AACP: Aesthetics Assessment of Children's Paintings Based on Self-Supervised Learning

Shiqi Jiang[1, 3], **Ning Li**[1], **Chen Shi**[1], **Liping Guo**[2], **Changbo Wang**[1], **Chenhui Li**[1*]

[1]School of Computer Science and Technology, East China Normal University
[2]Faculty of Education, East China Normal University
[3]Shanghai Institute of AI for Education, East China Normal University
{52265901032, 51215901058, 52205901019}@stu.ecnu.edu.cn
lpguo@pie.ecnu.edu.cn, {cbwang, chli}@cs.ecnu.edu.cn

Abstract

The **A**esthetics **A**ssessment of **C**hildren's **P**aintings (AACP) is an important branch of the image aesthetics assessment (IAA), playing a significant role in children's education. This task presents unique challenges, such as limited available data and the requirement for evaluation metrics from multiple perspectives. However, previous approaches have relied on training large datasets and subsequently providing an aesthetics score to the image, which is not applicable to AACP. To solve this problem, we construct an aesthetics assessment dataset of children's paintings and a model based on self-supervised learning. 1) We build a novel dataset composed of two parts: the first part contains more than 20k unlabeled images of children's paintings; the second part contains 1.2k images of children's paintings, and each image contains eight attributes labeled by multiple design experts. 2) We design a pipeline that includes a feature extraction module, perception modules and a disentangled evaluation module. 3) We conduct both qualitative and quantitative experiments to compare our model's performance with five other methods using the AACP dataset. Our experiments reveal that our method can accurately capture aesthetic features and achieve state-of-the-art performance.

Introduction

Aesthetics education plays a crucial role in the holistic development of children as it fosters the development of aesthetic skills, stimulates creativity, improves cultural literacy, and enhances social skills (Denac 2014). Children's painting is a way for children to express their emotions, feelings and understanding of things. It is a reflection of their cognitive and emotional development, as well as their ability to perceive and interpret the world around them (Robson and Rowe 2012; Chang 2005). Aesthetics assessment of children's paintings is a crucial component of aesthetics education, as it provides a way to measure and perceive the aesthetic qualities of children's paintings from multiple perspectives. By analyzing quantitative attributes, such as composition and color, researchers can gain a more comprehensive understanding of the cognitive processes and artistic abilities of children.

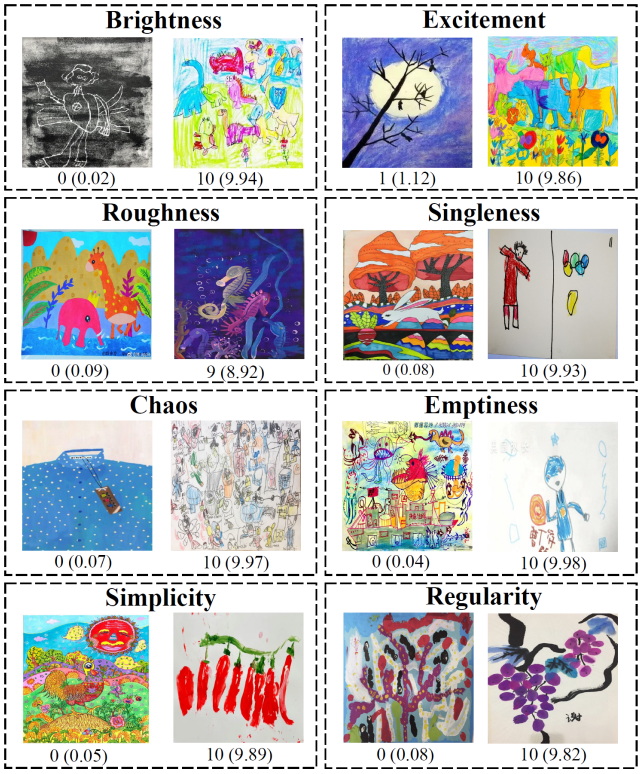


Figure 1: Examples of images and annotations in the proposed dataset with the ground truth (and predicted) scores at the bottom. We assess the aesthetics of children's paintings from 8 attributes.

Traditionally, the aesthetics assessment of children's paintings was performed by experts in the field of art or design, who would evaluate the content, colors, and other aspects to infer the meaning or message behind the painting (Sali, Akyol, and Baran 2014). However, such methods are inherently subjective and may be influenced by personal bias or preconceived notions. In addition, their assessment metric is relatively single and potentially insufficient to capture the full range of potential meanings in children's paintings.

In recent years, researchers have explored the application

*Corresponding author

of deep learning algorithms for assessing the aesthetic quality of paintings (Esfandarani and Milanfar 2018; She et al. 2021). This approach potentially simplifies the assessment process and offers a more objective and consistent way to assess the aesthetic value of paintings. Consequently, we have incorporated deep learning methods into the aesthetics assessment of children's paintings, aiming to improve the accuracy and objectivity of our assessment. However, this approach has also faced two challenges.

First, existing IAA datasets such as AVA (Murray, Marchesotti, and Perronnin 2012) and AADB (Kong et al. 2016) are not well-suited for the aesthetics assessment of children's paintings for two main reasons. 1) The majority of images in these datasets are natural images or photographs, which significantly differ from children's paintings. As shown in Figure 1, children's paintings often have unique characteristics, such as abstract and personalized styles, as well as simple and vivid compositions, which may not be effectively represented in existing datasets. 2) The single scores in these datasets inadequately represent children's paintings' broad aesthetic range and potential meanings that often express their emotions and cognitive development.

Second, previous IAA approaches used open datasets and direct image-to-score mapping (Esfandarani and Milanfar 2018), achieving state-of-the-art performance on natural images using various techniques, such as multi-scale representation (Ke et al. 2021), graph convolution networks (GNNs) (She et al. 2021), and other methods. However, such approaches may not be well suited for the aesthetics assessment of children's paintings, as they fail to adequately capture the intrinsic meanings and unique characteristics of these images, and it is insufficient to rely on a single score to evaluate the aesthetic quality of children's paintings.

To overcome existing challenges in assessing the aesthetics of children's paintings, we have made two key contributions. First, we have constructed the first AACP-specific dataset, as shown in Figure 1. Our dataset, the first of its kind specifically designed for AACP, provides a valuable resource for studying children's painting aesthetics and art-creation psychology. This dataset enables researchers to gain a more comprehensive understanding of these aesthetics, facilitating aesthetics education and children's artistic development. Second, we propose a network architecture for AACP that comprises a masked encoder, two perception modules and a disentangled evaluation module. This architecture offers an accurate method for assessing children's painting aesthetics, further prompting children's artistic growth and aesthetics education.

In summary, our main contributions include:

- We have constructed a novel and multi-attribute dataset for the aesthetics assessment of children's paintings, specifically tailored to support children's education.
- We propose an effective model based on self-supervised learning to extract aesthetic features for AACP, eliminating the need for extensive manual annotations.
- Our approach demonstrates outstanding results on the AACP dataset through extensive experimentation, surpassing the performance of other methods for assessing the aesthetic qualities of images.

Related Work

Self-Supervised Learning

Self-supervised learning, particularly contrastive learning and masked auto-encoding, is frequently utilized in computer vision. Contrastive learning aims to identify common features among positive samples while distinguishing differences between negative samples. One notable method is SimCLR (Chen et al. 2020), which employs a simple framework and has demonstrated superior performance over previous methods. Masked auto-encoding, on the other hand, encourages the model to learn local features of the data by partially masking the input data. Both MAE (He et al. 2022) and ConvMAE (Gao et al. 2022) have demonstrated success in obtaining excellent representations using masked auto-encoding. Furthermore, self-supervised learning can also be implemented through different masked views, as shown in works like data2vec (Baevski et al. 2022) and MaskFeat (Wei et al. 2022). These pre-trained models obtained through self-supervised learning can be effectively employed in downstream tasks such as classification and segmentation (Bao et al. 2022; Caron et al. 2021).

In our study, we employ self-supervised learning to primarily reduce the need for labeled data and thus mitigate the cost of manual labeling. We further optimize the training and inference speed by simplifying the ConvMAE structure, facilitating the efficient assessment of results.

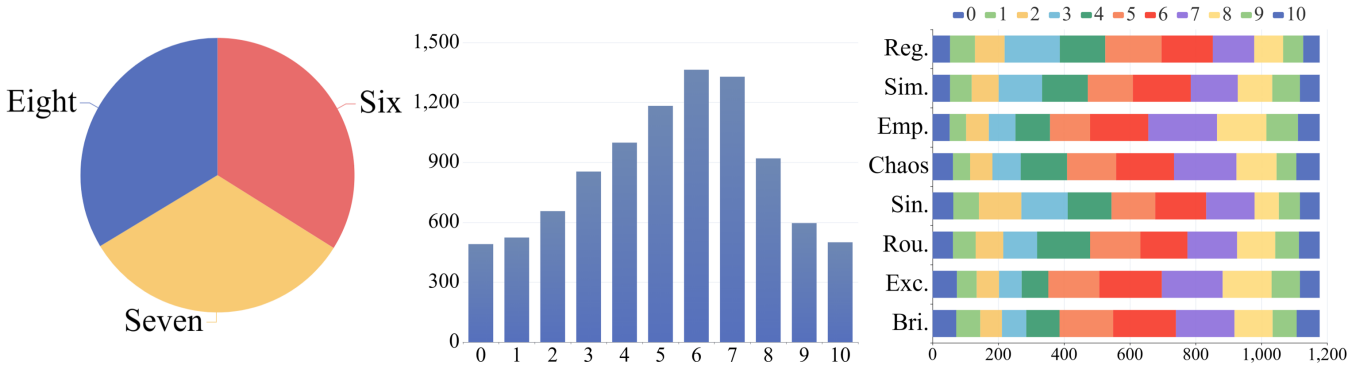
IAA Model

Early IAA models map aesthetics features from images to scores (Lu et al. 2014, 2015; Mai, Jin, and Liu 2016). Following this, A_LAMP (Ma, Liu, and Chen 2017) predicts aesthetics scores based on layout and patch. NIMA (Esfandarani and Milanfar 2018) produces a distribution of aesthetics ratings closely matching human ratings. MP_{ada}(Sheng et al. 2018) uses multi-patch and attention mechanisms to improve learning efficiency when only aesthetics labels are available. MLSP(Hosu, Goldlücke, and Saupe 2019) and MUSIQ (Ke et al. 2021) propose a structure that maintains original image resolution as input, handling images of arbitrary size without information loss caused by cropping or scaling. UGIAA (Lv et al. 2023) uses reinforcement learning and HGCN (She et al. 2021) employs graph networks, while PIAA (Zhu et al. 2022) leverages meta-learning. These diverse methods have advanced image aesthetics assessment from different perspectives.

However, the significance of spatial information in elements and the insufficiency of relying on labels from image aesthetics datasets for understanding children's paintings have been overlooked. Thus, we propose to explore spatial and channel features to improve AACP.

IAA Dataset

In recent years, several datasets have been constructed for image aesthetics assessment. AVA (Murray, Marchesotti, and Perronnin 2012) is a large-scale dataset that contains



(a) Age (b) Score (c) Attribute

Figure 2: The age, score and attribute distribution of AACP dataset.

over 250k images with aesthetics, semantic, and style labels, and has been widely used in aesthetics assessment. After this, AADB (Kong et al. 2016) includes annotations for eight aesthetic factors, but the labels are not detailed enough to capture the diversity of aesthetics. PCCD (Chang, Lu, and Chen 2017) first used linguistic comments with multiple aesthetic factors, which has comprehensive annotation but fewer data. Art500k (Mao, Cheung, and She 2017) and OmniArt (Strezoski and Worring 2017) are large-scale art painting datasets, but the data are collected from paintings in museums rather than photographs. SemArt (Garcia and Vogiatzis 2018) is a dataset for semantic art understanding that includes attributes and textual art comments for each image, but is not specific to children's paintings. While these datasets have advanced the field of image aesthetics assessment in different areas, they are not suitable for the aesthetics assessment of children's paintings.

AACP Dataset

Painting Collection

To collect high-quality children's paintings, we first invited hundreds of children to participate in painting, whose age distribution is shown in Figure 2 (a). We did not constrain the content or duration of painting, in order to ensure that the collected works are complete and diverse. The collected paintings include works created using various tools such as oil pastels, watercolor pens, crayons, and brushes, as well as composite materials made using collages.

Secondly, we invited ten experts in the design field to screen and annotate the paintings. The quantity of the manually labeled image was approximately 1.2k. To ensure the quality of the data, we set strict standards for the selection and labeling process. The labeled data was then used for training and testing our model. We believe that our approach of collecting and annotating children's paintings can be used as a reference for future studies in this area.

Painting Annotation

Each image is scored on a scale of 0 to 10 for each attribute, with a higher score indicating a closer alignment with that metric. Based on design theory and composition principles, experts have divided the assessment of children's paintings into four aspects: color, texture, composition, and conception. To make the annotation process easier, we use *brightness* and *excitement* to describe **color**, *roughness* and *singleness* to represent **texture**, *chaos*, *emptiness* and *simplicity* to analyze **composition**, and *regularity* to represent **conception**. In total, we use eight attributes to assess children's paintings. The distribution of scores and attributes in our dataset is shown in Figure 2 (b) and (c), respectively. We have balanced the amount of data for each score and attribute to avoid long-tailed distributions. Please refer to the **supplementary materials** for an explanation of this annotation method.

Specifically, each image is evaluated by multiple experts, and each attribute receives multiple scores. The final annotation value for each attribute is the average of these scores. By using this method, we can reduce the influence of subjective factors in the evaluation process.

Dataset Expansion

To overcome manual labeling constraints, we augment our children's painting dataset using generative models. While high-performance generators like StyleGAN2 (Karras et al. 2020) and DDPM (Ho, Jain, and Abbeel 2020) create realistic images, their lack of control limits their suitability. Instead, we focus on semantic-based image generation methods, such as DALL-E (Ramesh et al. 2022), Imagen (Saharia et al. 2022), and Stable Diffusion (Rombach et al. 2022) for controlled and high-quality generation. In our work, we specifically use DALL-E to generate children's painting images based on keyword combinations.

We carefully choose semantic keywords related to children's painting attributes—such as tone, pattern, and composition—to maximize image diversity and realism. Employing these keywords with DALL-E, we generate a diverse dataset of approximately 20,000 images.

We employ this approach for two primary reasons. Firstly, the Dall-E model produces diverse, compliant children's paintings in large quantities, with prompts ensuring diversity. Secondly, the Fréchet Inception Distance (FID) met-

ric (Heusel et al. 2017) scored 7.85, suggesting high similarity between generated and real paintings. This allows us to pre-train our model using generated images.

Annotation Explanation

Color We assess the color in children's paintings based on brightness and excitement. Brightness denotes the lightness or darkness of a color, with lighter colors having high brightness and darker colors having low brightness. Excitement pertains to color saturation and vibrancy; bright, saturated colors evoke excitement, whereas darker, unsaturated hues suggest calmness. Thus, we assess a painting's color using both brightness and excitement.

Texture The texture in children's paintings is analyzed based on factors such as roughness versus smoothness, and singularity versus complexity. Rough textures may indicate a child's anxiety or unrefined brush control, while smooth textures could suggest relaxation or confidence. A painting with a singular texture may point towards a lack of interest or creativity, while complex textures might reflect high levels of excitement or creativity.

Composition The composition of children's paintings, which includes element arrangement and tone relationships, conveys themes and aesthetics. We assess composition by examining the degrees of chaos, emptiness, and simplicity. Chaos might represent a child's emotional unrest or an unclear idea. Emptiness could signify introversion, a creativity deficit, or a lack of motivation. Simplicity, implying the level of detail, may suggest a limited imagination if the composition lacks complexity.

Conception We assess the concept of a child's painting by examining the level of regularity present in the work. A regular concept can be interpreted as an indication of the child's self-confidence in their abilities or trust in their judgment during the creative process, or it may reflect the child's imaginative thinking at the time of drawing. Conversely, an irregular concept may represent the child's curiosity or creativity, or it could be a sign of strong critical thinking skills demonstrated during the drawing process.

Model

Overview

The aesthetics assessment of children's paintings presents several challenges. First, insufficient labeled data may limit network learning capacity, inducing prediction biases. Second, the model needs to consider not only the aesthetic features of the images, but also the semantic and emotional content of the paintings. Third, the eight attributes of children's paintings present a complex mapping problem that can make model training unstable. To address these challenges, we propose a four-module network architecture (Figure 3). Through careful design and integration, we aim to enhance network performance on AACP.

Self-supervised Learning Network

To address the issue of limited labeled data, we employ a self-supervised learning strategy. As shown in Figure 3, we utilize a structure similar to ConvMAE (Gao et al. 2022). During the training phase, generated images of children's paintings are used to train the model. The masking ratio is set to 0.75. After training for 1500 epochs, our model achieves 86% accuracy on the test set. In the fine-tuning phase, we train the network on real children's paintings without masking. After fine-tuning for 20 epochs, the model obtains 93% accuracy on real children's paintings. The latent codes from this module will be used in the prediction model, which includes the perception modules and the evaluation module and is described as follows:

$$S = F(S_e(x, \theta_s), \theta_F), \tag{1}$$

where S represents the predicted aesthetic score, F indicates the prediction model, S_e means the encoder in self-supervised model, and x represents the input image.

Spatial Perception Network

The spatial perception module aims to preserve spatial information of children's paintings, which is a crucial factor reflecting a child's psychological state that may be lost during encoding, leading to inaccurate aesthetic scores. To enhance scoring precision, we implement a spatial perception module which, as depicted in Figure 3, fuses the latent encodings learned from the self-supervised module into each convolution layer. The process can be expressed as:

$$S = \omega_\sigma(e)\frac{F_i - \mu(F_i)}{\sigma(F_i)} + \omega_\mu(e), \tag{2}$$

where F_i denotes an intermediate feature map from each convolution layer. e is the latent encodings from the self-supervised module, and ω_σ and ω_μ are the learnable parameters for the standard deviation and mean, respectively. The function $\mu(F_i)$ and $\sigma(F_i)$ compute the channel-wise mean and standard deviation of F_i, respectively.

While our structure is similar to EQGAN-SA (Wang et al. 2022), their goal is to maintain spatial positions, requiring Gaussian distribution sampling. In contrast, our model incorporates features obtained from self-supervised learning into the convolution module, ensuring that spatial information is not lost and effectively captures the intrinsic meanings of children's paintings, improving the performance of the network on AACP.

Channel Perception Network

The channel perception network is designed to extract channel information. In the aesthetics assessment of children's paintings, the channel information affects the accuracy of the aesthetic score. Some methods such as SE-NET (Hu, Shen, and Sun 2018) and DAN (Fu et al. 2019) use attention mechanism to learn the channel information, but they reduce the number of channels to reduce the computational effort. Therefore, to capture the channel information, we use a learning-based cross-channel module.

First, the latent embeddings $x^{N\times L\times E}$ is transposed to $z^{N\times(LE)}$. Then, the output x_c of our perception can be described as:

$$x_c = \sigma(W^k(z)) \cdot x, \tag{3}$$

where W^k indicates a $1D$ convolution layer with kernel size $=$ k. In our experiments, we set k = 3 and padding $= 1$, which are common parameter settings for obtaining weights.

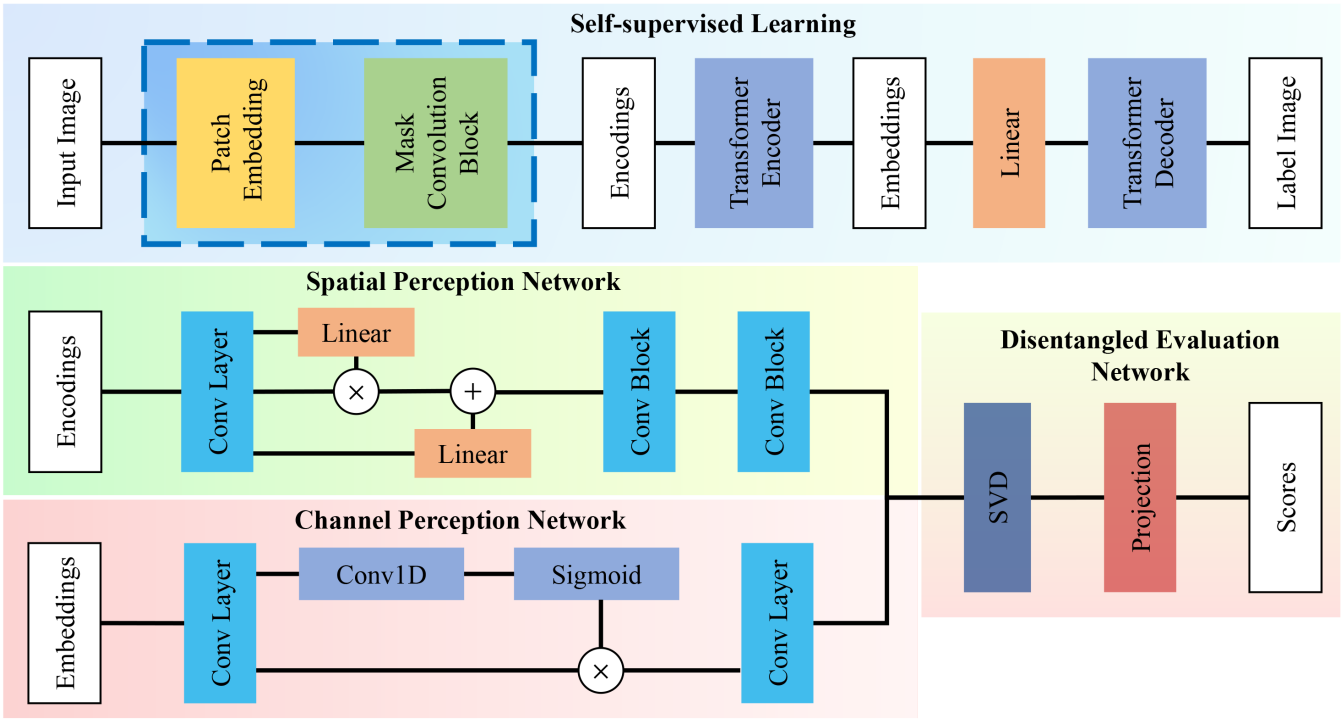


Figure 3: Our network consists of four parts: self-supervised learning, a spatial perception network, a channel perception network and a disentangled evaluation network.

Disentangled Evaluation Network

The evaluation module is designed to map spatial features and channel features to aesthetic scores. However, manually labeled aesthetic scores often contain small errors on each attribute. Previous methods that directly learn the mapping through fully connected layers result in direct interactions between different attributes, leading to lower accuracy. To address this issue, we adopt feature decoupling techniques in the model. There are various ways of feature decoupling, such as Principal Component Analysis (PCA) (Ke and Sukthankar 2004) and Singular Value Decomposition (SVD) (Aharon, Elad, and Bruckstein 2006). PCA treats the features as a high-dimensional vector and finds the correlation between feature maps by computing the covariance matrix and performing eigenvalue decomposition. This results in a disentangled representation. Similarly, SVD decomposes the features into matrices and retains the largest singular values to obtain a disentangled representation.

Due to the non-square nature of the acquired feature maps in our study, SVD was employed as an alternative to PCA, which is unable to handle non-square data directly. We fuse spatial and channel features and apply SVD to project the results onto the eight aesthetic attributes.

Training Details

Our model is trained on an NVIDIA RTX 3090 using PyTorch, and takes 256×256 fixed images as the input. The training process is divided into two steps. In the first step, we use the default configurations to obtain the parameters of the self-supervised representation model. In the second step, we train the remaining modules without data augmentation. The Adam algorithm is used to optimize the model, and the Mean Squared Error (MSE) is used as the loss function. The model converges after 400 epochs of training with a learning rate of 1×10^{-4} and a batch size of 64.

Evaluation

Qualitative Evaluation

Figure 4 illustrates our proposed method's performance evaluation on spatial features, leveraging SmoothGrad-Cam (Omeiza et al. 2019) for feature extraction. We compare the feature maps of the last convolution layer in the spatial perception module. Our method identifies more activation regions than both NIMA and AADB, and when compared to PIAA, it shows varied activation region differences across attributes, showing our method's accuracy in feature extraction and disentanglement. An ablation study, removing the spatial perception module, confirmed its effectiveness by revealing a significant decrease in activation regions, attesting to the module's capability in aesthetic feature perception.

Quantitative Evaluation

Spearman's rank correlation coefficient (SRCC) (Esfandarani and Milanfar 2018) and the linear correlation coefficient (LCC) (Esfandarani and Milanfar 2018) are commonly used metrics for quantifying the results of evaluation models. The Earth Mover's Distance (EMD) (Esfandarani and

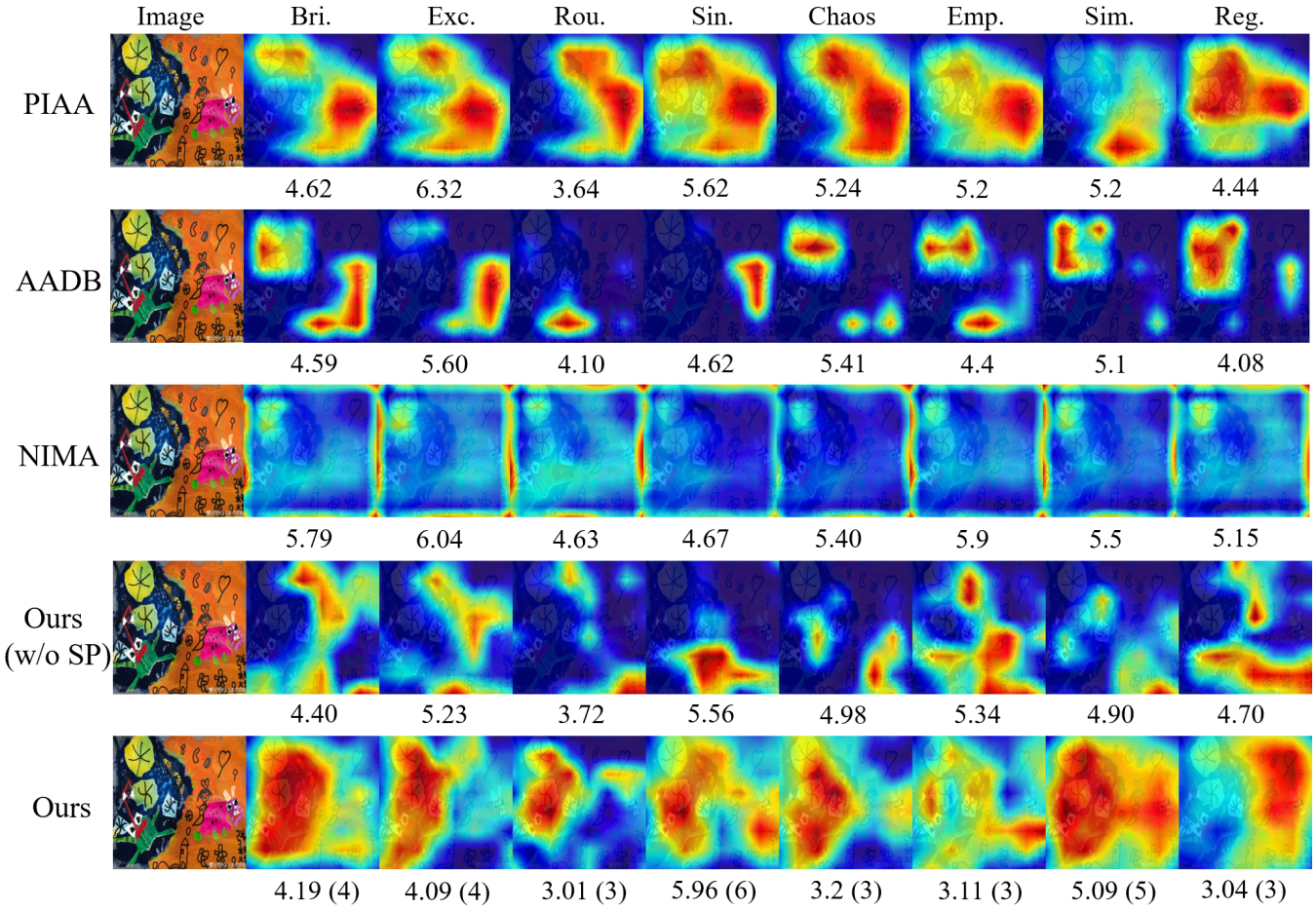


Figure 4: Spatial visual attention at intermediate layers, visualized by SmoothGradCAM. A strong contribution to the final score is indicated by a bright color. The last line shows the result of our method (ground truth).

Milanfar 2018) and MSE are two commonly utilized methods for calculating errors. In our quantitative experiments, we primarily employ these four metrics for comparison with other IAA models. All models were trained on the children's painting dataset and utilized the recommended parameter settings. As shown in Table 1, our method outperformed in most metrics, except EMD, where A LAMP was slightly better. This indicates that incorporating spatial and channel information and utilizing feature disentanglement in the analysis of children's paintings fundamentally improves the perception of aesthetics in such images.

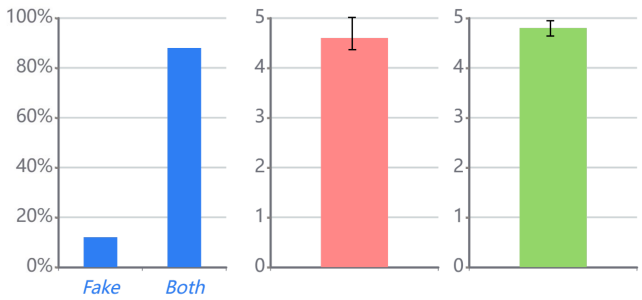
User Study

We conducted user studies to evaluate our dataset and model in terms of three aspects: *dataset expansion*, *painting annotation*, and *aesthetics assessment*. We invited 15 participants, consisting of 10 non-experts and 5 experts.

Dataset Expansion To assess the dissimilarity between generated and real children's paintings, we conducted an experiment where participants first viewed 20 real children's paintings. Subsequently, they were presented with 15 pairs of images, each pair containing a generated and a real painting, and asked to identify the real one. They were allowed to select both if both appeared real. Additionally, participants rated 15 generated paintings on a scale of 0 to 5, with 0 indicating significant deviation and 5 denoting strong similarity to real paintings.

As shown in Figure 5 (a), 12% of participants identified the generated paintings as real, while 88% selected both. Additionally, Figure 5 (b) indicates the dataset expansion yielded an average score of 4.6, suggesting that users found it challenging to distinguish between real and generated images. Thus, we conclude that the generated paintings, due to their similarity to real ones, are appropriate for dataset augmentation.

Painting Annotation To verify the validity and accuracy of our dataset annotations, we randomly selected 10 images. Participants were asked to observe and rate the reasonableness of scores for each attribute on a scale of 0 to 5. A rating of 0 indicates that the annotations are unreasonable, while a rating of 5 indicates that they are reasonable. As depicted in Figure 5 (c), the average rating for our painting annotations was found to be 4.8, indicating that our annotations are consistent with human perception and reflect a high level of



(a) Distinction (b) Expansion (c) Annotation

Figure 5: Statistic results of the user study.

accuracy. Therefore, our dataset is reliable.

Aesthetics Assessment We conducted a user study involving 15 participants to validate our method, comparing it with five other well-known methods. Participants evaluated the aesthetic assessment results of 25 randomly selected children's paintings, providing ratings on a scale of 1 to 10, with 10 indicating the most reasonable aesthetic score. Considering the subjective nature of aesthetics, diverse ratings provide a general consensus on the aesthetic qualities of the paintings.

We collect approximately 1200 judgments. The mean and standard error of these assessments are presented in Table 2. Our method obtains the highest mean score, indicating effective feature extraction from children's paintings can improve assessment accuracy. Moreover, our method achieves the lowest standard error, showing stability due to its disentangled nature.

Ablation Study

In the ablation study, we evaluate the effectiveness of each component of our model. Our results, as presented in Table 3, indicate a significant decrease in performance when the self-supervised learning component is removed. This finding confirms the importance of including self-supervised learning in our model. We also found that the disentangled network component plays a crucial role, as evidenced by a significant decrease in the MSE metric when it is removed.

Additionally, our experiments demonstrated that the spatial and channel perception modules contribute to the overall performance of the model. Without these components, there is a slight decrease in all metrics. Therefore, we conclude that all four parts of our model are necessary to enhance the assessment ability of our method.

Metric	SRCC ↑	LCC ↑	EMD ↓	MSE ↓
NIMA	0.17	0.17	1.38	0.55
AADB	0.21	0.23	0.59	0.49
MLSP	0.36	0.39	0.62	0.42
A_LAMP	0.05	0.04	**0.14**	0.46
PIAA	0.27	0.30	0.15	0.45
Ours	**0.61**	**0.65**	0.38	**0.08**

Table 1: Comparison of 5 state-of-the-art IAA models on the CP dataset. For all models with publicly available codes, we use the recommended parameter settings.

Method	Mean ↑	Std ↓
MLSP	9.4	2.1
A_LAMP	9.5	1.6
PIAA	9.2	0.9
AADB	8.7	1.3
NIMA	8.9	1.1
Ours	**9.7**	**0.8**

Table 2: Results of user study comparing the performance of various methods on AACP.

Type (W/O)	SRCC ↑	LCC ↑	MSE ↓
SSL	0.16	0.20	0.12
DE	0.49	0.48	0.14
CP	0.55	0.56	0.11
SP	0.41	0.42	0.13
Full Model	**0.61**	**0.65**	**0.08**

Table 3: Ablation studies of our network on AACP dataset. Each score is calculated as the average of the scores of the 8 attributes.

Conclusion

In this paper, we first construct a dataset consisting of 20k unlabeled generated children's paintings and 1.2k manually labeled real children's paintings with eight attributes. Then, we design a model that includes a self-supervised learning module, a spatial perception module, a channel perception module and a disentangled evaluation module. Both quantitative experiments and user studies show that our method achieves SOTA performance on the aesthetics assessment of children's paintings. We also conduct ablation studies to investigate the impact of each module.

In the future, we will explore the aesthetic attributes of children's paintings more comprehensively and from varied dimensions to better understand children's aesthetic standards and creative expression. Furthermore, we plan to investigate the impact of environmental factors on AACP to uncover how various environmental factors may influence the development and expression of children's aesthetic sensibilities, providing valuable insights into how to cultivate creativity in young children.

Acknowledgments

This work was supported by the NSSFC under Grant 22ZD05, the Shanghai Committee of Science and Technology under Grant 22511104600, and the ECNU-funded Project on the Analysis of Musical Paintings.

References

Aharon, M.; Elad, M.; and Bruckstein, A. M. 2006. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11): 4311–4322.

Baevski, A.; Hsu, W.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 1298–1312. PMLR.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 9630–9640. IEEE.

Chang, K.; Lu, K.; and Chen, C. 2017. Aesthetic Critiques Generation for Photos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 3534–3543. IEEE Computer Society.

Chang, N. 2005. Children's Drawings: Science Inquiry and beyond. *Contemporary Issues in Early Childhood*, 6: 104 – 106.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Denac, O. 2014. The Significance and Role of Aesthetic Education in Schooling. *Creative Education*, 05: 1714–1719.

Esfandarani, H. T.; and Milanfar, P. 2018. NIMA: Neural Image Assessment. *IEEE Transactions on lmage Processing.*, 27(8): 3998–4011.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual Attention Network for Scene Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 3146–3154. Computer Vision Foundation / IEEE.

Gao, P.; Ma, T.; Li, H.; Lin, Z.; Dai, J.; and Qiao, Y. 2022. ConvMAE: Masked Convolution Meets Masked Autoencoders. *CoRR*, abs/2205.03892.

Garcia, N.; and Vogiatzis, G. 2018. How to Read Paintings: Semantic Art Understanding with Multi-modal Retrieval. In Leal-Taixé, L.; and Roth, S., eds., *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11130 of *Lecture Notes in Computer Science*, 676–691. Springer.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 15979–15988. IEEE.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6626–6637.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hosu, V.; Goldlücke, B.; and Saupe, D. 2019. Effective Aesthetics Prediction With Multi-Level Spatially Pooled Features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 9375–9383. Computer Vision Foundation / IEEE.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7132–7141. Computer Vision Foundation / IEEE Computer Society.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 8107–8116. Computer Vision Foundation / IEEE.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 5128–5137. IEEE.

Ke, Y.; and Sukthankar, R. 2004. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA*, 506–513. IEEE Computer Society.

Kong, S.; Shen, X.; Lin, Z. L.; Mech, R.; and Fowlkes, C. C. 2016. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, 662–679. Springer.

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z. 2014. RAPID: Rating Pictorial Aesthetics using Deep Learning. In Hua, K. A.; Rui, Y.; Steinmetz, R.; Hanjalic, A.; Natsev, A.; and Zhu, W., eds., *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, 457–466. ACM.

Lu, X.; Lin, Z.; Shen, X.; Mech, R.; and Wang, J. Z. 2015. Deep Multi-patch Aggregation Network for Image Style, Aesthetics, and Quality Estimation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santi-*

ago, Chile, December 7-13, 2015, 990–998. IEEE Computer Society.

Lv, P.; Fan, J.; Nie, X.; Dong, W.; Jiang, X.; Zhou, B.; Xu, M.; and Xu, C. 2023. User-Guided Personalized Image Aesthetic Assessment Based on Deep Reinforcement Learning. *IEEE Trans. Multim.*, 25: 736–749.

Ma, S.; Liu, J.; and Chen, C. W. 2017. A-Lamp: Adaptive Layout-Aware Multi-patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 722–731. IEEE Computer Society.

Mai, L.; Jin, H.; and Liu, F. 2016. Composition-Preserving Deep Photo Aesthetics Assessment. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 497–506. IEEE Computer Society.

Mao, H.; Cheung, M.; and She, J. 2017. DeepArt: Learning Joint Representations of Visual Arts. In Liu, Q.; Lienhart, R.; Wang, H.; Chen, S. K.; Boll, S.; Chen, Y. P.; Friedland, G.; Li, J.; and Yan, S., eds., *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 1183–1191. ACM.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2408–2415. IEEE Computer Society.

Omeiza, D.; Speakman, S.; Cintas, C.; and Weldemariam, K. 2019. Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models. *CoRR*, abs/1908.01224.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR*, abs/2204.06125.

Robson, S.; and Rowe, V. 2012. Observing young children's creative thinking: engagement, involvement and persistence. *International Journal of Early Years Education*, 20: 349 – 364.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 10674–10685. IEEE.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR*, abs/2205.11487.

Sali, G.; Akyol, A. K.; and Baran, G. 2014. An Analysis of Pre-school Children's Perception of Schoolyard through their Drawings. *Procedia - Social and Behavioral Sciences*, 116: 2105–2114.

She, D.; Lai, Y.; Yi, G.; and Xu, K. 2021. Hierarchical Layout-Aware Graph Convolutional Network for Unified Aesthetics Assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 8475–8484. Computer Vision Foundation / IEEE.

Sheng, K.; Dong, W.; Ma, C.; Mei, X.; Huang, F.; and Hu, B. 2018. Attention-based Multi-Patch Aggregation for Image Aesthetic Assessment. In Boll, S.; Lee, K. M.; Luo, J.; Zhu, W.; Byun, H.; Chen, C. W.; Lienhart, R.; and Mei, T., eds., *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, 879–886. ACM.

Strezoski, G.; and Worring, M. 2017. OmniArt: Multi-task Deep Learning for Artistic Data Analysis. *CoRR*, abs/1708.00684.

Wang, J.; Yang, C.; Xu, Y.; Shen, Y.; Li, H.; and Zhou, B. 2022. Improving GAN Equilibrium by Raising Spatial Awareness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 11275–11283. IEEE.

Wei, C.; Fan, H.; Xie, S.; Wu, C.; Yuille, A. L.; and Feichtenhofer, C. 2022. Masked Feature Prediction for Self-Supervised Visual Pre-Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 14648–14658. IEEE.

Zhu, H.; Li, L.; Wu, J.; Zhao, S.; Ding, G.; and Shi, G. 2022. Personalized Image Aesthetics Assessment via Meta-Learning With Bilevel Gradient Optimization. *IEEE Transactions on Cybernetics.*, 52(3): 1798–1811.