyeol Yun and Jaegul Choo

Korea Advanced Institute of Science and Technology (KAIST)
blizzard072@kaist.ac.kr jchoo@kaist.ac.kr

Abstract. The task of personalized image aesthetic assessment seeks to tailor aesthetic score prediction models to match individual preferences with just a few user-provided inputs. However, the scalability and generalization capabilities of current approaches are considerably restricted by their reliance on an expensive curated database. To overcome this long-standing scalability challenge, we present a unique approach that leverages readily available databases for general image aesthetic assessment and image quality assessment. Specifically, we view each database as a distinct image score regression task that exhibits varying degrees of personalization potential. By determining optimal combinations of task vectors, known to represent specific traits of each database, we successfully create personalized models for individuals. This approach of integrating multiple models allows us to harness a substantial amount of data. Our extensive experiments demonstrate the effectiveness of our approach in generalizing to previously unseen domains—a challenge previous approaches have struggled to achieve—making it highly applicable to real-world scenarios. Our novel approach significantly advances the field by offering scalable solutions for personalized aesthetic assessment and establishing high standards for future research. [1]

1 Introduction

Personalized image aesthetic assessment (PIAA) is an emerging field dedicated to developing aesthetic score prediction models that closely match an individual's aesthetic preferences based on a small set of user-provided samples. While they function as simple models that assign scores based on aesthetic quality, PIAA models enhance user experience in digital environments by personalizing tasks such as managing photo albums [17], curating web-scale databases like LAION-Aesthetics [33], and even guiding generative models such as Stable Diffusion [32] to create images that align with individual tastes [34].

In recent years, there has been a notable trend in PIAA [21, 41, 44] that involves the integration of meta-learning techniques, as personalization is inherently related to few-shot learning. Specifically, these approaches entail pretraining a model on a database consisting of image collections from diverse individuals. However, these strategies encounter significant challenges in terms of scalability and generalization to unseen domains, primarily due to the prohibitive cost associated with data collection.

[1] https://yeolj00.github.io/personal-projects/personalized-aesthetics/

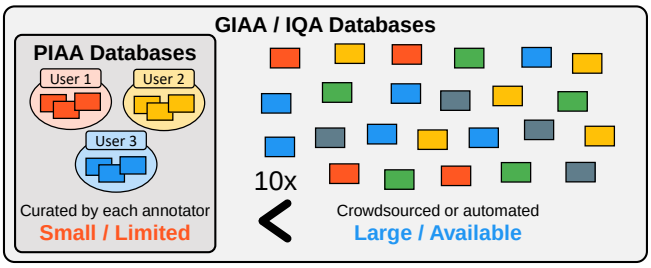

Fig. 1: Comparison between PIAA and other image regression databases. Our approach has the flexibility to leverage any image regression databases.

While meta-learning typically demands an extensive number of meta-training tasks, often exceeding 10,000, to achieve high performance [1], currently available PIAA databases [19, 31] offer a limited number of tasks fewer than 200. Even with the recently collected PIAA database [40], which contains rich attributes from each annotator, the number of training tasks remains below 500. In the context of PIAA, a single meta-training task corresponds to learning the preferences of a single individual. Scaling such a database would require obtaining a vast number of images and corresponding personalized aesthetic scores from numerous individuals, which poses severe limitations in the scalability and generalization of existing approaches.

In this paper, we unveil a data-rich approach for personalizing aesthetic assessment models by leveraging existing general image aesthetic assessment (GIAA) and image quality assessment (IQA) databases. In contrast to prior work [20, 21, 41, 44, 45], our approach is not limited to training on databases that track individual annotators, as depicted in Figure 1. Instead, we have the flexibility to utilize multiple image score regression databases, which are readily available. By employing broader resources, we overcome the scalability and generalization challenges that have *long been unaddressed* in this field.

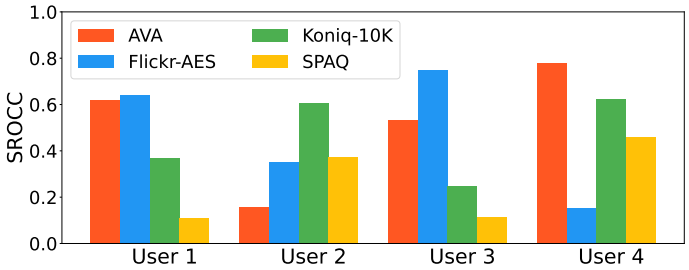

Fig. 2: Zero-shot personalization performance for four users in the Flickr-AES database [31]. Each colored bar refers to the personalization performance of a model trained on a specific database.

The key idea is to view each GIAA and IQA data as a unique image score regression task. This perspective arises from the distinct characteristics and tendencies of each database due to different data collection procedures and the population from which they originate. As visualized in Figure 2, image score regression models trained on four different GIAA [27, 31] and IQA [6, 12] databases

exhibit varying levels of personalization for users within the PIAA database. For instance, User 3's aesthetic preferences closely align with the score regression model trained on the Flickr-AES [31] while diverging from the model trained on the SPAQ [6] database. In other words, even without any personalized training, we find that general image assessment models inherently possess different degrees of personalization capabilities.

In light of these insights, we extend an emerging approach known as task arithmetic [14], which was originally devised for developing multi-task models. This method involves directly adding or subtracting the weights of models fine-tuned on specific tasks. In our context, task vectors from diverse image assessment models encapsulate distinct features and behaviors related to specific tasks, such as recognizing different aspects of image quality or aesthetics. By carefully combining task vectors, we explore the potential to selectively amplify or refine the capabilities of a pre-trained model.

However, there has been limited research on determining the optimal adjustments necessary to achieve a certain behavior. Thus, we present a personalization approach by introducing *trainable coefficients* for each task vector derived from models trained on distinct GIAA and IQA databases, thereby determining the weights for combining these vectors. Once the coefficients are trained with user-provided inputs, the weighted sum of task vectors creates a model tailored to the user's aesthetic preferences. Since task vectors obtained from large databases already capture preferences across various themes and contents, we find that training only the coefficients is sufficient for precise personalization, making our approach *highly parameter-efficient*. To the best of our knowledge, our approach marks the first to customize task vectors for specific model behaviors that were previously unattainable.

By carefully customizing task vectors, we incorporate large databases into the training process, granting our approach a strong generalization capability for unseen domains, a scenario often encountered in real-world applications. Furthermore, we find that our approach is also practical for learning a user's aesthetic preference with only a few samples, as our approach of merging task vectors guides the model to make *well-informed updates* and prevents overfitting. We provide extensive experiments demonstrating the efficacy of our approach for personalization in diverse scenarios.

Our contributions are threefold:

- We introduce a novel data-rich approach for PIAA that tackles the long-standing scalability challenge, eliminating the dependency on expensive, manually-curated databases that have hindered progress in this field.
- Demonstrating exceptional effectiveness in cross-database evaluations, our method showcases robust generalization capabilities that surpass those of existing approaches, setting high standards for future work.
- By learning optimal combinations of task vectors, we present a parameter-efficient approach for precise personalization, leveraging the comprehensive information embedded in these vectors.

2 Related Work

2.1 Personalized Image Aesthetic Assessment

Personalized image aesthetic assessment (PIAA) [13,20,21,24,25,35,39–41,44,45] aims to learn the personal aesthetic preference of individual users based on the image-score pairs from one's image collection. Since collecting a large number of samples from a single user is impractical, PIAA is closely linked to few-shot learning methods. One line of work [20,40,45] incorporates additional attributes (*e.g.*, personality traits, age, and gender) to facilitate user-specific aesthetic predictions. These approaches rely on the correlation between user attributes and aesthetic preferences, which may not always be clearly defined. Consequently, many high-performing approaches for PIAA [21, 41, 44] leverage meta-learning techniques to train models that can be easily fine-tuned to accommodate the aesthetic preferences of new users with only a limited number of samples.

Nevertheless, applying meta-learning techniques to PIAA encounters scalability challenges owing to the limited number of tasks available during the meta-learning phase. In traditional meta-learning, achieving a high performance typically requires a substantial number of meta-training tasks [1], often exceeding 10,000. In PIAA databases [19,31], the number of training tasks ranges from 100 to fewer than 500, which can impose limitations on the performance. Expanding the task pool in PIAA is prohibitively expensive since each task corresponds to the image collection of a single user, posing significant scalability challenges that has long been unaddressed. We tackle this issue in previous PIAA approaches and present a scalable fine-tuning method that leverages existing general image aesthetic assessment (GIAA) and image quality assessment (IQA) databases.

2.2 Task Vectors and Task Arithmetic

Ilharco *et al.* [14] presented a new paradigm for creating multi-task models, utilizing the parameters of multiple fine-tuned models. They define the *task vector* τ as the difference between the fine-tuned and pre-trained weights,

$$\tau = \theta_{\text{ft}} - \theta_{\text{pre}},$$

where θ_{ft} and θ_{pre} represent the weights of the fine-tuned and pre-trained model, respectively. They demonstrate that adding and subtracting task vectors to pre-trained models can encourage or suppress the ability to perform specific tasks. For example, negating the task vector derived from a language model fine-tuned on toxic data reduces the frequency of generating toxic text. However, it remains unclear what the optimal adjustments are for achieving a specific behavior.

In our paper, we investigate methods to determine necessary adjustments to achieve a certain behavior defined by a few user-provided samples. By introducing learnable coefficients to task vectors, we effectively customize models to closely match individual preferences. We also demonstrate that finding the optimal coefficients is possible with only a few samples, as the task vectors encompass comprehensive knowledge derived from large databases.

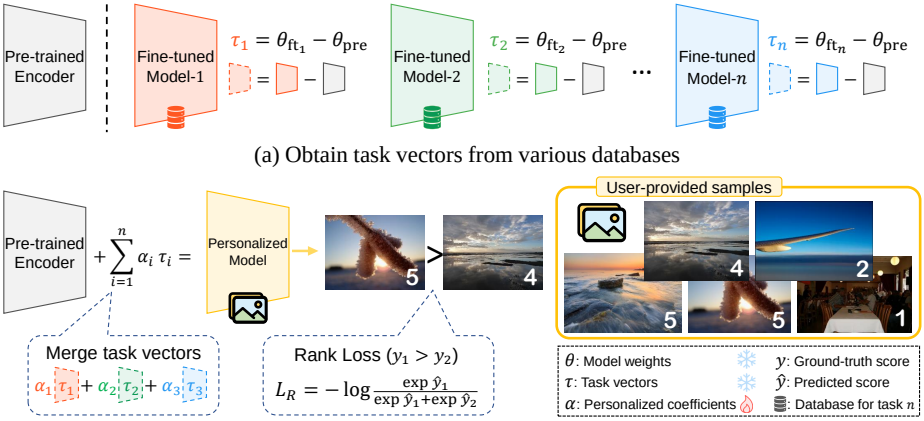

Fig. 3: (a) Multiple models are fine-tuned from a single pre-trained model on different tasks to obtain task vectors. For simplicity, we omit the discussion of layers. (b) Given a small number of user-provided samples, the coefficients corresponding to each task vector are optimized with the rank loss.

3 Method

3.1 Obtaining Layer-wise Task Vectors from Various Databases

To obtain task vectors, we train multiple models with the same architecture on various tasks (*i.e.*, different databases), as illustrated in Figure 3 (a). Specifically, we fine-tune the weights from a pre-trained backbone, adjusting the parameters from θ_{pre} to θ_{ft_i} for each task i. Moreover, while the specific roles of model layers in capturing unique features are inherently complex, their distinct contributions are undeniable [3,29]. Thus, we compute task vectors across *all layers*, enhancing the flexibility of our approach for personalization (See appendix Section A.2).

Accordingly, we expand the concept of original task vectors to include layer-wise differentiation by,

$$\tau_i^l = \theta_{\text{ft}_i}^l - \theta_{\text{pre}}^l, \tag{1}$$

where τ_i^l is the task vector for the l-th layer of the i-th task. Each vector comprehensively integrates patterns learned from large databases, acting as basis vectors for robust and precise personalization.

It is important to note that this process is conducted *only once* and is not required again for personalizing to individual users, thereby not contributing to the count of learnable parameters for personalization.

3.2 Learning Optimal Combinations for Personalization

Given the layer-wise task vectors, our goal is to find a combination of these vectors that accurately reflects a user's aesthetic preferences. To achieve this, we introduce learnable coefficients α_i^l for each task vector τ_i^l. Once the coefficients

are trained, the weights of the personalized model for each layer are derived as,

$$\theta_{\mathrm{p}}^{l} = \theta_{\mathrm{pre}}^{l} + \sum_{i=1}^{n} \alpha_i^l \tau_i^l \quad \text{for } l = 1, 2, \ldots, L, \tag{2}$$

where n is the number of tasks, and L is the number of layers in the model. We keep all parameters other than the $n \times L$ coefficients *frozen*. This is essential since the task vectors contain comprehensive insights into aesthetic preferences, enabling the model to make *well-informed updates*. By harnessing the rich information embedded in these vectors, our method efficiently learns complex aesthetic preferences from merely a few user-provided samples.

Our objective function is designed to penalize the misordering of aesthetic scores. Specifically, we adopt the Bradley-Terry model [2], which estimates the probability of pairwise comparison between two samples. We frame the prediction of the two scores as a binary classification problem, where the objective is to predict which sample has the higher score. The coefficients of each task vector are trained to minimize the following loss function:

$$L_{\mathrm{R}} = -\log \frac{\exp \hat{y}_1}{\exp \hat{y}_1 + \exp \hat{y}_2} \quad (y_1 \geq y_2), \tag{3}$$

where $\hat{y}_1$ and $\hat{y}_2$ represent the aesthetic score predictions for images x_1 and x_2, whose ground-truth scores are y_1 and y_2. We prefer this loss function over the Earth Mover Distance (EMD) or the MSE loss, due to its sample efficiency and fast convergence (See appendix Section A.3).

3.3 Adaptive Coefficient Initialization

We observe that appropriately initializing the coefficients plays a crucial role in determining the coefficients that best align with the user's aesthetic preferences. When a high coefficient is assigned to a specific task vector, it indicates a strong correlation between the user's aesthetic preference and that particular task. Thus, we adaptively initialize the coefficients based on the zero-shot personalization performances of each image score regression model associated with the task vector.

To do this, we calculate the Spearman's rank-order correlation coefficient (SROCC) [28] between the user-provided samples and the scores predicted by each image score regression model. We then apply the softmax function to initialize the coefficients α as,

$$\alpha_i^l = \frac{e^{\mathrm{SROCC}(S, \theta_{\mathrm{ft}_i})}}{\sum_{j=1}^{n} e^{\mathrm{SROCC}(S, \theta_{\mathrm{ft}_j})}} \quad \text{for } l = 1, 2, \ldots, L, \tag{4}$$

where $\mathrm{SROCC}(S, \theta_{\mathrm{ft}_i})$ indicates the SROCC calculated on the user-provided set S using the model with parameters θ_{ft_i}. Note that we initialize α values for a single task i with the same value for all layers.

Table 1: Description of the databases included in our experiments. The AADB and REAL-CUR databases are exclusively reserved as a test database.

Database	Regression task	Annotator tag	Number of images
KonIQ-10K [12]	IQA	✗	10,073
SPAQ [6]	IQA	✗	11,125
AVA [27]	GIAA	✗	255,500
TAD66K [11]	GIAA	✗	67,125
Flickr-AES [31]	GIAA/PIAA	✓	40,000
PARA [40]	GIAA/PIAA	✓	31,220
AADB [19]	GIAA/PIAA	✓	9,958
REAL-CUR [31]	PIAA	✓	2,871

4 Experiments

4.1 Database Overview

The key advantage of our approach is the flexibility to utilize multiple image score regression databases freely, even those without annotator tags. In our experiments, we fine-tune models using six different databases in Table 1, reserving the AADB [19] and REAL-CUR [31] database exclusively for testing. While the Flickr-AES [31] and PARA [40] databases do contain annotator tags, our approach does not require distinguishing between individuals. Therefore, we treat these two databases as GIAA databases. In total, we effectively leverage **415,043** samples from six databases, marking the most extensive utilization of samples in PIAA within our knowledge. Detailed descriptions and performance on these databases are provided in the appendix (See Section A.5). However, our approach is not limited to training solely on these six databases, having the scalability to include additional IQA and GIAA databases [4, 8, 10, 16, 42].

For testing, we use the following three most widely-used PIAA test sets.
Flickr-AES [31] database consists of 40,000 images collected from 210 users. For testing purposes, we utilized a test set containing 4,737 images rated by 37 users, with each user contributing image collections ranging from 105 to 171.
AADB [19] database consists of 9,958 images from 190 users. We selected a test set from 22 users, each of which contributed individual image collections containing 110 to 190 images.
REAL-CUR [31] database consists of personal image collections from 14 users, totaling 2,871 images. Each user's collection ranges from 197 to 222 images.

4.2 Implementation Details

To obtain task vectors from the models fine-tuned on each of the six databases, it is essential to start from an identical pre-trained model. Recent approaches in IQA [18], GIAA [38], and PIAA [13] emphasize the importance of pre-training on large databases for training powerful image score regression models. Thus, we choose the publicly available OpenCLIP models ViT-B/16 and ViT-L/14 [15] as the pre-trained model. Full details regarding the architectural design and

Table 2: Cross-database evaluation on the REAL-CUR [31] database. For the baseline approaches, we report the best cross-domain performances.

Method	SROCC	
	10-shot	100-shot
PA-IAA [20]	0.443±0.004	0.562±0.013
BLG-PIAA [44]	0.448±0.007	0.578±0.015
PIAA-SOA [45]	0.487±0.003	0.589±0.015
TAPP-PIAA [21]	-	0.580
MTCL [41]	0.495±0.007	0.599±0.012
Ours	**0.577±0.005**	**0.621±0.007**

training hyperparameters are provided in the appendix (See Section A.5 and A.6). We will also release the codes upon acceptance.

If not stated otherwise, we use $n = 6$ fine-tuned models to derive the task vectors. Each model comprises $L = 296$ layers when using the ViT-L/14 model and $L = 152$ layers when using the ViT-B/16 model. While the number of 'layers' usually refers to the number of transformer blocks, we include all computational elements within these blocks, such as linear layers and normalization layers. The total number of learnable parameters, $n \times L = 1776$ and $n \times L = 912$, is significantly lower compared to a full fine-tuning approach, which would require training over 23 million parameters even for small ResNet-50 models [9]. We *do not* consider the layer-wise task vector itself as a learnable parameter, as obtaining these vectors is not repeated for each individual.

4.3 Evaluation Metric for PIAA

PIAA approaches are often evaluated through the Spearman's rank-order correlation coefficient [28] (SROCC), which is the correlation coefficient between the predicted and ground-truth rank variables. Specifically, given the ground-truth and predicted scores y and $\hat{y}$, the SROCC is calculated as,

$$\text{SROCC} = 1 - \frac{6\sum_{i=1}^{N}(r_i - \hat{r}_i)^2}{N(N^2-1)}, \tag{5}$$

where r_i and $\hat{r}_i$ represent the rank of the i-th sample within the ground-truth and predicted scores, respectively, and N denotes the number of samples.

4.4 Cross-database Evaluation

PIAA evaluation can be categorized based on whether training and testing occur within the same database (*i.e.*, intra-database evaluation) or involve different databases (*i.e.*, cross-database evaluation). In most real-world scenarios involving personal image collections, user-provided samples often diverge from the statistics or tendencies of the training database. Therefore, cross-database evaluation becomes pivotal in assessing the generalization capabilities of personalization approaches, which is an aspect that remains relatively *underdeveloped* in comparison to intra-database evaluations.

Table 3: Cross-database evaluation on the AADB [19] database. Results marked with * indicate intra-database performance (*i.e.*, directly trained on the AADB database).

Method	SROCC	
	10-shot	100-shot
PA-IAA [20]	0.469±0.002	0.524±0.006
BLG-PIAA [44]	0.486±0.004	0.536±0.006
PIAA-SOA [45]	0.509±0.003	0.557±0.007
TAPP-PIAA [21]	-	0.540
MTCL [41]	0.533±0.004	0.572±0.007
*TAPP-PIAA [21]	*0.534±0.004	*0.612±0.007
*MTCL [41]	*0.540±0.005	*0.622±0.007
Ours	**0.556±0.004**	**0.654±0.007**

For cross-database evaluation, we employ the AADB [19] and REAL-CUR [31] databases as the test databases, neither of which were among the six databases from which we derived our task vectors. Following established protocols [31], we randomly select K samples for each individual to form the training set, also referred to as the support set, while the remaining images form the test set. In our experiments, we set K as either 10-shot or 100-shot. Given the potential variability inherent in few-shot learning depending on the chosen K support set, we conduct 10 independent trials for each user and report the average SROCC with the standard deviation.

As illustrated in Table 2 and Table 3, our method significantly surpasses existing PIAA models in cross-database performance for both the 10-shot and 100-shot protocols. Remarkably, *it even exceeds the performance of models trained excplicitly on the AADB database*, as compared in Table 3. Such results underscore the exceptional generalization capabilities of our approach, in contrast to previous methods that are constrained by their inability to leverage GIAA and IQA databases. This superior performance is primarily due to the scalability of our approach, which effectively harnesses an extensive amount of 415,043 samples across multiple databases, each contributing unique insights and tendencies.

4.5 Intra-database Evaluation

For intra-database evaluation, we test our approach for the 37 individuals in the Flickr-AES [31] database, which is the most widely used PIAA database. Since the Flickr-AES [31] database is one of the six databases used to derive the task vectors, this experiment falls under the category of intra-database evaluation. However, we do not use the annotator information in the Flickr-AES.

In Table 4, we report the average SROCC for each approach along with the standard deviation across 10 independent trials. These experiments underscore the effectiveness of training personalized task vectors through the use of multiple fine-tuned models. Our approach consistently outperforms existing methods that rely on meta-learning techniques or incorporate additional user attributes (*e.g.*, personality traits) to enhance personalized score predictions.

Table 4: 10-shot and 100-shot personalization results on the Flickr-AES database [31], with the average SROCC of 37 users and standard deviation across 10 trials.

Method	SROCC	
	10-shot	100-shot
PAM (attribute only) [31]	0.511±0.004	0.516±0.003
PAM (content only) [31]	0.512±0.002	0.516±0.010
PAM [31]	0.513±0.003	0.524±0.007
PASS [35]	0.516±0.003	0.521±0.007
USAR-PPR [25]	0.521±0.002	0.544±0.007
USAR-PAD [25]	0.520±0.003	0.537±0.003
USAR-PPR&PAD [25]	0.525±0.004	0.552±0.015
PA-IAA [20]	0.543±0.003	0.639±0.011
BLG-PIAA [44]	0.561±0.005	0.669±0.013
UG-PIAA [24]	0.559±0.002	0.660±0.013
PIAA-SOA [45]	0.618±0.006	0.691±0.015
TAPP-PIAA [21]	0.591±0.007	0.685±0.012
IM-PIAA [13]	0.620	0.708
MTCL [21]	0.667±0.005	0.737±0.014
Ours	**0.668±0.004**	**0.748±0.012**

5 Analysis and Ablation Studies

We perform an in-depth analysis and ablation studies on each component of our approach, including model size, scalability, and coefficient initialization.

5.1 Task Similarities Across Databases

To gain insights into the derived task vectors, we visualize the similarity matrix of the task vectors in Figure 4. For simplicity, we aggregate the L task vectors derived from each layer of a single fine-tuned model as a single task vector by concatenating them into one vector. In line with the findings provided by Ilharco *et al*. [14], task vectors are typically orthogonal, even though the databases are collected with similar intentions (*i.e.*, GIAA and IQA). This observation highlights that treating each database as a unique task remains a valid choice, owing to variations in their data collection methodologies and the diverse nature of the populations from which they are sourced.

5.2 Architecture Choices and Variations

Selecting the architecture. Weight mixing techniques, including task vectors [14, 36, 37], have been recognized for their effectiveness when fine-tuning models pre-trained on large databases, such as the LAION-5B [33] or DataComp [7]. Vision Transformers (ViT) [5] is a suitable architecture for accommodating web-scale databases, in contrast to ResNets [9], which often struggle to achieve high performance despite training on large databases. Therefore, given

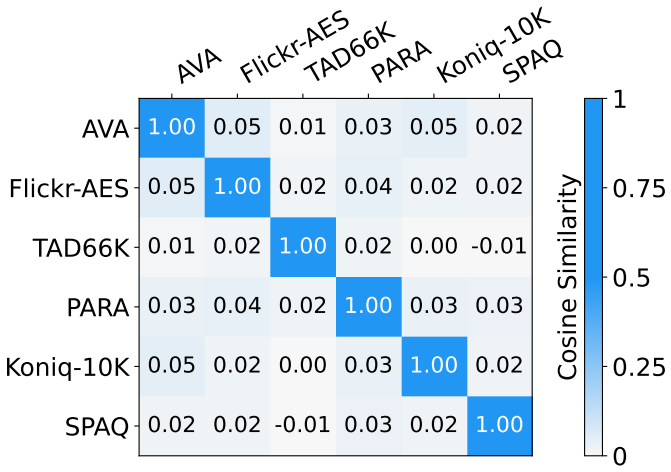

Fig. 4: Similarity matrix between task vectors derived from each database. The similarities are measured with the cosine similarity.

Table 5: Performance and inference speed of the ViT-B/16 model and the Best-fit FT method (*i.e.*, directly fine-tuning ViTs) on the REAL-CUR [31] database.

Method	SROCC		FPS
	10-shot	100-shot	
Best-fit FT	0.303±0.008	0.394±0.008	94.41
PIAA-SOA [45]	0.487±0.003	0.589±0.015	-
MTCL [41]	0.495±0.007	0.599±0.012	359.05
Ours (ViT-B/16)	0.562±0.004	0.607±0.007	**417.95**
Ours	**0.577±0.005**	**0.621±0.007**	94.41

its scalability and performance advantages, we choose the publicly available pre-trained ViT as our architecture for our experiments. Regardless of the chosen architecture, our parameter-efficient approach only trains a minimal set of coefficients corresponding to each layer-wise task vector.

Scaling down for efficient computation. While our primary experiments focus on combining task vectors derived from ViT-L/14 models, our approach is adaptable to smaller models as well. As shown in Table 5, our approach yields strong results even on smaller models like ViT-B/16, which are well-suited for deployment on commercial computers. Although fast inference was not our primary goal, we achieve real-time inference speeds ($\geq$ 30 FPS) on the RTX 3090, as our weight mixing technique combines multiple models into a single ViT model, eliminating the need to pass through multiple models. Acceleration of ViTs [22, 23, 43] beyond our implementation is also actively supported in recent frameworks [26] and communities [30], which ensures the practicality of our approach across a wide range of computing platforms such as mobile devices.

Vision Transformers without task vectors. Our experimental results in Table 5 demonstrate that the performance gains are not solely attributed to

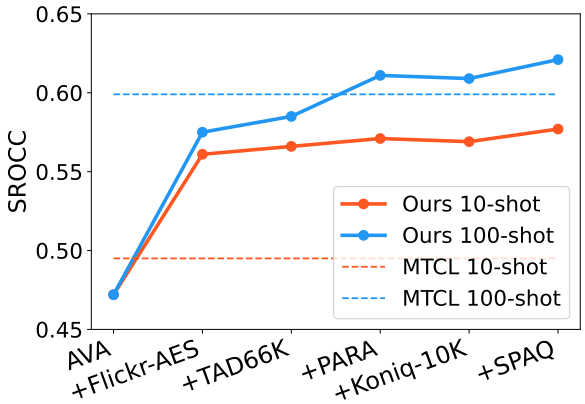

Fig. 5: Performance gains with respect to the number of databases used. The dotted lines indicate the previous state-of-the-art performance.

the inherent strengths of the ViT architecture. When directly fine-tuning the aesthetic assessment model that best matches the user's preference (*i.e.*, Best-fit FT), we observe distinctly low personalization performances despite using a large pre-trained model. This highlights the pivotal role of keeping the layer-wise task vectors frozen during personalization, as altering the vectors would indicate *discarding* the rich preferences learned from a large database.

5.3 Scalability of Our Approach

How does each database contribute? We assess the scalability of our approach by varying the number of task vectors. Specifically, we evaluate our approach on the REAL-CUR database [31] by increasing the number of databases used. Starting from the GIAA model trained on the AVA [27] database, we accumulate the databases in the following order: Flicker-AES [31], TAD66K [11], PARA [40], KonIQ-10K [12], and SPAQ [6]. As demonstrated in Figure 5, both 10-shot and 100-shot performance significantly improves as the number of task vectors increases. This indicates that our scalable approach benefits from the expansion of task vectors, which allows us to capture a broader range of image characteristics and nuances.

How can we increase the number of task vectors? The advantage of our approach can be amplified with the availability of additional image score regression databases. For example, scaling up the number of task vectors can be further pursued by incorporating readily available IQA [4,8,42] and GIAA [10,16] databases not included in our current study. Additionally, partitioning large databases, such as AVA [27] and PaQ-2-PiQ [42], can be an alternative method for generating additional task vectors.

These strategies for obtaining a diverse set of task vectors may lead to performance gains in the 10-shot evaluation, which is considered highly challenging due to the limited number of user-provided samples. However, we leave this as future work, given that our current *work already utilizes the largest number of*

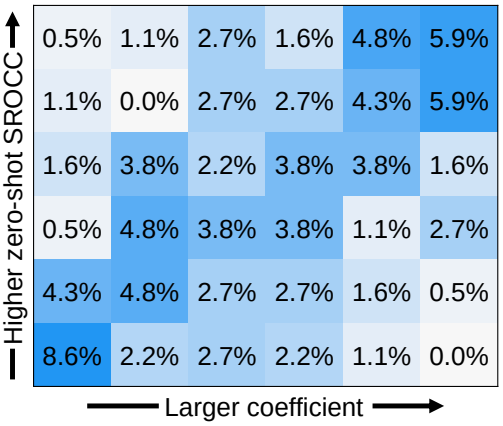

Fig. 6: 2D-histogram of coefficients and zero-shot SROCC pairs. The size of the coefficients are correlated with the relevance of the task.

samples for PIAA, 10 times larger than the most widely used and highly curated Flickr-AES database [31].

5.4 Task Relevance and Coefficient Initialization

We conduct an analysis to better understand the distinguishing characteristics of coefficients, specifically, what sets apart those with high values from those with low values. For a straightforward examination, we calculate the average coefficient for each task (*i.e.*, database), averaged across all the layers of the model, which results in six average coefficients for each user ($n = 6$). Subsequently, we assess the six zero-shot SROCC scores associated with the average coefficients and plot the histogram of the coefficient-SROCC pairs from all test users in Figure 6. We observe that coefficients for tasks with a high zero-shot SROCC are trained to have high values, while those for less-relevant tasks (*i.e.*, low zero-shot SROCC) exhibit low values. This analysis justifies our adaptive initialization of the coefficients, which assigns large values to coefficients associated with closely related tasks.

We investigate two alternative initialization strategies: a uniform initialization and a best-fit initialization. Both cases can be viewed as an extreme case of the adaptive initialization, which can be written in a generalized form:

$$\alpha_i^l = \frac{e^{\mathrm{SROCC}(S,\theta_{\mathrm{ft}_i})/T}}{\sum_{j=1}^{n} e^{\mathrm{SROCC}(S,\theta_{\mathrm{ft}_j})/T}} \quad \text{for } l = 1, 2, \ldots, L, \tag{6}$$

where T is the temperature for the softmax. The uniform initialization corresponds to the case where the $T \to \infty$, while the best-fit initialization is when $T \to 0$. As demonstrated in Table 6, the adaptive initialization outperforms other initialization strategies. Nevertheless, we consider further exploration of the initialization strategies as a potential avenue for future research.

Table 6: Comparison between three different coefficient initialization strategies tested on the REAL-CUR [31] database.

Method	SROCC	
	10-shot	100-shot
Uniform init. ($T \to \infty$)	0.520±0.002	0.602±0.002
Best-fit init. ($T \to 0$)	0.464±0.012	0.596±0.002
Adaptive init. ($T = 1$)	**0.577±0.005**	**0.621±0.007**

6 Limitation

We have observed that when the zero-shot SROCC for all six image regression models are low for an individual (*i.e.*, SROCC $\approx$ 0), our approach may struggle to identify an optimal combination of task vectors for that user. This phenomenon suggests that when there is a limited correlation between the user's preferences and the predictions of these models, the task vectors derived from them may be insufficient for finding an appropriate personalized task vector.

However, we find this case to be extremely rare (See appendix Figure 2-7.), with only *1 among 73* cases having a low SROCC under 0.2, demonstrating the broad coverage of our approach. Also, increasing the number of task vectors will effectively mitigate this issue, leveraging the inherent scalability of our approach, which can accommodate a growing number of task vectors to improve personalization even in challenging scenarios.

7 Conclusion

We address the scalability issue in PIAA, a critical yet previously unaddressed aspect necessary for enhancing the generalization performance. Unlike previous studies limited to training on PIAA databases curated by individual annotators, our approach enables the flexible use of multiple image score regression databases, thereby providing greater scalability and generaliation capabilities.

We demonstrate that GIAA and IQA databases exhibit distinct personalization potentials, which have motivated us to combine GIAA and IQA models to achieve desired behaviors. By introducing learnable parameters for optimal model combinations, we enable the precise personalization of models to individual users' aesthetic preferences. Our approach significantly outperforms existing approaches, demonstrating exceptional efficacy in generalizing to unseen domains which is a critical requirement for real-world applications. This work opens up new avenues for personalized image aesthetic assessment, offering valuable insights and practical solutions for this challenging task.

Acknowledgement

This work was supported by the Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913), and the Institute of Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2021-II212068, Artificial Intelligence Innovation Hub).

References

1. Al-Shedivat, M., Li, L., Xing, E., Talwalkar, A.: On data efficiency of meta-learning. In: International Conference on Artificial Intelligence and Statistics. PMLR (2021)
2. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika (1952)
3. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: CVPR (2021)
4. Ciancio, A., da Silva, E.A., Said, A., Samadani, R., Obrador, P., et al.: No-reference blur assessment of digital pictures based on multifeature classifiers. IEEE Transactions on Image Processing (2010)
5. 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)
6. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: CVPR (2020)
7. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108 (2023)
8. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. IEEE Transactions on Image Processing (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. He, S., Ming, A., Li, Y., Sun, J., Zheng, S., Ma, H.: Thinking image color aesthetics assessment: Models, datasets and benchmarks. In: ICCV (2023)
11. He, S., Zhang, Y., Xie, R., Jiang, D., Ming, A.: Rethinking image aesthetics assessment: Models, datasets and benchmarks. IJCAI (2022)
12. 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)
13. Hou, J., Lin, W., Yue, G., Liu, W., Zhao, B.: Interaction-matrix based personalized image aesthetics assessment. IEEE Transactions on Multimedia (2022)
14. Ilharco, G., Ribeiro, M.T., Wortsman, M., Schmidt, L., Hajishirzi, H., Farhadi, A.: Editing models with task arithmetic. In: ICLR (2023)
15. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773
16. Jin, X., Wu, L., Zhao, G., Li, X., Zhang, X., Ge, S., Zou, D., Zhou, B., Zhou, X.: Aesthetic attributes assessment of images. In: ACM MM (2019)
17. Karlsson, K., Jiang, W., Zhang, D.Q.: Mobile photo album management with multiscale timeline. In: ACM MM (2014)
18. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV (2021)
19. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: ECCV. Springer (2016)
20. 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)

21. 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)
22. Lin, Y., Zhang, T., Sun, P., Li, Z., Zhou, S.: Fq-vit: Post-training quantization for fully quantized vision transformer. IJCAI (2022)
23. Liu, Z., Wang, Y., Han, K., Zhang, W., Ma, S., Gao, W.: Post-training quantization for vision transformer. NeurIPS (2021)
24. Lv, P., Fan, J., Nie, X., Dong, W., Jiang, X., Zhou, B., Xu, M., Xu, C.: User-guided personalized image aesthetic assessment based on deep reinforcement learning. IEEE Transactions on Multimedia (2021)
25. Lv, P., Wang, M., Xu, Y., Peng, Z., Sun, J., Su, S., Zhou, B., Xu, M.: Usar: An interactive user-specific aesthetic ranking framework for images. In: ACM MM. pp. 1328–1336 (2018)
26. Microsoft: Deepspeed (2023), https://www.deepspeed.ai/
27. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: CVPR. IEEE (2012)
28. Myers, J.L., Well, A.D., Lorch Jr, R.F.: Research design and statistical analysis. Routledge (2013)
29. Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F., Yang, M.H.: Intriguing properties of vision transformers. NeurIPS (2021)
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019)
31. Ren, J., Shen, X., Lin, Z., Mech, R., Foran, D.J.: Personalized image aesthetics. In: ICCV (Oct 2017)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. 2022 ieee. In: CVPR (2021)
33. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. NeurIPS (2022)
34. Wallace, B., Gokul, A., Ermon, S., Naik, N.: End-to-end diffusion latent optimization improves classifier guidance. ICCV (2023)
35. Wang, G., Yan, J., Qin, Z.: Collaborative and attentive learning for personalized image aesthetic assessment. In: IJCAI. pp. 957–963 (2018)
36. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: ICML. PMLR (2022)
37. Wortsman, M., Ilharco, G., Li, M., Kim, J.W., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. 2022 ieee. In: CVPR (2021)
38. 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)
39. Yan, X., Shao, F., Chen, H., Jiang, Q.: Hybrid cnn-transformer based meta-learning approach for personalized image aesthetics assessment. Journal of Visual Communication and Image Representation **98** (2024)
40. Yang, Y., Xu, L., Li, L., Qie, N., Li, Y., Zhang, P., Guo, Y.: Personalized image aesthetics assessment with rich attributes. In: CVPR (2022)
41. 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)

42. Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In: CVPR (2020)
43. Yuan, Z., Xue, C., Chen, Y., Wu, Q., Sun, G.: Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In: ECCV. Springer (2022)
44. 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)
45. Zhu, H., Zhou, Y., Li, L., Li, Y., Guo, Y.: Learning personalized image aesthetics from subjective and objective attributes. IEEE Transactions on Multimedia (2021)