On the Role of Individual Differences in Current Approaches to Computational Image Aesthetics

Li-Wei Chen[1,2]
li-wei.chen@kuleuven.be
Ombretta Strafforello[1]
ombretta.strafforello@kuleuven.be
Anne-Sofie Maerten[1]
annesofie.maerten@kuleuven.be
Tinne Tuytelaars[2]
Tinne.Tuytelaars@esat.kuleuven.be
Johan Wagemans[1]
johan.wagemans@kuleuven.be

[1] Brain and Cognition, KU Leuven, Belgium
[2] Department of Electrical Engineering (ESAT), KU Leuven, Belgium, KU Leuven, Belgium

Abstract

Image aesthetic assessment (IAA) evaluates image aesthetics, a task complicated by image diversity and user subjectivity. Current approaches address this in two stages: Generic IAA (GIAA) models estimate mean aesthetic scores, while Personal IAA (PIAA) models adapt GIAA using transfer learning to incorporate user subjectivity. However, a theoretical understanding of transfer learning between GIAA and PIAA, particularly concerning the impact of group composition, group size, aesthetic differences between groups and individuals, and demographic correlations, is lacking. This work establishes a theoretical foundation for IAA, proposing a unified model that encodes individual characteristics in a distributional format for both individual and group assessments. We show that transferring from GIAA to PIAA involves extrapolation, while the reverse involves interpolation, which is generally more effective for machine learning. Extensive experiments with varying group compositions, including sub-sampling by group size and disjoint demographics, reveal substantial performance variation even for GIAA, challenging the assumption that averaging scores eliminates individual subjectivity. Score-distribution analysis using Earth Mover's Distance (EMD) and the Gini index identifies education, photography experience, and art experience as key factors in aesthetic differences, with greater subjectivity in artworks than in photographs. Code is available at https://github.com/lwchen6309/aesthetics_transfer_learning.

1 Introduction

Assessing the aesthetics of images, known as Image Aesthetics Assessment (IAA), is a challenging task due to the inherent complexity of image diversity and individual subjectivity [13, 33, 35, 42, 43, 46, 48]. IAA has become significant due to the success of image

generation [23, 25, 27, 29], which has amplified the need to adjust images according to personal aesthetics [6]. Extensive research [2, 16, 37, 38, 40] has explored how image aesthetics correlate with various image attributes, *e.g.*, spatial composition [16, 37, 38], figure-ground organization [40], and symmetry [2]. Image aesthetics are also influenced by image type, with artworks generally considered more subjective than photographs [39]. Additionally, individual differences in perception play a role in image aesthetic [8, 30, 36], contributing to individual subjectivity that correlates with demographic factors. This adds further challenges for modeling IAA.

Current IAA approaches tackle image diversity and individual subjectivity separately in two stages. First, Generic IAA (GIAA) models [13, 35, 42] estimate averaged user aesthetic scores or score distributions across a broad range of images, aiming to capture a mean score without individual subjectivity. Subsequently, Personal IAA (PIAA) models [17, 18, 31, 41, 42, 43, 46, 47, 48] adapt these generic models, fine-tuning them with a small amount of data (*i.e.* few-shot learning), with or without incorporating personal traits to handle the subjectivity. This process represents a form of transfer learning [50] from GIAA to PIAA, even though it is not explicitly defined as such in the existing PIAA literature.

However, the existing approaches present several limitations. *1)* Existing PIAA approaches make it difficult to analyze aesthetic differences between groups and individuals, as well as their correlation with demographic factors, since these differences can only be inferred through parameter shifts observed during fine-tuning of PIAA. *2)* Although existing GIAA approaches [13, 35, 42] assume that individual subjectivity can be minimized by averaging scores, the bias caused by demography may persist within group averages. Furthermore, these methods often oversimplify group composition by overlooking variations in demographic factors and group size, which can introduce bias and affect the fine-tuning of PIAA. For example, a GIAA dataset with a group size of 2 is clearly closer to personal aesthetics (PIAA) than a group of size 100, resulting in better transfer learning. This challenge remains largely unexplored. *3)* Furthermore, the existing GIAA [13, 35, 42] and PIAA [17, 18, 31, 41, 42, 43, 46, 47, 48] approaches do not adequately address generalization to unseen users, *i.e.* zero-shot learning. Given the high subjectivity of image aesthetics, it is important to investigate model generalization on unseen test users and how it correlates with demographic differences between training and test users.

To address these issues, we propose a novel IAA approach by encoding personal traits in a distributional format that accounts for both individual and group characteristics. Our method is capable of inferencing both GIAA and PIAA with a single model by receiving the corresponding trait distribution as input. This approach reveals the geometry of the IAA domain, where the input space (personal traits) and output space (aesthetic scores) form distinct convex hulls based on personal data for given images, as depicted in Figure 1. We refer to these convex hulls as the trait convex hull and the score convex hull, respectively. In this context, GIAA maps the average trait distribution located at the inner regime of the trait convex hull to the average score distribution located at the inner regime of the score convex hull. In contrast, PIAA maps each vertex of the trait convex hull to corresponding points in the score convex hull. Based on this insight, we claim that transfer learning from GIAA to PIAA represents an extrapolation within the characteristic domain, whereas the reverse direction constitutes an interpolation—a generally more effective approach for machine learning models. If this holds, we would expect PIAA models to perform well on GIAA data without any GIAA pre-training. This hypothesis is further supported by our experimental results. We extend the GIAA and PIAA baseline models by conditioning them on a distributional trait encoding and demonstrate that direct training on PIAA data achieves performance compa-

rable to the GIAA baseline, whereas training on GIAA data underperforms relative to the PIAA baseline. Additionally, we demonstrate that models trained directly on PIAA data match the performance of the GIAA baseline and even exceed the PIAA baseline. These results validate our proposed theory regarding interpolation and extrapolation, offering a more precise analysis of the aesthetic differences between groups and individuals.

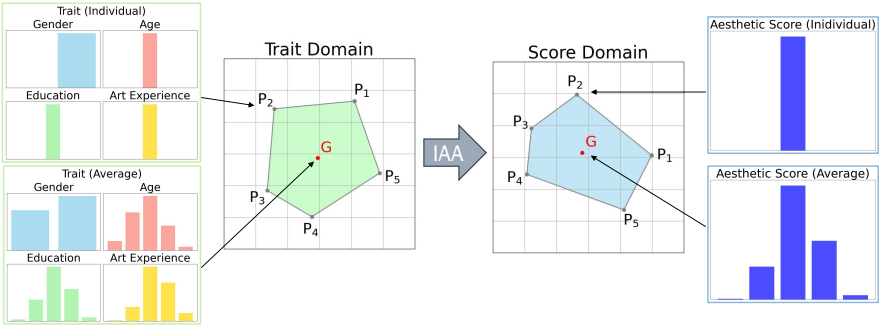


Figure 1: The geometry of IAA. For a given input image, the input trait and output score of the IAA are represented by two convex hulls. Left: The input convex hull in the space of trait distribution, *e.g.*, age, gender, and education; Right: The output convex hull in the space of aesthetic scores distribution. The averaged personal traits distribution of different subsets of individuals all lie within the convex hull formed by the individual traits (*Trait Domain*). These traits can be provided as input to the IAA model, in addition to the input image. Similarly, the average aesthetic score distribution given by a group of individuals all lie within the convex hull formed by the aesthetic scores given by single individuals (*Score Domain*). For both convex hulls, each point P_i represents an individual data point, where $i = 1, 2, 3, \ldots$, and G represents the averaged data.

Then, we demonstrate the transfer learning from GIAA can be significantly affected by group's composition. We introduce *sGIAA*, a data augmentation method that sub-samples GIAA data, increasing diversity in both demography and group size. By sampling between 2 and the maximum number of users per image, we show that this approach maintains GIAA performance while significantly improving zero-shot PIAA performance by at most 44%. This improvement arises from increased demographic variation and from the closer training domain to individual users with smaller group sizes.

Lastly, we investigate the model 's ability to generalize to unseen users by dividing training and test users into distinct groups, according to demography such as gender and education level. We observe a significant performance decrease of up to 80.6% in this setup compared to seen users, with a larger decrement on the artwork dataset than on the photo dataset. Our findings reveal that, for both photos and artworks, education and experience with photography and art are the primary factors driving distinct aesthetic judgments. This further emphasizes the challenge of PIAA when generalizing to unseen users with varying demographic profiles. To our knowledge, ours is the first model enabling both GIAA and PIAA. It matches the performance of the GIAA baseline and surpasses state-of-the-art PIAA models that require GIAA pre-training. Furthermore, it is the first theoretical framework to address aesthetic differences between groups and individuals, account for diverse demographic factors and group size. Our work is the first to comprehensively investigate the model's generalization, *i.e.*, its zero-shot performance on unseen users.

2 Related Work

2.1 PIAA models

The existing PIAA approaches [17, 31, 45, 48, 49] adapt pre-trained GIAA models through fine-tuning. A frequently chosen model is NIMA [35], which predicts score distribution along with predicting aesthetic attribute [31, 48, 49] or personal trait [17]. Other GIAA models that utilize score-based comparison [14] or score regression [13, 18, 44, 46] are considered less often. The fine-tuning for personalization can be improved using several methods. One method uses a meta-learner [18, 46], either alone or combined with a prior model that predicts personal traits [17, 47] or aesthetic attributes [41, 49]. Importantly, these models take only images as input. Other approaches involve models that receive additional personal traits as input [31, 43, 48]. For example, PIAA-MIR [48] and PIAA-ICI [31] involve learning personal scores by computing the interaction between aesthetic attributes and demographic features for personalization. Specifically, PIAA-ICI goes further by extracting aesthetic attributes from both images and demographic features, constructing separate graphs for each, and computing both internal interactions within each graph and external interactions between the two graphs. In contrast, Multi-Level Transitional Contrast Learning (MCTL) [43] uses contrastive learning to learn trait embeddings from personal aesthetic scores, without explicit demographic features. Unlike PIAA-MIR and PIAA-ICI, which can infer image aesthetics for unseen users without fine-tuning, MCTL cannot generalize to unseen users due to its trait embeddings being tied to specific personal scores.

Despite the success of these approaches, the performance of PIAA models when directly evaluated on unseen users remains unclear, as they are typically evaluated under a meta-learning scheme where the model is fine-tuned on each user. While this scheme is appropriate for models that take only images as input [18, 41, 46], it becomes redundant for models that incorporate personal traits [31, 48]. These models should ideally infer image aesthetics for unseen users without requiring fine-tuning, thus performing zero-shot inference on unseen users. Moreover, the current evaluation scheme emphasizes performance variation across image sampling while overlooking variation due to user sampling. This work focuses on PIAA models like PIAA-MIR and PIAA-ICI, which are capable of zero-shot inference, and explores performance variation specifically based on demographic factors.

2.2 Transfer learning

Transfer learning leverages prior knowledge from a source task to improve performance on a related target task with limited data [9, 22, 24, 34, 45, 50]. This approach is essential to existing PIAA research [17, 31, 45, 48, 49], where personal aesthetic data is typically scarce. To quantify source–target alignment, task vectors serve as a key metric by capturing the directional shifts in parameter space needed to adapt a model from a source to a target task, thus providing insight into task alignment and suitability for transfer [1, 11]; Yun and Choo [45] utilize task vectors (*i.e.* model parameters of GIAA models) to facilitate the metric comparisons between GIAA datasets. While their approach successfully analyzes multiple GIAA datasets, demographic differences of the individuals across these datasets cannot be further examined using these task vectors. Complementary distributional metrics—such as Maximum Mean Discrepancy (MMD) [7], Kullback–Leibler (KL) divergence, and Earth Mover's Distance (EMD) [28]—quantify dataset divergence, while performance-based metrics assess transfer effectiveness by measuring target-task accuracy after source pre-training,

thereby guiding optimal model adaptation. Notably, although these metrics have not been used to study transfer learning in existing IAA works, EMD is widely adopted as a training loss in standard IAA protocols [35, 42].

3 On the Geometry of Image Aesthetics

Notation Let s represent the aesthetic score. The function $\hat{P}(s)$ denotes the output score distribution generated by an IAA model, which takes images as inputs in a d-dimensional space, where d is the number of score intervals. The symbol $\delta_i(s)$ represents the groundtruth score distribution for PIAA, expressed as a one-hot vector for an individual score, with i indicating the user.

We demonstrate that for the PIAA models without incorporating personal traits [18, 41, 46], their performance on PIAA serves as the upper bound for their performance on GIAA, as shown in Theorem 3.1.

Theorem 3.1. *With the notation $\hat{P}(s)$ and $\delta_i(s)$ as defined above, and where n is the total number of users, the GIAA and PIAA loss functions are given by*

$$\mathcal{L}_{GIAA} = \left\| \hat{P}(s) - \frac{1}{n}\sum_{i=1}^{n} \delta_i(s) \right\|, \tag{1}$$

$$\mathcal{L}_{PIAA} = \frac{1}{n}\sum_{i=1}^{n} \left\| \hat{P}(s) - \delta_i(s) \right\|. \tag{2}$$

respectively. Then, we have

$$\mathcal{L}_{GIAA} \leq \mathcal{L}_{PIAA}. \tag{3}$$

Note that this result holds not only when $\hat{P}(s)$ and $\delta_i(s)$ represent score distributions but also when they are scalar scores.

See the proof in Suppl. Section 1. This theory suggests that IAA models perform better on GIAA tasks than on PIAA tasks when unconditioned on the user. It follows immediately that PIAA models can generalize well to GIAA data without GIAA pre-training.

Next, we further show that the same statement holds even when the model is conditioned to the demographic traits (hereafter referred to as traits), *e.g.*, PIAA-MIR [48] and PIAA-ICI [31]. Under this setup, the PIAA models map a pair of images and traits to scores. To make this setup compatible with GIAA, we extend the definition of GIAA such that it maps pairs of the averaged traits distribution and images to the average score distribution across users. This extension is reasonable because, without accounting for traits, a GIAA model is likely to overfit the training data by simply memorizing preferences linked to the averaged traits distribution. This definition provides a clear way to model group preferences. With this extension, GIAA maps averaged traits distribution to average score distribution, while PIAA maps individual traits to individual scores for given images. Moreover, both the input space (traits) and output space (scores) of IAA form distinct convex hulls based on all personal data, as illustrated in Figure 1. GIAA projects the average trait distribution located at the inner regime of the trait convex hull to the average score distribution located at the inner regime of the score convex hull, whereas PIAA projects each vertex from the input convex hull to the corresponding point in the output convex hull. This shifts the transfer learning

between GIAA and PIAA into a domain generalization problem, revealing the explicit geometry of IAA. Under this framework, transfer learning from GIAA to PIAA can be viewed as extrapolation in the trait domain, while the reverse is interpolation, which is generally more effective in machine learning. Thus, we conclude that PIAA models can generalize well to GIAA data without GIAA pre-training, even when the model is conditioned on users. We provide experimental results to verify this in Section 4.

4 Experimental Results

4.1 Evaluation Scheme of PIAA

Prior PIAA methods [18, 31, 42, 43, 46, 48] risk data leakage if the same datasets are used for both pre-training and PIAA fine-tuning. This risk arises because GIAA splits data by images, whereas PIAA divides training and test sets by users, potentially resulting in the same images being present in both the GIAA training phase and the PIAA testing phase. To avoid this, we unify the evaluation: all models share a conventional GIAA image split into train/validation/test sets; PIAA uses the full training set (*i.e.*, no few-shot) and is trained collectively across users. For zero-shot generalization to unseen users, we keep the image split but segregate users by demographic (*e.g.*, train on females, test on males). See Suppl. Section 2 for details.

4.2 Datasets

Despite the abundance of image aesthetics resources in GIAA [9, 10, 12, 15, 21, 26, 42, 44], only a few datasets provide personal aesthetic scores for PIAA, such as FLICK-AES [26], PARA [42], and LAPIS [20]. In this work, we demonstrate our method on the PARA and LAPIS datasets, which include photos and artworks, respectively. A sample from these datasets is shown in Suppl. Section 3. Notably, FLICK-AES is excluded from this study as it lacks personal trait data.

PARA dataset [42] includes 31,220 photos and 438 users—each (image, user) pair includes an aesthetic score and aesthetic attributes. The dataset consists of the demographics of users such as age, gender, education, Big-5 personality traits [1], and experience in art and photography. We adopt a unified GIAA/PIAA evaluation (Section 4.1) with 25,398 training, 2,822 validation, and 3,000 test images, and segment users by gender (male/female), age (18–21, 22–25, 26–29, 30–34, 35–40), education (junior high, senior high, technical secondary, junior college, university), and photo/art experience (beginner, competent, proficient, expert).

LAPIS dataset [20] contains 11 723 artworks rated by 578 users—each (image, user) pair includes an aesthetic score, art style, demographics (age, gender, education, nationality), and 11 Vienna Art Interest and Art Knowledge (VAIAK) values (VAIAK1-7 and 2VAIAK1-4) [32]. Images are split into 7,074 train, 2,358 validation, and 2,358 test samples; users are segmented by gender (female/male), age (18–27, 28–38, 39–49, 50–60, 61–71), education (primary, secondary, Bachelor's, Master's, Doctorate), nationality (44 countries detailed in Suppl. Section 4), and 2 VAIAK levels (low ≤ 3 and high > 3)

[1] Openness to experience (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N)

4.3 Trait Encoding and Models

Trait Encoding. Unlike existing PIAA work [31, 48] that uses numeric traits directly, we apply onehot encoding to all traits, including numeric ones like the Big-5 in the PARA dataset and the VAIAKs in the LAPIS dataset. This allows our models to access the full distribution of each trait when computing GIAA, rather than relying on mean values. For PARA, we encode gender (2), age (5), education (5), photography experience (4), art experience (4), and Big-5 personality (50), yielding 70 dimensions. For LAPIS, we encode gender (4), color-blindness (2), age (5), education (5), nationality (44), and VAIAK art-interest/knowledge scores (77), totaling 137 dimensions. By comparison, conventional encodings produce just 25 dimensions for PARA and 71 for LAPIS (see Suppl. Section 6).

Models. As outlined in Section 2, we reimplemented PIAA-MIR and PIAA-ICI as our PIAA baselines—reproducing their results (Suppl. Section 5)—and adopt NIMA [35] as our GIAA baseline. We choose NIMA over newer GIAA models [9, 44] because our PIAA models build on NIMA and the newer models' use of extra attributes (theme, style) would preclude a fair comparison. We adapt NIMA, PIAA-MIR, and PIAA-ICI to onehot encoding, enabling inference for both GIAA and PIAA, which we refer to as NIMA-trait, PIAA-MIR (Onehot enc.), and PIAA-ICI (Onehot enc.), respectively. For NIMA-trait, these traits are integrated into NIMA's predictions via an additional two-layer multilayer perceptron (MLP) as detailed in Suppl. Section 6. These models are built on pretrained backbones, including ResNet-50 [8], Swin-Tiny [19], and ViT-Small [6]. For the GIAA inference of NIMA-trait, PIAA-MIR (Onehot enc.), and PIAA-ICI (Onehot enc.), **we adjust the inference method so that models receive the average trait distribution across all training users during evaluation.** This ensures a fair comparison to the GIAA scenario, where images are the only input. We follow the standard IAA training protocol [35, 42], training these models with EMD loss and evaluating them using SROCC and PLCC.

4.4 Model Evaluation on overlapped users

The SROCC performance of NIMA, PIAA-ICI, PIAA-MIR, and their corresponding onehot-encoded models are shown in Figure 2. The numeric values for both SROCC and PLCC are demonstrated in Suppl. Section 7. When trained on GIAA, onehot-encoded models achieves performance comparable to the NIMA baseline. On the other hand, **when trained on PIAA, these models even outperform the PIAA baselines such as PIAA-MIR and PIAA-ICI, which require GIAA pre-training.** These results demonstrate that our simple yet effective approach performs well in both the GIAA and PIAA settings. Moreover, the generalization between GIAA and PIAA is evident, where the zero-shot PIAA performance ($G \rightarrow P$) is significantly worse than the PIAA baselines while the zero-shot GIAA performance ($P \rightarrow G$) is generally comparable to the GIAA baseline, with only a few exceptions observed on the LAPIS dataset. This observation aligns with our analysis in Section 3.

Building on the generalization discussed above, the number of users involved in GIAA likely affects generalization, as fewer users shifts the training domain toward individual preferences, while more users provides greater confidence in the annotated scores. Although existing works have explored GIAA, the effect of group size has rarely been addressed. To investigate this, we propose sub-sampling the GIAA dataset, referred to as sGIAA, with user groups ranging from 2 to the maximum number of users per image, as a form of data augmentation. The model performances in SROCC for this approach are shown in Figure 3, while the numeric values for both SROCC and PLCC are demonstrated in Suppl. Section

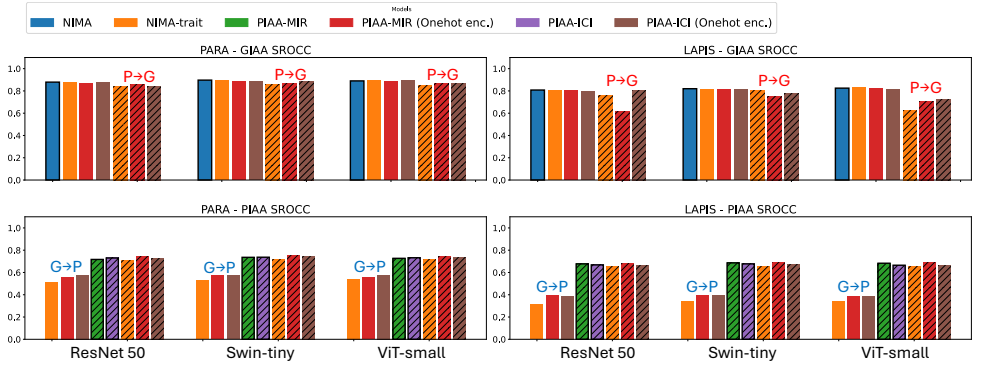


Figure 2: **GIAA/PIAA SROCC of IAA models using the PARA and LAPIS datasets.** Hatched bars indicate models trained on PIAA data, solid bars those trained on GIAA data, and bars with a black outline denote the respective GIAA/PIAA baselines. Zero-shot PIAA performance (trained on GIAA, tested on PIAA) is shown as $G \rightarrow P$, while zero-shot GIAA performance (trained on PIAA, tested on GIAA) is shown as $P \rightarrow G$.

7. These results demonstrate that **sub-sampling significantly improves zero-shot PIAA performance** by up to 44% while maintaining GIAA performance. This reveals how sub-sampling helps strike a balance between individual and group preferences, further emphasizing the importance of the number of users in the GIAA dataset.

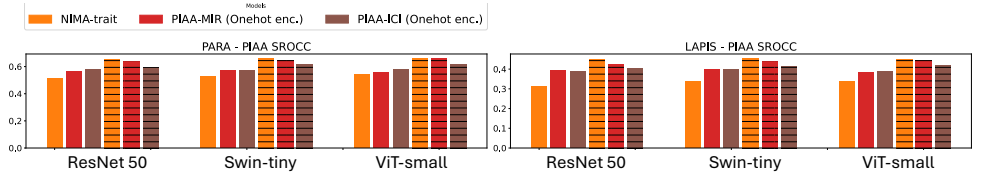
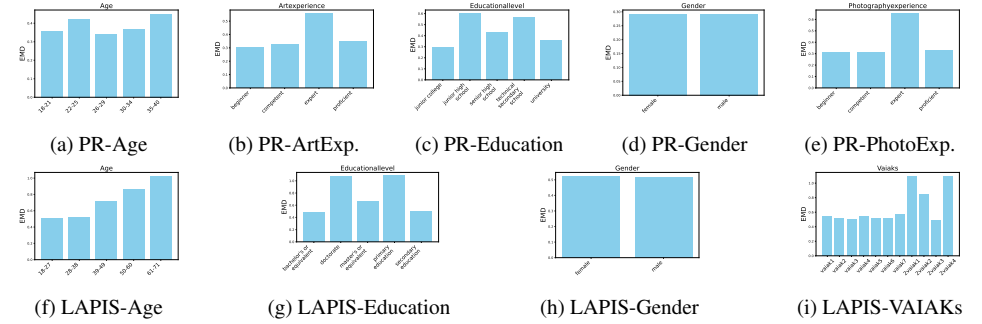


Figure 3: **Zero-shot PIAA SROCC of IAA models using the PARA and LAPIS datasets.** Hatched bars indicate models trained on sGIAA data and solid bars those trained on GIAA data. The performance improves markedly when using sGIAA as augmentation compared to training on GIAA alone.

4.5 Model Evaluation on Disjoint Users across Demography

Analysis of demographic differences. To further assess the aesthetic differences and model generalization across the demographic split, we select users with a specific trait (e.g., users aged 18-21) as the test users, while all other users serve as the training users as described in section 4.1. We hereafter refer to this setup as the disjoint user split. We then compute the Earth Mover's Distance (EMD) between the aesthetic score distributions of the train and test groups for various demographic splits, as shown in Figure 4. A higher EMD indicates a greater distinction in the aesthetic preferences of the test users compared to the training users. The EMD values split by gender are the lowest, while splits based on art experience, photography experience, and educational level show higher EMD values, reaching up to around 0.8. Specifically, experts in both photography and art, as well as users with only high



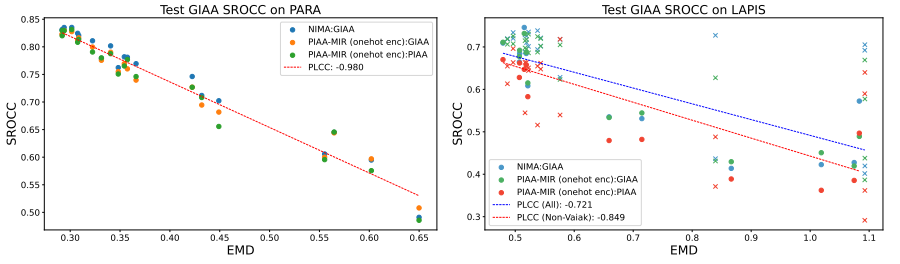
(a) PR-Age (b) PR-ArtExp. (c) PR-Education (d) PR-Gender (e) PR-PhotoExp.

(f) LAPIS-Age (g) LAPIS-Education (h) LAPIS-Gender (i) LAPIS-VAIAKs

Figure 4: EMD between disjoint users split by demographic groups on PARA (a–e) and LAPIS (f–i) datasets. "PR" denotes the PARA dataset.

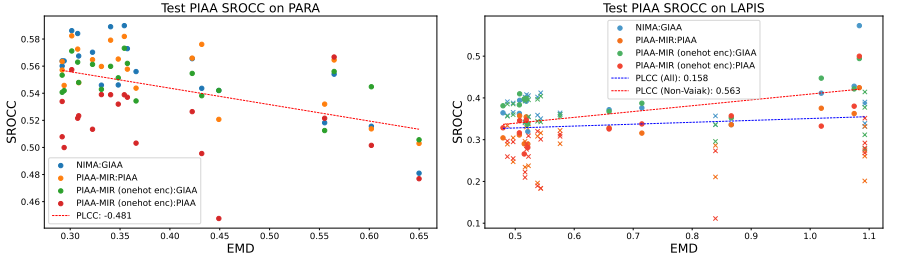
school education, demonstrate the greatest aesthetic distinction. For the LAPIS dataset, splits based on age, educational level, 2VAIAK1, and 2VAIAK4 yield even higher EMD values, reaching up to approximately 1.2. Overall, EMD values in the LAPIS dataset are higher than those in the PARA dataset, suggesting that aesthetic preferences for artworks are more subjective than for photographs—consistent with previous findings [59]. We also compute the Gini index to quantify the relative importance of each demographic factor. The results are consistent with the EMD analysis, indicating that education level and photography/art experience are the strongest distinguishing factors in the score distributions (Suppl. Section 8).

Generalization to new users. Using the disjoint user split, we evaluate NIMA, PIAA-MIR, and PIAA-MIR (Onehot-enc.) across various demographic splits, focusing on models with a ResNet-50 backbone. Further details are provided in Suppl. Section 9. In Figure 5, we plot GIAA SROCC against EMD to show that aesthetic differences driven by demographics are substantial across all models. Performance varies markedly, with SROCC ranging from 0.486 to 0.835 on the PARA dataset and from 0.292 to 0.746 on the LAPIS dataset—equivalent to performance gaps of 41.8% and 60.9%, respectively. **These performance differences across demographic groups indicate systematic variation in aesthetic preferences, showing that even aggregated group scores cannot fully remove individual subjectivity.** Furthermore, the results reveal a strong negative correlation between aesthetic distinction (EMD) and model generalization (SROCC), with PLCCs of –0.980 for PARA and –0.721 for LAPIS. Removing demographic outliers in LAPIS—where certain VAIAK groups are highly imbalanced—improves the PLCC to –0.849. **These results demonstrate that EMD is an effective metric for estimating GIAA model generalization to unseen users.** A similar analysis on PIAA is conducted and illustrated in Figure 6. We observe a smaller variation in SROCC values for the PARA dataset, ranging from 0.448 to 0.590, while the LAPIS dataset exhibits a significantly larger variation, with SROCC values ranging from 0.111 to 0.573. Namely, the model performance can vary by as much as 24.1% and 80.6% on the two datasets, respectively. The greater variation in performance on the LAPIS dataset further underscores the higher individual subjectivity associated with artworks compared to photographs in both GIAA and PIAA models. In particular, the LAPIS dataset shows stronger performance variation for PIAA than for GIAA, indicating a greater challenge in achieving model generalization for unseen users with diverse demographic profiles. The weaker negative correlation between PIAA SROCC values and EMD, compared to GIAA, may indicate improved generalization to unseen users.



(a) PARA dataset (b) LAPIS dataset.

Figure 5: Correlation between GIAA SROCC and EMD across various demographic splits. For LAPIS dataset, the symbol 'x' denotes data split by VAIAKs, with the same color as the 'o' symbols representing the same model.



(a) PARA dataset (b) LAPIS dataset

Figure 6: Correlation between PIAA SROCC and EMD across various demographic splits. The format follows that of Figure 5.

5 Conclusion

We propose the first model capable of supporting both GIAA and PIAA, matching GIAA baseline performance and even surpassing state-of-the-art PIAA models that require GIAA pre-training. For GIAA, while our method requires additional trait data during training, it demonstrates efficacy by using the averaged trait distribution from the training data as a fixed input, enabling the model to rely solely on image inputs during inference. Additionally, our model introduces the first theoretical framework for addressing aesthetic differences between groups and individuals, accounting for diverse demographics and group size. Our comprehensive experiments investigate the transfer learning between GIAA and PIAA under these factors. The results support our theory that transferring from GIAA to PIAA involves extrapolation, while the reverse—interpolation—is generally more effective for machine learning. Additionally, sub-sampled GIAA (sGIAA) improves zero-shot PIAA performance by 44%, underscoring the importance of group size variation, especially for PIAA fine-tuning in the current scenario. For unseen users from diverse demographics, large performance variations highlight the limited generalizability of IAA models and challenge the long-standing assumption that score averaging suppresses subjectivity. We also present the first quantitative evidence that artworks exhibit greater subjectivity than photographs.

Acknowledgment

This research was funded by the European Union (ERC Advanced Grant GRAPPA, 101053925, awarded to Johan Wagemans). We thank Yongzhen Ke for providing the pretrained PIAA-ICI model.

References

[1] Alessandro Achille, Michael Lam, Raman Tewari, Avinash Ravichandran, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Int. Conf. Comput. Vis.*, pages 6430–6439, 2019.

[2] Claudia Damiano, John Wilder, Elizabeth Yue Zhou, Dirk B Walther, and Johan Wagemans. The role of local and global symmetry in pleasure, interest, and complexity judgments of natural scenes. *Psychology of Aesthetics, Creativity, and the Arts*, 17(3): 322, 2023.

[3] Lee de Wit and Johan Wagemans. Individual differences in local and global perceptual organization. 2014.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2018.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5744–5753, 2019.

[7] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arxiv e-prints. *arXiv preprint arXiv:1512.03385*, 10, 2015.

[9] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *International Joint Conferences on Artificial Intelligence*, pages 942–948, 2022.

[10] Heng Huang, Xin Jin, Xinning Li, Shuai Cui, and Chaoen Xiao. Aesthetic evaluation of asian and caucasian photos with overall and attribute scores. *Computers and Electrical Engineering*, 103:108341, 2022.

[11] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Simon Kornblith, Rebecca Roelofs, Benjamin Recht, Ludwig Schmidt, and Awni Hannun. Editing models with task arithmetic. In *ICML*, pages 9638–9652. PMLR, 2022.

[12] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. Eva: An explainable visual aesthetics dataset. In *Joint workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends*, pages 5–13, 2020.

[13] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Int. Conf. Comput. Vis.*, pages 5148–5157, 2021.

[14] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10041–10051, 2023.

[15] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Eur. Conf. Comput. Vis.*, pages 662–679. Springer, 2016.

[16] Lisa Koßmann and Johan Wagemans. Composition and spatial layout of images of artworks in relation to their aesthetic appreciation. 2023.

[17] 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.

[18] Yaohui Li, Yuzhe Yang, Huaxiong Li, Haoxing Chen, Liwu Xu, Leida Li, Yaqian Li, and Yandong Guo. Transductive aesthetic preference propagation for personalized image aesthetics assessment. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 896–904, 2022.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[20] Anne-Sofie Maerten, Li-Wei Chen, Stefanie De Winter, Christophe Bossens, and Johan Wagemans. Lapis: A novel dataset for personalized image aesthetic assessment. *arXiv preprint arXiv:2504.07670*, 2025.

[21] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2408–2415. IEEE, 2012.

[22] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[23] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[24] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. In *OpenAI*, 2018.

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[26] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *Int. Conf. Comput. Vis.*, Oct 2017.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[28] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000.

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[30] Celine Samaey, Johan Wagemans, and Pieter Moors. Individual differences in processing orientation and proximity as emergent features. *Vision Research*, 169:12–24, 2020.

[31] Huiying Shi, Jing Guo, Yongzhen Ke, Kai Wang, Shuai Yang, Fan Qin, and Liming Chen. Personalized image aesthetics assessment based on graph neural network and collaborative filtering. *Knowledge-Based Systems*, 294:111749, 2024.

[32] Eva Specker, Michael Forster, Hanna Brinkmann, Jane Boddy, Matthew Pelowski, Raphael Rosenberg, and Helmut Leder. The vienna art interest and art knowledge questionnaire (vaiak): A unified and validated measure of art interest and art knowledge. *Psychology of Aesthetics, Creativity, and the Arts*, 14(2):172, 2020.

[33] Ombretta Strafforello, Gonzalo Muradas Odriozola, Fatemeh Behrad, Li-Wei Chen, Anne-Sofie Maerten, Derya Soydaner, and Johan Wagemans. Backflip: The impact of local and global data augmentations on artistic image aesthetic assessment. *arXiv preprint arXiv:2408.14173*, 2024.

[34] Nima Tajbakhsh, Jae Shin, Suryakanth R Gurudu, Randall T Hurst, Charles B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[35] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[36] Eline Van Geert and Johan Wagemans. Individual differences in attractive and repulsive context effects on shape categorization. *Journal of Vision*, 21(9):1980–1980, 2021.

[37] Eline Van Geert, Rong Ding, and Johan Wagemans. A cross-cultural comparison of aesthetic preferences for neatly organized compositions: Native chinese-vs. native dutch-speaking samples. 2021.

[38] Eline Van Geert, Christophe Bossens, and Johan Wagemans. The order & complexity toolbox for aesthetics (octa): A systematic approach to study the relations between order, complexity, and aesthetic appreciation. *Behavior Research Methods*, 55(5):2423–2446, 2023.

[39] Edward A Vessel, Natalia Maurer, Alexander H Denker, and G Gabrielle Starr. Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition*, 179:121–131, 2018.

[40] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger Von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[41] Xingao Yan, Feng Shao, Hangwei Chen, and Qiuping Jiang. Hybrid cnn-transformer based meta-learning approach for personalized image aesthetics assessment. *Journal of Visual Communication and Image Representation*, 98:104044, 2024.

[42] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19861–19869, 2022.

[43] Zhichao Yang, Leida Li, Yuzhe Yang, Yaqian Li, and Weisi Lin. Multi-level transitional contrast learning for personalized image aesthetics assessment. *IEEE Transactions on Multimedia*, 2023.

[44] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22388–22397, 2023.

[45] Jooyeol Yun and Jaegul Choo. Scaling up personalized aesthetic assessment via task vector customization. *arXiv preprint arXiv:2407.07176*, 2024.

[46] 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*, 52(3):1798–1811, 2020.

[47] Hancheng Zhu, Yong Zhou, Leida Li, Yaqian Li, and Yandong Guo. Learning personalized image aesthetics from subjective and objective attributes. *IEEE Transactions on Multimedia*, 2021.

[48] Hancheng Zhu, Yong Zhou, Zhiwen Shao, Wenliang Du, Guangcheng Wang, and Qiaoyue Li. Personalized image aesthetics assessment via multi-attribute interactive reasoning. *Mathematics*, 10(22):4181, 2022.

[49] Hancheng Zhu, Zhiwen Shao, Yong Zhou, Guangcheng Wang, Pengfei Chen, and Leida Li. Personalized image aesthetics assessment with attribute-guided fine-grained feature representation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6794–6802, 2023.

[50] Fuzhen Zhuang, Zhiqiang Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.