DA3Attacker: A Diffusion-based Attacker against Aesthetics-oriented Black-box Models

Shuai He, *Member, IEEE*, Shuntian Zheng, Anlong Ming*, *Member, IEEE*, Yanni Wang, Huadong Ma, *Fellow, IEEE*

***Abstract*—The adage "Beautiful Outside But Ugly Inside" resonates with the security and explainability challenges encountered in image aesthetics assessment (IAA). Although deep neural networks (DNNs) have demonstrated remarkable performance in various IAA tasks, how to probe, explain, and enhance aesthetics-oriented "black-box" models has not yet been investigated to our knowledge. This lack of investigation has significantly impeded the commercial application of IAA. In this paper, we investigate the susceptibility of current IAA models to adversarial attacks and aim to elucidate the underlying mechanisms that contribute to their vulnerabilities. To address this, we propose a novel diffusion-based framework as an attacker (DA3Attacker), capable of generating adversarial examples (AEs) to deceive diverse black-box IAA models. DA3Attacker employs a dedicated Attack Diffusion Transformer, equipped with modular aesthetics-oriented filters. By undergoing two unsupervised training stages, it constructs a latent space to generate AEs and facilitates two distinct yet controllable attack modes: restricted and unrestricted. Extensive experiments on 26 baseline models demonstrate that our method effectively explores the vulnerabilities of these IAA models, while also providing multi-attribute explanations for their feature dependencies. To facilitate further research, we contribute the evaluation tools and four metrics for measuring adversarial robustness, as well as a dataset of 60,000 re-labeled AEs for fine-tuning IAA models. The resources are available here.**

***Index Terms*—Image Aesthetics Assessment, Adversarial Attack, Deep Learning.**

I. Introduction

"*Everyone desires beauty.*" From the early focus on factors related to compression, transmission, and image processing [1], to directly addressing imaging measurements of user-generated content quality [2] (e.g., photos and videos taken with smartphones), and moving on to the recently popular AI-Generated Content (AIGC) [3] and the self-media market [4], at every stage, accurately evaluating visual aesthetics remains an indispensable need to the computer vision field. However, most existing image aesthetics assessment (IAA) studies have primarily focused on employing deep learning techniques to learn aesthetic perception, while neglecting the problems posed by black-box deep learning methods.

Security Concerns. Most existing IAA models appear to be vulnerable to adversarial attacks in real-world scenarios, especially when confronted with adversarial examples (AEs) designed to mislead them (Fig. 1). This susceptibility arises from two primary factors: 1) IAA models often have to train on imperfect datasets that suffer from sample bias, long-tail distribution, and limited size. Consequently, these models fail to achieve generalization and robustness for IAA. This limitation weakens the models' ability to handle scarce, anomalous or adversarial images. 2) A lack of rigorous investigation of IAA models' vulnerabilities and adversarial robustness. Specifically, concerns arise regarding their sensitivity, output dependencies, and application limits.

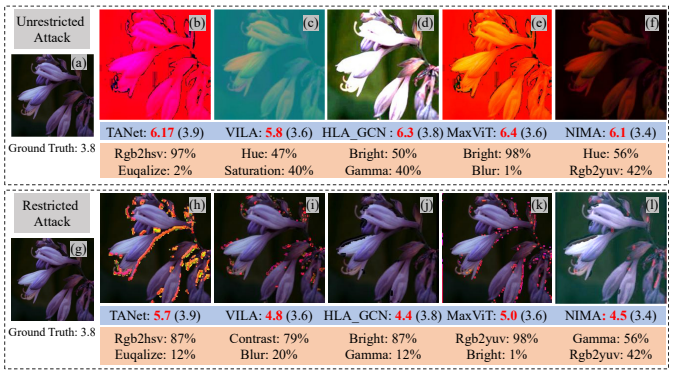


Fig. 1. The AEs of unrestricted (b-f) and restricted (h-l) attack on the target IAA models. show the predicted scores of each AE and each original image by a target IAA model. show the Top-2 adopt filters in the corresponding AE.

Explainability Concerns. Understanding the outcomes generated by IAA models can be challenging, especially when they yield unexpected or anomalous results. This challenge arises from two primary factors: 1) The "black-box" nature of IAA models makes it difficult to comprehend the features they rely on for assessment, as well as the methods they employ to accomplish their tasks. This lack of transparency can lead to spurious correlations and misalignment with human expectations or intuitions. 2) Most IAA model designers are not experts in photography aesthetics, which leads them to rely on datasets with human-annotated labels for model training. These labels, however, can be highly subjective and introduce noise into the training data, resulting in unpredictable and unconvincing outcomes from data-driven IAA models.

To address these concerns, we propose a novel diffusion-based framework as an attacker (DA3Attacker), aiming to identify vulnerabilities in IAA models through adversarial attacks and elucidate the rationale behind their assessments. Specifically, we employ 14 differentiable filters with aesthetics-oriented characteristics as attacking strategies and integrate them with a dedicated Attack Diffusion Transformer (ADT) to generate adversarial examples (AEs). These AEs serve as probes to expose security vulnerabilities in IAA

S. He, S. Zheng, A. Ming, Y. Wang and H. Ma are with School of Computer Science, Beijing University of Posts and Telecommunications (e-mail:{hs19951021; zhengshuntian; mal; yanni.wang; mhd}@bupt.edu.cn).

TABLE I
SUMMARY OF 26 MODELS ON THE AVA DATASET. WE ONLY COUNT 14 SOTA IAA MODELS WITH PUBLICLY AVAILABLE OFFICIAL CODE. ADDITIONALLY, WE HAVE RETRAINED 12 MODELS OF OTHER VISION TASKS (INDICATED BY '*') THAT DEMONSTRATED OUTSTANDING PERFORMANCE ON IAA TASKS. THE REGRESSION PERFORMANCE (PREDICT MOS) IS MEASURED BY THE SPEARMAN RANK CORRELATION COEFFICIENT (SRCC) AND THE LINEAR CORRELATION COEFFICIENT (LCC). FOR BINARY CLASSIFICATION ACCURACY (AESTHETICALLY NEGATIVE OR POSITIVE), WE ADOPT THE METRIC ACC.

No.	Model	Years	Pub.	Method	Acc↑	SRCC↑	LCC↑
1	AADB [5]	2016	ECCV	photo aesthetics rank CNN	0.77	0.56	0.58
2	ALamp [6]	2017	CVPR	Adaptive Layout-Aware Multi-Patch CNN	0.82	0.67	0.67
3	NIMA [7]	2018	TIP	score distribution predicting CNN	0.81	0.61	0.64
4*	SE_Net [8]	2018	CVPR	channel-wise feature recalibrating block	0.78	0.65	0.66
5	U_IAA [9]	2019	TIP	unified IAA statistical framework	0.80	0.72	0.72
6	MLSP [10]	2019	CVPR	staged training,multi-level features	0.79	0.76	0.76
7*	GhostNet [11]	2020	CVPR	intrinsic features revealing layers	0.78	0.65	0.68
8	Relic [12]	2020	CVIU	a fully connected graph based on deep CNN	0.82	0.75	0.76
9*	Swin [13]	2021	CVPR	shifted windows transformer	0.81	0.73	0.76
10	MUSIQ [14]	2021	ICCV	multi-scale image quality transformer	0.81	0.73	0.74
11*	Coat-Net [15]	2021	CVPR	co-scale conv-attentional transformers	0.75	0.54	0.52
12*	DeiT [16]	2021	ICML	competitive convolution-free transformer	0.79	0.67	0.69
13	KonIQ++ [17]	2021	BMVC	refining task-specific features framework	0.79	0.69	0.70
14*	SAMP [18]	2021	BMVC	saliency-augmented multi-pattern pooling	0.79	0.70	0.71
15	MaxVit [19]	2022	ECCV	multi-axis vision transformer	0.81	0.71	0.75
16*	DAT [20]	2022	CVPR	ViT with deformable attention	0.81	0.74	0.74
17*	ConvNext [21]	2022	CVPR	pure ConvNet hierarchical	0.81	0.74	0.73
18	hyperIQA [22]	2022	CVPR	self-adaptive hyper network architecture	0.78	0.65	0.66
19*	VCRnet [23]	2022	TIP	visual compensation restoration network	0.79	0.69	0.70
20	GraphIQA [24]	2022	IEEE TMM	distortion graph representationlearning	0.78	0.61	0.62
21*	GAT [25]	2022	ICPR	feature-semantic two-stage GNN	0.82	0.75	0.76
22*	EdgeNext [26]	2022	ECCV	split depth-wise transpose attention	0.79	0.68	0.70
23*	TReS [27]	2022	WACV	GPU-Dedicated architecture	0.80	0.73	0.72
24	HLA_GCN [28]	2022	CVPR	layout-Aware graph network	0.83	0.67	0.69
25	TANet [29]	2022	IJCAI	theme-oriented baseline model	0.82	0.76	0.77
26	VILA [30]	2023	CVPR	image-comment aesthetic semantics network	0.82	0.77	0.77

models. To improve the efficiency of generation, ADT incorporates domain-specific knowledge as conditions to construct adaptive latent space representations for each filter. Moreover, unsupervised pre-training further enhances ADT's comprehension of how each filter alters an image's aesthetic qualities. We also employ a lightweight Filter Coordinator to execute restricted (less perceptible) or unrestricted (more aggressive) attacks, which applies regularization to achieve the desired attacks and allocates weights to coordinate the various filters. These weights reveal the dependencies of IAA models on particular features and offer multi-attribute explanations for the generated assessment scores. In summary, our contributions are as follows:

- This is the first time, to our knowledge, that security vulnerabilities have been revealed in Aesthetics-oriented Black-box Models while also clarifying their underlying reasons for abnormal behaviors. Additionally, our work offers multi-attribute explanations and reveals feature dependencies in these models.
- DA3Attacker is proposed, which includes 14 aesthetics-oriented attack filters, two stages of unsupervised training, and the dedicated Attack Diffusion Transformer. Our framework efficiently generates AEs from latent spaces, enabling both restricted and unrestricted attacks.
- Our evaluation of adversarial robustness encompasses results from 26 baseline models, affirming the effectiveness of DA3Attacker. Consequently, our benchmark stands as the most comprehensive in the field of IAA to date. Additionally, we have curated the Aesthetics Attack Adversarial Examples (3AE) dataset to fortify defenses against adversarial attacks.

Our three contributions constitute a comprehensive benchmark suite equipped with essential tools for evaluating adversarial robustness, which is expected to stimulate further research in this direction and provide deeper insights into explainable IAA.

II. RELATED WORK

A. Image Aesthetics Assessment

General IAA includes three types of tasks: binary classification (aesthetically positive or negative) [31], [32], aesthetic score regression [33], [34], and score distribution prediction [1], [35]–[38]. In contrast, personalized IAA adapts a generic aesthetics model for individual preference [39]–[41]. These methods (Table I) learn from human labeled IAA datasets where images are paired with aesthetic ratings, and most models are trained to regress towards **the mean opinion scores (MOS)**. To our knowledge, most existing methods emphasize assessment performance, while typically overlooking the security and explainability issues of IAA.

B. Adversarial Attacks

There are generally two categories of adversarial attacks: white-box and black-box attacks [42], [43]. White-box attacks

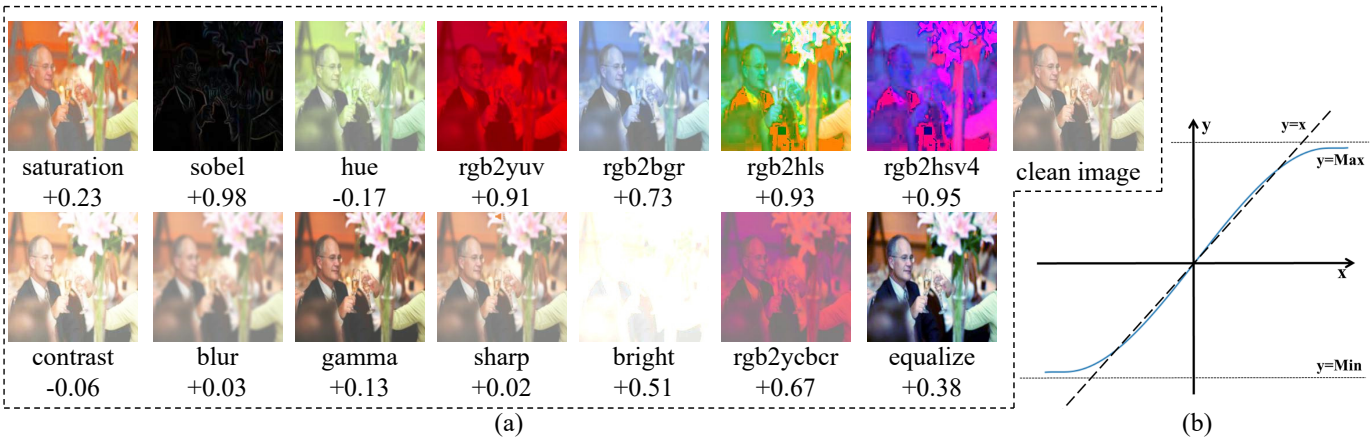


Fig. 2. The visualization of our filter functions and normalized function. (a) Results of the clean image after being processed by 14 filters with certain parameters; (b) Parameter scaling function.

have full access to the information of target models, like model structure and weights [44], [45]. In contrast, black-box attacks [46]–[48] only have access to the inputs and outputs of the target system without requiring knowledge of the model's internal structure or parameters.

Despite the abundance of research in classic computer vision tasks, there is currently no corresponding work in the IAA field, and adversarial attacks in IAA present unique challenges. Firstly, IAA is a highly subjective task, lacking a clear objective criterion. Beyond the model being a black box, even the meaning of the ground truth is obscured. This dual ambiguity makes defining an effective attack AE particularly challenging. Secondly, image aesthetics are affected by multiple attributes which make the model highly sensitive to even minor changes in these attributes. For instance, a minor variation in image brightness can significantly affect the aesthetic score given by the model. As a result, the search space for AEs becomes vast.

C. Model Explainability

One previous definition of explainability as **the ability to provide explanations in understandable terms to a human** [49], [50]. While in the field of Deep Neural Networks (DNNs), explainability refers to explaining the internal structure [51], feature dependency [52], receptive fields [53], [54], etc., of a learning model to humans.

Compared to others, IAA tasks necessitate more comprehensive explanations of the model's outputs due to their inherent subjectivity. Outputs that are difficult to explain may contradict human intuition, thereby posing challenges in establishing trust in these models. However, to the best of our knowledge, previous research has not yet to offer explanations for the outputs generated by IAA models.

D. Diffusion Models

Diffusion models [55]–[58] have recently become a popular method for learning the data distribution, particularly for generative tasks. The core concept of diffusion models is to start with features sampled from a Gaussian noise distribution, and then iteratively denoise and deform the feature distribution until it converges to the original distribution.

The proposed ADT distinguishes itself from conventional diffusion techniques in two key aspects. **1)** Instead of generating noise from the Gaussian distribution, our method overlays perturbations to clean images by specific filters. This approach empowers our framework to control perturbations, identify specialized noise, and recycle it to yield AEs. **2)** Our strategy adopts a guidance prediction method that operates independently of a classifier [59]. We incorporate prior information into various filters to improve the effectiveness of AE generation.

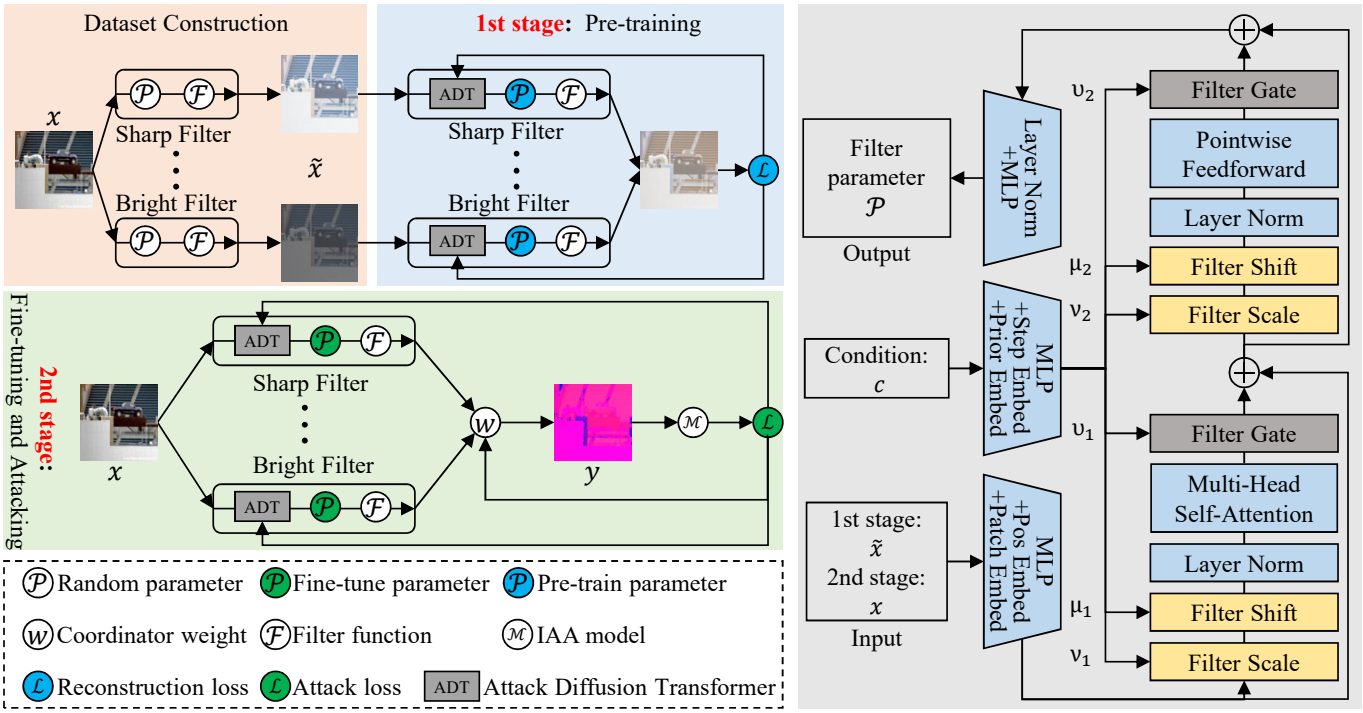
III. DA3Attacker

To generate adversarial examples effectively using various filters, our framework must **1)** understand the distinct impact of each filter on image features, and **2)** establish a mapping relationship between adversarial samples and images that are deemed to possess high aesthetic value by targeted models.

Fig. 3 illustrates the pipeline of the proposed framework. Given the clean images $x \in \mathbb{R}^{H\times W\times 3}$, we adopt 14 predefined filters f to construct pre-training dataset samples $\tilde{x} \in \mathbb{R}^{H\times W\times 3}$. Each filter consists of a filter function $\mathcal{F}$ and parameter $\mathcal{P}$, where $\mathcal{P}$ is used to control $\mathcal{F}$ for editing images. Subsequently, we employ the proposed ADT as latent spaces for each filter and feed $\tilde{x}$ into ADT to generate $\mathcal{P}$ and further reconstruct the original x. The weights θ_f of f are pre-trained by the reconstruction loss (formula (3)). This stage is designed to provide a comprehensive understanding of the effects of different filters. Finally, to obtain AE $y \in \mathbb{R}^{H\times W\times 3}$, we design a Filter Coordinator that learns a weight $w \in \mathbb{R}^{1\times 14}$ to coordinate the output of the pre-trained f to generate AE y. y is then sent to the IAA model for score generation and fine-tuning w and θ_f through formula (6). This stage learns how to construct AEs by fine-tuning learned latent spaces and coordinating filter parameters to deceive different IAA models.

A. Filter Design

Within our framework, the term "filters" refers to a collection of operations that can adjust an image locally or globally,



(a) Dataset Construction and Two Training Stages (b) ADT

Fig. 3. Overview of DA3Attacker. (a) Two-Stage Training Process. **Dataset Construction**: clean images are perturbed by predefined filters (filter function with random parameters) to construct pre-training dataset; **1-st stage**: clean images are recovered from their perturbed images to pre-train the latent spaces of each filter; **2-nd stage**: the Filter Coordinator fine-tunes the framework using the learned latent spaces to generate AEs. (b) The Attack Diffusion Transformer (ADT): based on the inputs, ADT dynamically adjusts the scaling and shifting parameters of layer normalization in the latent space.

thereby altering its aesthetic qualities. To ensure effective alignment between image modifications and aesthetic features for attack purposes, it is crucial to select aesthetics-oriented filters as attack operations.

1) Selection of Filters: Incorporating filters into the training process requires that the filter function's parameters are continuous, ensuring differentiability. This paper presents 14 filters, whose specific effects are illustrated in Fig. 2 (a). Additionally, the quantity of filters influences the framework's training efficiency. A limited number of filters restricts the framework's ability to identify the best attack strategy. Conversely, having numerous filters enriches action combinations, yet increasing action dimensions poses significant training challenges.

2) Selection Principles: **Differentiable.** For a given clean image, the i-*th* filter $f_i(x;\theta_f^i)$ learns a set of corresponding network parameters θ_f^i, which necessitates differentiability of the filter parameters for gradient-based optimization of the task loss. Therefore, all filters should be modeled as fundamental neural network layers. **Specifically**, each filter learns parameters within the ADT, combined with a filter function to execute operations using these parameters. **Resolution-independent.** Varying input resolutions in targeted IAA models and applications may compromise the attack performance of a trained framework. Additionally, AEs with altered resolutions compared to their original clean images are easier to detect. To address these issues, the filter network's design should adaptively process input images of various resolutions and generate outputs at consistent resolution levels, ensuring uniform performance across different scenarios.

B. Attack Diffusion Transformer

As the core of our framework, ADT consists of four primary components (Fig. 3).

Encoder and Decoder. The encoder's first layer is the patch embedding layer, which converts the spatial input x into a sequence of $32 \times 32 \times 3$ tokens. After the patch embedding, we incorporate standard ViT frequency-based positional embeddings to all input tokens. Besides the image inputs, ADT processes two additional conditional data: time steps and filter prior information. In the decoder, layer normalization and linear layers decode the sequence of image tokens into output filter parameter $\mathcal{P}$.

Filter Prior. We have tailored dedicated prior information c for each type of filter, as elaborated in Table II. This prior information is typically derived through mathematical and statistical functions to extract initial filter characteristics.

Adaptive Latent Features. Rather than setting hyperparameters manually, we regress the scale v and shift μ parameters for each filter's latent space features using the embedding vectors of c. We then incorporate these into the adaptive layer normalization [60]. This adaptive approach not only enhances the generalization to variations in feature distributions across different latent spaces (Fig. 4), but also allows it to learn filter-specific inductive biases adeptly.

Filter Gate. Finally, within the latent space, we regress a filter gate matrix through the condition to selectively activate certain features. This facilitates focused attention on important task-related information and enables training in a lower dimensional

TABLE II
THE PRIOR INFORMATION OF EACH FILTER.

Filter	Prior information
Bright	Weighted sum of RGB channels
Contrast	The maximum difference in the pixels
Hue	Relative relationship of RGB channels
Saturation	Pixels's distance from gray
Gamma	Pixel changes in different areas of image
Sobel	Prominence of the main lines in the image
Blur	Insignificance of different areas of image
Sharp	Prominence of the main lines in the image
Equalize	Similarity in the areas of the pixels
Rgb2hsv Rgb2ycbcr Rgb2bgr Rgb2hls Rgb2yuv	Combination of bright, hue, and saturation

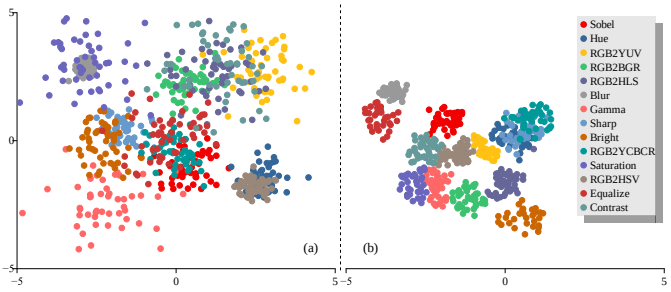


Fig. 4. Adjustment of various filter features in the latent space are achieved through our learnable scale and shift parameters. (a) Original distribution of features. (b) Adjusted distribution of features.

space with significantly improved computational efficiency. Meanwhile, outputs of the Filter Gate are normalized in a similar manner, and our normalized function is based on the Sigmoid as illustrated in Fig. 2 (b). Most input values align closely with their outputs (around y=x), with only a few outliers or anomalies adjusted to default Min or Max values.

C. Filter Coordinator

Besides assigning weights to each filter during an attack, more importantly, the Filter Coordinator employs regularization to achieve diverse attacks.

Filter Distribution. Applying only a few filters during an attack might result in underutilizing of filter types and encountering local minima. To address this, we introduce an additional term $\mathcal{R}$ that penalizes distributions of n filter weights w exhibiting high concentration and low entropy. This concept is mathematically expressed as follows:

$$\mathcal{R}(w) = \log|n| + \sum_{i=1}^{n} w_i \log w_i, \tag{1}$$

where w_i is the learnable weights assigned to the f_i filter.

Unrestricted and Restricted Attack. Our framework demonstrates remarkable flexibility in addressing two distinct tasks: restricted and unrestricted adversarial attacks, both of which are controlled by the regularizer $\mathcal{G}$:

$$\mathcal{G}(x, y) = \sum_{(i,j)\in S} \|x_{i,j} - y_{i,j}\|_2^2, \tag{2}$$

where S represents random sampled pixels from the image. When setting the hyper-parameter α to 1 in formula (6) and making S equal to x, the framework minimizes the pixel difference between clean images and AEs, effectively altering only a sparse set of pixels to achieve attacks. Conversely, when α is set to 0, it removes any constraints on the attack, thereby allowing for a more aggressive approach (the primary attacks applied in this paper).

Filter Coordinator Network. The Filter Coordinator network is designed based on a simple architecture consisting of MLPs. It acquires the initial AEs' outputs from each trained filter in the latent space and produces coordinating weights to generate the final AEs.

D. Loss Function

Pre-training Loss. The first stage of pre-training is dedicated to reconstructing the original image x from its perturbed version $\tilde{x}$. The parameters θ_f of the latent space can be optimized using a specific loss function:

$$\mathcal{L}_{rec} = \sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{3} \|x_{i,j,k} - \mathcal{F}(\tilde{x}, c; \theta_f)_{i,j,k}\|_2^2, \tag{3}$$

where $\mathcal{F}(\tilde{x}, c; \theta_f)$ is the reconstruction image, c is a condition containing time steps and filter prior information.

Fine-tuning Loss. In the second stage of fine-tuning, the primary objective is to ensure that an adversarial image y attains a higher score than the clean image x after the attack, as evaluated by a specific IAA regression model $\mathcal{M}$. The adversarial image y is defined as the output of a series of filter functions applied to the clean image x, expressed as:

$$y = \sum_{i=1}^{n} w_i \mathcal{F}_i(x, c; \theta_f^i), \tag{4}$$

where $\mathcal{F}_i$ represents the i-th filter function, w_i denotes the weight of the i-th filter, c is the context or condition of the attack, and θ_f^i are the learnable parameters of the i-th filter's latent space. The loss function for this objective is formulated as:

$$\mathcal{L}(x) = (\mathcal{M}(x) + \eta) - \mathcal{M}(y), \tag{5}$$

where η is a parameter designed to promote **the generation of AEs to achieve a higher MOS than the clean image**. By further trading off against two regularization terms $\mathcal{G}$ and $\mathcal{R}$ to control attacks, the IAA attack problem can be formulated as optimizing the following loss function:

$$\mathcal{L}_{attack} = \mathcal{L}(x) + \alpha\mathcal{G}(x, y) + \lambda\mathcal{R}(w), \tag{6}$$

where $\alpha \geq 0$ and $\lambda \geq 0$ are the regularization parameters.

TABLE III

THE ADVERSARIAL ROBUSTNESS METRICS OF 26 METHODS WERE EVALUATED UNDER **UNRESTRICTED ATTACKS** ($\alpha = 1$ IN FORMULA (6)). THE TOP-3 RESULTS ARE HIGHLIGHTED IN **RED**, **BLUE** AND **GREEN**, RESPECTIVELY. DETAILS OF METRICS ARE SHOWN IN SEC. IV-B. ↑ INDICATES THAT HIGHER VALUES OF THE METRICS CORRESPOND TO BETTER ADVERSARIAL ROBUSTNESS OF IAA MODELS. ALL MODELS UTILIZE OFFICIALLY AVAILABLE CODE AND EMPLOY OUR FRAMEWORK WITH UNIFORM PARAMETER SETTINGS TO OBTAIN THESE METRICS.

	Model	2016-2021												
		AADB	ALamp	NIMA	SE_Net	U_IAA	MLSP	GhostNet	Relic	Swin	MUSIQ	Coat-Net	DeiT	KonIQ++
Metric	BASR↓	48.3	46.3	89.1	53.8	69.4	53.2	66.6	31.1	**24.0**	63.2	83.7	**24.1**	48.4
	RASR↓	**83.7**	91.2	93.7	**83.3**	92.8	94.2	92.9	90.6	86.1	93.1	98.0	85.5	94.1
	MSI↓	0.87	0.91	0.96	0.57	0.94	0.94	1.03	0.83	**0.33**	0.91	1.30	**0.36**	0.67
	ART↑	0.23	0.46	0.78	0.91	0.83	0.03	0.31	0.48	**1.32**	**1.38**	0.93	0.94	0.46
	Model	2021-2023												
		SAMP	MaxViT	DAT	ConvNext	hyperIQA	VCRnet	GraphIQA	GAT	EdgeNext	TReS	HLA_GCN	TANet	VILA
Metric	BASR↓	62.9	76.1	38.9	**20.4**	46.7	45.8	47.4	63.1	44.7	31.1	49.4	39.7	55.9
	RASR↓	90.0	97.0	87.7	93.1	87.2	92.8	89.9	93.1	88.4	88.8	**81.1**	91.2	97.0
	MSI↓	0.96	1.13	0.66	0.41	0.73	0.91	0.94	1.13	**0.24**	0.68	0.75	1.34	0.48
	ART↑	0.78	0.91	0.83	0.03	0.31	0.48	1.31	0.52	**1.37**	0.93	0.35	0.93	0.67

TABLE IV

THE ADVERSARIAL ROBUSTNESS METRICS OF 26 METHODS WERE EVALUATED UNDER **RESTRICTED ATTACKS** ($\alpha = 0$ IN FORMULA (6) AND $S = x$ IN FORMULA (2)). THE TOP-3 RESULTS ARE HIGHLIGHTED IN **RED**, **BLUE** AND **GREEN**, RESPECTIVELY. DETAILS OF METRICS ARE SHOWN IN SEC. IV-B. ↑ INDICATES THAT HIGHER VALUES OF THE METRICS CORRESPOND TO BETTER ADVERSARIAL ROBUSTNESS. ALL MODELS UTILIZE OFFICIALLY AVAILABLE CODE AND EMPLOY OUR FRAMEWORK WITH UNIFORM PARAMETER SETTINGS TO OBTAIN THESE METRICS.

	Model	2016-2021												
		AADB	ALamp	NIMA	SE_Net	U_IAA	MLSP	GhostNet	Relic	Swin	MUSIQ	Coat-Net	DeiT	KonIQ++
Metric	BASR↓	37.8	20.3	13.2	44.2	67.1	31.2	43.4	20.3	**11.9**	16.9	12.3	15.0	**7.7**
	RASR↓	**76.6**	87.9	82.4	82.9	91.6	90.9	88.9	85.6	**67.2**	85.7	89.8	81.4	82.6
	MSI↓	0.50	0.63	0.72	0.38	0.72	0.67	0.76	0.58	**0.12**	0.41	0.87	0.32	0.27
	ART↑	0.25	0.47	0.81	0.83	0.80	0.11	0.33	0.45	**1.39**	**1.43**	0.85	0.90	0.35
	Model	2021-2023												
		SAMP	MaxViT	DAT	ConvNext	hyperIQA	VCRnet	GraphIQA	GAT	EdgeNext	TReS	HLA_GCN	TANet	VILA
Metric	BASR↓	58.6	65.1	24.0	12.0	15.5	26.4	43.4	23.9	37.2	22.3	**1.3**	23.0	51.7
	RASR↓	82.1	94.1	86.5	86.1	83.4	76.7	84.9	81.3	82.3	82.0	**49.0**	90.6	87.7
	MSI↓	0.73	0.99	0.30	0.41	0.44	0.32	0.57	0.80	**0.19**	0.44	**0.09**	0.96	0.39
	ART↑	0.69	0.95	0.86	0.12	0.50	0.48	1.25	0.58	**1.36**	0.86	0.87	0.99	0.60

IV. EXPERIMENTS

A. Experimental Settings

Target Models. We selected 26 IAA baselines (Table I) according to the following criteria: 1) classical architectures with **available code**, and 2) SOTA models in a specific IAA field, e.g., general or personalized IAA. 3) Moreover, we included other vision-based task models that perform well in IAA tasks. These baselines were trained with the recommended parameter settings (e.g., optimizer and batch size).

Dataset. 1) The dataset used for training target IAA models conducted on the AVA dataset, a widely used and large-scale IAA dataset that contains nearly 255,000 images. This dataset served as the training ground for all target IAA models to ensure optimal performance and generalization. Specifically, the dataset was divided into 235,528 images for training and 20,000 images for testing. 2) The dataset for the pre-training (1st stage) and fine-tuning (2nd stage) of our framework. Our 1st and 2nd stages do not impose any requirements on the dataset samples. We extracted 10,000 clean images from the **testing set** of AVA. By applying random perturbations to these images (Fig. 2) , we generated 100,000 perturbed images. 3) The dataset for evaluating attack performance. To evaluate the performance of our trained framework, and assess the adversarial robustness of target IAA models on its trained dataset, we extracted 10,000 clean images from the **training set** of the AVA dataset. Of these, 63% fall within the [3, 5.5] score range, and the remaining portion within the (5.5, 7] range.

Training Process. The Adam optimizer was employed throughout the model's training phase, with the learning rate initialized at a value of 2×10^{-4}. To adjust the learning rate dynamically, we utilized the ReduceLROnPlateau scheduler available in PyTorch with a patience of 4 epochs and a learning rate decrease factor of 0.1. We set the batch size to 8 with an input and output size of 224×224. The η in formula (5) is set to 3.

B. Designed Metrics

In our study, a successful attack is determined when the generated AEs outscore the clean images on the target IAA model. To reflect both the adversarial robustness of different

TABLE V
COMPARISON AMONG VARIOUS BACKBONES IN OUR FRAMEWORK.

		NIMA	HLA-GCN	MaxViT	TANet
ViT	BASR	62.7 (-30%)	48.1 (-4%)	61.8 (-19%)	30.4 (-23%)
	RASR	91.4 (-2%)	73.6 (-9%)	90.8 (-6%)	86.5 (-5%)
	MSI	0.84 (-13%)	0.63 (-16%)	0.91 (-20%)	1.31 (-2%)
UNet	BASR	53.6 (-40%)	42.8 (-13%)	57.1 (-25%)	30.1 (-24%)
	RASR	76.9 (-18%)	71.0 (-13%)	81.4 (-16)%	86.4 (-5%)
	MSI	0.63 (-34%)	0.59 (-21%)	0.83 (-27%)	1.30 (-3%)

TABLE VI
AN ABLATION STUDY CONDUCTED ON THE PRE-TRAINING APPROACH OF OUR FRAMEWORK (1ST STAGE).

	NIMA	HLA-GCN	MaxViT	TANet
BASR	81.7 (-6%)	42.4 (-14%)	70.3 (-8%)	37.0 (-7%)
RASR	93.3 (-1%)	78.5 (-3%)	94.8 (-2%)	86.5 (-5%)
MSI	0.94 (-2%)	0.65 (-13%)	1.01 (-11%)	1.23 (-1%)
ART	1.03 (+32%)	0.79 (+126%)	1.24 (+36%)	1.28 (+38%)

IAA models and the attack effectiveness of our framework, we introduce the following four metrics.

1) Binary Attack Success Rate (BASR, %). In IAA tasks, the binary classification task [1], [10] typically defines the score range [0, 5.5] as aesthetically negative and the range (5.5, 10] as aesthetically positive. Therefore, the proposed metric indicates the ratio of negative images (in the [0, 5.5] range) that, after being attacked, are misclassified as positive images. A higher BASR indicates a more significant impact on the model's **coarse-grained** predictions, especially for binary classification accuracy.

2) Regression Attack Success Rate (RASR, %). The proposed metric quantifies the proportion of clean images that exhibit enhanced scores following an attack by the framework. For RASR, any increase in score, no matter how small, is counted as an increase. A higher RASR indicates that the attacks affect the model's **fine-grained** predictions more significantly, especially for regression performance.

3) Mean Score Improvement (MSI). The proposed metric measures the average score improvement of all clean images after being attacked by the framework. A higher MSI indicates that the model is less robust to attacks.

4) Attack Reach Time (ART, hour). The proposed metric indicates the minimum duration required for our framework to reach the highest and converged RASR in targeted model attacks.

C. Attack Results

Unrestricted Attack Performance. Table III shows that our framework can generate AEs against 80% clean images (RASR>0.8) on all targeted models within an hour or less. Furthermore, the corresponding AEs increase MOS by over 0.24 (MSI>0.24) compared to the clean images. Although most unrestricted AEs suffer from a noticeable degradation in aesthetic quality from a human perspective, as illustrated in Fig. 1(b-f), existing IAA models tend to assign higher MOS to these AEs than original ones. This contradiction reveals significant flaws inherent in current IAA models.

Restricted Attack Performance. The restricted attack necessitates minimal alterations to the AEs in visual perception, enhancing their stealthiness. Table IV shows that our framework achieves an attack success rate of 80% (RASR>0.8) within a time frame of less than an hour for most models; however, the BASR and MSI metrics show a decline compared to the unrestricted attack. Nevertheless, the AEs produced by the restricted attack have the potential to inflict greater harm on IAA applications than those resulting from the unrestricted attack. These samples (Fig. 1(h-l)) pose increased challenges for prevention and detection, even for human observers.

Analysis of IAA Models' Adversarial Robustness. Based on the results of both unrestricted and restricted attacks, two observations are made: firstly, adopting a larger receptive field is proves advantageous in countering attacks. Across multiple metrics, EdgeNext [26], Swin [13], and HLA_GCN [28] demonstrate superior adversarial robustness. Specifically, EdgeNext introduces split depth-wise transpose attention to expand the receptive field, while the shifting operation of Swin ensures comprehensive coverage of local self-attention across the entire image space, thereby increasing its receptive field. HLA_GCN, on the other hand, uses FCN [61] and a larger kernel size. To evaluate the effects of different receptive fields, we selected TANet and NIMA, both of which are CNN-based models. We adjusted their receptive fields by changing the kernel size in the network. As indicated in Table XIII, adopting a larger receptive field proves advantageous against attacks.

Secondly, overfitting results in lower adversarial robustness. Interestingly, certain models not even designed for IAA tasks exhibit stronger adversarial robustness than many SOTA IAA models. Most SOTA models on IAA datasets exhibit poor adversarial robustness, with an SRCC above 0.75. In contrast, some non-SOTA models, such as Swin (SRCC = 0.73), EdgeNext (SRCC = 0.68), and HLA_GCN (SRCC = 0.67), demonstrate strong resilience against adversarial attacks. One reason for this discrepancy is that SOTA methods tend to overfit, especially when trained on datasets with a significant long-tail effect. This imbalance is evident in the AVA dataset, where the majority class has 2,700 times more samples than the minority class. As a result of this imbalance, models tend to develop prediction bias and focus more on the majority of samples, making it difficult for them to generalize well when encountering uncommon or unseen adversarial examples.

TABLE VII
PERFORMANCE OF ALTERNATIVE ATTACK METHODS COMPARED TO OURS.

		NIMA	HLA-GCN	MaxViT	TANet
BlackVIP [62]	BASR	14.3	7.7	10.1	5.5
	RASR	5.0	3.1	4.8	3.7
RLB-MI [63]	BASR	9.3	3.6	5.3	3.9
	RASR	2.7	2.1	1.7	1.1

AE Visualization. Fig. 1 presents typical AEs. The first row exhibits AEs under unrestricted attacks. Although these images have notable alterations, they obtain higher predicted

TABLE VIII
PERFORMANCE OF DIFFERENT VARIATIONS.

Modules	BASR	RASR	MSI	ART
no Filter Shift	33.2	80.7	1.20	0.98
no Filter Scale	34.3	83.1	1.27	0.96
no Filter Gate	31.8	79.9	1.15	1.01
no Condition	29.4	73.2	1.06	1.28

MOS than the clean images. The second row displays AEs under restricted attacks, where only individual pixels were modified. We deliberately chose AEs with alterations that remained visible compared to clean images. These results indicate that both IAA and non-IAA models are susceptible to vulnerabilities and potential flaws in IAA tasks. These visible AEs may reduce people's trust in the models, while invisible AEs could be used to attack or deceive IAA-related applications.

D. Ablation Study

Different Backbone. To evaluate the performance of different backbones within our framework, we used two well-established models: the Transformer-based ViT [64] and the CNN-based UNet [65]. Compared to our ADT, the integration of these backbones into our framework, as shown in Table V, resulted in a significant decrease in attack efficacy. This decline is attributed to the backbones' limited capacity to effectively understand and utilize filter-related features.

Pre-training Stage. Table VI provides validation for the initial pre-training stage of our framework. Omitting the initial pre-training stage, we observed a significant decrease in our framework's effectiveness, attributable to insufficient knowledge of how filters impact features.

Different Attack Method. To our knowledge, no other methods are specifically designed to attack IAA models. Thus, we adopt two relatively general attacks, namely BlackVIP [62] and RLB-MI [63], as alternatives to our method. Table VII shows the distinct advantages achieved by our method under unrestricted attacks. This is mainly because IAA tasks are highly sensitive to image pixels, and the output of the attacked IAA model is unpredictable. There's a significant challenge in developing an effective attack method without a comprehensive understanding of how pixel tampering affects global aesthetic features, compounded by the absence of implanted prior knowledge.

Different Component. To verify the impact of key components in ADT on attack performance, we use the attack on TANet as an example. Ablation studies presented in Table VIII illustrate the architectural variations of our attack framework. The results indicate that each component significantly influences attack performance. Specifically, Filter Shift, Scale, and Gate primarily affect the flexibility of feature changes and robustness to variations in feature distributions. On the other hand, the absence of the condition markedly diminishes the ability to perceive the correlation between feature changes and aesthetic alterations, leading to a significant decline in performance.

Different Dataset. We also conduct adversarial robustness evaluations on two other representative and challenging datasets: TAD66K [29] and FLICKR-AES [39]. TAD66K, the largest theme-oriented general aesthetic dataset, features refined annotations with distinct evaluation criteria for each theme. FLICKR-AES, the largest personalized aesthetic dataset for IAA tasks, focuses on individual aesthetic preferences and involves few-shot learning for personalization. Within less than an hour, our method effectively attacks two representative IAA methods, theme-oriented TANet [29] and personalized BLG-PIAA [66], as shown in Table IX. This demonstrates its generalizability across various datasets.

Furthermore, to evaluate the adversarial transferability of our framework, we conducted cross-dataset assessments. In these experiments, the ground truth of both datasets was normalized to the same score range. The targeted models and the attack framework were trained using the training and testing sets from the corresponding datasets. As shown in Table X, the attack capability of the framework remains largely unaffected when transferred across different datasets. For the targeted models, training on the larger TAD66K dataset enhances adversarial robustness.

TABLE IX
ADVERSARIAL ROBUSTNESS ON TAD66K AND FLICKR-AES.

Dataset	Method	RASR	MSI	ART
TAD66K [29]	BLG-PIAA [66]	95.1	1.57	0.74
	TANet [29]	87.3	1.12	0.90
FLICKR-AES [39]	BLG-PIAA [66]	96.0	1.15	0.49
	TANet [29]	90.7	1.36	0.77

TABLE X
CROSS-DATASET EVALUATIONS OF ADVERSARIAL ROBUSTNESS ON TAD66K AND FLICKR-AES.

Cross-dataset (RASR)	Training Dataset	Testing Dataset	
		TAD66K	FLICKR-AES
BLG-PIAA	TAD66K	95.1	94.3
	FLICKR-AES	95.6	96.0
TANet	TAD66K	87.3	86.4
	FLICKR-AES	87.8	90.7

Setting of η. The parameter η in formula (5) represents the maximum improvement score we expect the framework to achieve on a IAA model. Compared to the optimal setting (η = 3), Table XI indicates that a small η significantly restricts the framework's attack capacity. However, this results in shorter convergence time. Conversely, a large η exceeds the maximum improvement potential of the framework without noticeable impact on results but leads to longer convergence time.

E. Enhancing Adversarial Robustness

We analyzed successful attack samples across 26 models and curated the Aesthetics Attack Adversarial Examples (3AE) dataset, containing 60,000 images (Fig. 6). Experts were invited to reassess the aesthetic scores of these adversarial examples, allowing us to quantify the perceptual impact of adversarial attacks. Our results indicate that 96.9% of the adversarial examples show a substantial decline in aesthetic quality

TABLE XI
COMPARISON OF DIFFERENT η IN FORMULA (5) WITH THE DEFAULT VALUE OF η = 3.

		NIMA	HLC-GCN	MaxViT	TANet
η=0.5	BASR	48.6 (-46%)	26.8 (-46%)	33.4 (-56%)	1.7 (-96%)
	RASR	84.0 (-10%)	73.1 (-10%)	82.7 (-15%)	76.9 (-16%)
	MSI	0.24 (-75%)	0.24 (-68%)	0.42 (-63%)	0.28 (-79%)
	ART	0.32 (-59%)	0.27 (-22%)	0.65 (-29%)	0.68 (-27%)
η=1.5	BASR	89.0 (-1%)	48.9 (-1%)	69.8 (-8%)	37.7 (-5%)
	RASR	93.7 (0%)	81.0 (-0.1%)	95.9 (-1%)	90.1 (-1%)
	MSI	0.96 (0%)	0.75 (0%)	1.10 (-3%)	1.28 (-5%)
	ART	0.63 (-19.2%)	0.29 (-17%)	0.83 (-9%)	0.82 (-12%)
η=5	BASR	89.1 (0%)	49.1 (-0.6%)	76.2 (+0.1%)	39.9 (+0.5%)
	RASR	94.2 (+0.5%)	81.2 (+0.1%)	97.5 (+0.5%)	91.2 (0%)
	MSI	0.96 (0%)	0.75 (0%)	1.13 (0%)	1.34 (0%)
	ART	1.03 (+32%)	0.60 (+71%)	1.22 (+34%)	1.24 (+33%)

TABLE XII
THE ADVERSARIAL ROBUSTNESS CONTRIBUTED BY THE PROPOSED FINE-TUNED 3AE DATASET.

	NIMA	HLA_GCN	MaxViT	TANet
BASR	2.2 (-98%)	2.3 (-95%)	1.4 (-98%)	2.6 (-94%)
RASR	6.7 (-93%)	7.3 (-91%)	2.6 (-93%)	5.4 (-94%)
MSI	-0.67 (-169%)	-0.36 (-148%)	-0.31 (-127%)	-0.53 (-140%)

relative to clean images, though approximately 3.1% exhibit aesthetic enhancement from a visual perspective. Leveraging dataset, we propose two ways to enhance the adversarial robustness of IAA models. 1) By employing only AEs as negative samples, we train a lightweight network based on Mobilenetv2 [67] with a mere parameter count of 0.6M for deployment in front-end IAA applications. Following training on 3AE, this network achieves an accuracy rate exceeding 98% in AE identification. Further augmentation of data may elevate this accuracy even higher. 2) Through simple aesthetics annotations applied to the 3AE dataset followed by fine-tuning of IAA models, our results presented in Table XII demonstrate significant improvements in their adversarial resilience. By implementing these methodologies, our community can enhance the adversarial robustness of IAA applications.

F. Model Explanation

Through white-box attacks, our framework can expose the reliance of black-box models on specific features. We provide explanations from following aspects.

Explanation of MOS for a Single Image. When predicting an image's aesthetic MOS, does the model identify attributes where the image either lacks or excels in terms of aesthetic quality? To offer explanations from the perspective of a targeted IAA model, we make the following assumption: during the generation of the corresponding AE, deficient attributes will be significantly modified to optimize the MOS. In contrast, attributes that already demonstrate strong aesthetics will undergo only minor adjustments.

Based on this assumption, in our framework, the filter weights of the Filter Coordinator indicate the extent of modification for each filter, based on the assumption that a higher weight w_i suggests poor performance of the image in attributes related to that particular filter f_i. For example, as shown in Fig. 1, the VILA model (c) suggests that the color of the image (a) lacks aesthetics; hence, Hue and Saturation Filters are dominate. However, the HLA_GCN model (d) indicates insufficient brightness in the image. A more concrete example is shown in Fig. 5, where $(1 - w_i)\%$ of target models are plotted during the AE generation.

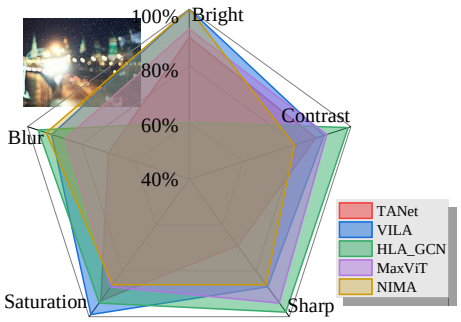


Fig. 5. The attacks allow us to indirectly infer the target model's perception of multiple attributes. If a significant proportion of w_i is assigned to generate AEs with high MOS, it suggests that the target model considers the aesthetics associated with f_i as subpar.

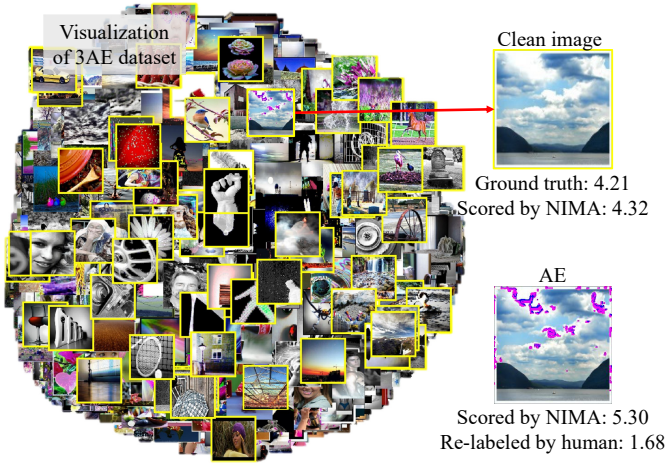


Fig. 6. The visualization of our 3AE dataset involves the manual re-labeling of all AEs.

To quantify the effectiveness of our explainability approach, we selected 1,000 images from the AVA dataset and identified the top-3 attributes that our framework considered most impactful on aesthetic quality perception (from six attributes: brightness, contrast, sharpness, saturation, and blur). Three experts were invited to evaluate these rankings for consistency with human perception, and a prediction was deemed correct if at least two experts agreed with the ranking. Subjective experimental results revealed a prediction accuracy of 76% across the 1,000 images, indicating our framework has a certain capacity for explainability.

Explanation for the Model's Aesthetic Preference. Do specific models have aesthetic preferences or heightened sensitivity to certain features? To investigate this, we examined the most frequently applied filters in the AE generation process

TABLE XIII
COMPARISON OF DIFFERENT KERNEL SIZE WITH THE DEFAULT VALUE OF NIMA AND TANET.

	kernel size	BASR ↓	RASR ↓	MSI ↓	ART ↑
NIMA	× 0.5	93 (+4%)	96 (+3%)	1.20 (+25%)	0.62 (-21%)
	× 1.5	80 (-10%)	84 (-10%)	0.83 (-14%)	0.88 (+13%)
	× 2.0	77 (-14%)	80 (-15%)	0.79 (-18%)	0.89 (+14%)
TANet	× 0.5	46 (+16%)	94 (+3%)	1.52 (+13%)	0.81 (-13%)
	× 1.5	32 (-19%)	80 (-12%)	1.14 (-15%)	0.99 (+6%)
	× 2.0	30 (-24%)	77 (-16%)	1.09 (-19%)	1.06 (+14%)

on the AVA dataset. We find that over 40% of IAA models prefer color-related filters, while approximately 20% favor contrast-related filters and about 25% tend to use brightness-related filters. These models heavily prioritize an image's aesthetics in these attributes, exceeding what would be deemed acceptable by human standards. Biases towards these attributes are evident. More details are shown in Table XIV.

TABLE XIV
TOP-2 FILTERS APPLIED IN THE AE GENERATION ON THE AVA DATASET.

No.	Model	First	Second
1	AADB	Bright	RGB2HLS
2	ALamp	RGB2HSV	Equalize
3	NIMA	RGB2HLS	Bright
4	SE_Net	RGB2HSV	Sobel
5	U_IAA	Bright	RGB2YCBCR
6	MLSP	RGB2HLS	Bright
7	GhostNet	RGB2HLS	Contrast
8	Relic	Sobel	Blur
9	Swin	Blur	Gamma
10	MUSIQ	Contrast	RGB2HSV
11	Coat-Net	Equalize	Bright
12	DeiT	Sobel	Contrast
13	KonIQ++	Hue	Bright
14	SAMP	Equalize	RGB2HSV
15	MaxVit	Bright	Gamma
16	DAT	Bright	Saturation
17	ConvNext	RGB2YCBCR	Gamma
18	hyperIQA	RGB2HSV	Contrast
19	VCRnet	RGB2HSV	Equalize
20	GraphIQA	Bright	RGB2HSV
21	GAT	Equalize	Bright
22	EdgeNext	Hue	RGB2HSV
23	TReS	RGB2HLS	Equalize
24	HLA_GCN	Blur	RGB2HLS
25	TANet	RGB2HSV	Sharp
26	VILA	RGB2HSV	RGB2HLS

G. Discussion

The adversarial robustness issue on real content platforms. The term "aesthetic economy" emphasizes the significant role of aesthetics in various industries such as photography, short videos, advertising, art, crafts and clothing. In these non-material goods sectors, beauty often holds the primary value. However, one obstacle to its commercial use is the vulnerability to adversarial attacks that can cheat IAA systems for various purposes. Two evaluation instances on real content platforms are denoted as follows: 1) ***Huawei HiAI*** (link). In photography and photo management platforms like Huawei HiAI on HarmonyOS, IAA methods are already utilized. We discovered a 69.3% success rate in RASR attacks, indicating a significant risk of economic impact due to decreased user trust and lower app ratings. 2) ***AIGC applications***, such as Stable Diffusion (link), rely on built-in IAA methods to filter generated images. However, almost half of the unsatisfactory and unfiltered generations exhibit artifacts and over-exposure that align with our AEs' characteristics.

The differences between bias and adversarial attack. In fact, a biased IAA model is reasonable due to aesthetic diversity and the bias is consistent. However, adversarial attacks cause disasters not because of bias, but rather due to irregularity and outright errors. Moreover, adversarial attacks can also expose sensitive aesthetic attributes in black-box models. Thus, adversarial attacks for IAA are essential and not directly related to bias.

Alternative attack method: To further validate the attack capability of our framework, we tested an additional attack objective, which involves assigning a low score to an image of high aesthetic quality after an attack. In this case, our framework achieved an attack success rate exceeding 97% on BASR and RASR across several representative models, including TANet, MaxViT, HLA-GCN, and NIMA. While this task is relatively simple, it further highlights the adversarial robustness issues in existing IAA models.

The potential of DA3Attacker for aesthetic enhancement: During the re-labeling of the 3AE dataset, we discovered that around 3.1% of the examples demonstrate aesthetic enhancement from a visual perspective. This indicates that, with proper adjustments, our framework may have the potential to improve image aesthetics.

V. CONCLUSIONS

This paper investigates the security and explainability of the IAA model from a novel perspective: adversarial attacks. To our knowledge, our work introduces a new research direction for the community. However, several challenges remain, e.g., how to utilize this attack framework to refine the IAA model design and enhance its performance. We hope that these contributions will provide the IAA community an opportunity to explore novel techniques in a safe environment. We have tried to cover the most important works. Nevertheless, it is impractical to thoroughly investigate all IAA models in this vast field. In the future, we will optimize our framework by incorporating more attack strategies.

VI. Acknowledgment

This work is supported by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation under Grant Number GZC20251056, and the National Natural Science Foundation of China under Grant U24B20176.

References

[1] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *TIP*, vol. 27, no. 8, pp. 3998–4011, 2018.

[2] E. Baudin, F.-X. Bucher, L. Chanas, and F. Guichard, "Dxomark objective video quality measurements," *Electronic Imaging*, vol. 2020, no. 9, pp. 166–1, 2020.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[4] W.-T. Sun, T.-H. Chao, Y.-H. Kuo, and W. H. Hsu, "Photo filter recommendation by category-aware aesthetic learning," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1870–1880, 2017.

[5] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*, 2016.

[6] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," 2017.

[7] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, aug 2018.

[8] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2019.

[9] H. Zeng, Z. Cao, L. Zhang, and A. C. Bovik, "A unified probabilistic formulation of image aesthetic assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 1548–1561, 2019.

[10] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *CVPR*, 2019.

[11] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," 2020.

[12] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell, "Representation learning via invariant causal mechanisms," 2020.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv:2103.14030*, 2021.

[14] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *ICCV*, 2021, pp. 5148–5157.

[15] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," 2021.

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[17] S. Su, V. Hosu, H. Lin, Y. Zhang, and D. Saupe, "Koniq++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects," 2021, pp. 1–12.

[18] B. Zhang, L. Niu, and L. Zhang, "Image composition assessment with saliency-augmented multi-pattern pooling," *arXiv preprint arXiv:2104.03133*, 2021.

[19] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," *ECCV*, 2022.

[20] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4794–4803.

[21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[22] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[23] Z. Pan, F. Yuan, J. Lei, Y. Fang, X. Shao, and S. Kwong, "Vcrnet: Visual compensation restoration network for no-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 31, pp. 1613–1627, 2022.

[24] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "Graphiqa: Learning distortion graph representations for blind image quality assessment," *IEEE Transactions on Multimedia*, 2022.

[25] K. Ghosal and A. Smolic, "Image aesthetics assessment using graph attention network," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 3160–3167.

[26] M. Maaz, A. Shaker, H. Cholakkal, S. Khan, S. W. Zamir, R. M. Anwer, and F. Shahbaz Khan, "Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications," in *European Conference on Computer Vision*. Springer, 2022, pp. 3–20.

[27] T. Ridnik, H. Lawen, A. Noy, E. B. Baruch, G. Sharir, and I. Friedman, "Tresnet: High performance gpu-dedicated architecture," 2020.

[28] D. She, Y.-K. Lai, G. Yi, and K. Xu, "Hierarchical layout-aware graph convolutional network for unified aesthetics assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8475–8484.

[29] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," *IJCAI*, 2022.

[30] J. Ke, K. Ye, J. Yu, Y. Wu, P. Milanfar, and F. Yang, "Vila: Learning image aesthetics from user comments with vision-language pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 041–10 051.

[31] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Eur. Conf. Comput. Vis.*, 2008, pp. 386–399.

[32] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*. Springer, 2006, pp. 288–301.

[33] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *CVPR*, 2017, pp. 722–731.

[34] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *ACMMM*, 2018.

[35] D. She, Y.-K. Lai, G. Yi, and K. Xu, "Hierarchical layout-aware graph convolutional network for unified aesthetics assessment," in *CVPR*, 2021, pp. 8475–8484.

[36] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," in *CVPR*, 2020, pp. 14 114–14 123.

[37] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks." in *IJCAI*, 2022, pp. 942–948.

[38] S. He, A. Ming, S. Zheng, H. Zhong, and H. Ma, "Eat: An enhancer for aesthetics-oriented transformers," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 1023–1032.

[39] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *ICCV*, 2017, pp. 638–647.

[40] P. Lv, J. Fan, X. Nie, W. Dong, X. Jiang, B. Zhou, M. Xu, and C. Xu, "User-guided personalized image aesthetic assessment based on deep reinforcement learning," *arXiv:2106.07488*, 2021.

[41] H. Zhong, S. He, A. Ming, and H. Ma, "Rethinking personalized aesthetics assessment: Employing physique aesthetics assessment as an exemplification," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2935–2944.

[42] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[43] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[44] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 231–10 241.

[45] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh, "On the adversarial robustness of vision transformers," *arXiv preprint arXiv:2103.15670*, 2021.

[46] X. Wang, Z. Zhang, K. Tong, D. Gong, K. He, Z. Li, and W. Liu, "Triangle attack: A query-efficient decision-based adversarial attack," in *European Conference on Computer Vision*. Springer, 2022, pp. 156–174.

[47] J. Zhang, Y. Huang, W. Wu, and M. R. Lyu, "Transferable adversarial attacks on vision transformers with token gradient regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 415–16 424.

[48] J. Zhang, J.-t. Huang, W. Wang, Y. Li, W. Wu, X. Wang, Y. Su, and M. R. Lyu, "Improving the transferability of adversarial samples by path-augmented method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8173–8182.

[49] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[50] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.
[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
[52] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui, "Attention interpretability across nlp tasks," *arXiv preprint arXiv:1909.11218*, 2019.
[53] A. Mahendran and A. Vedaldi, "Visualizing deep convolutional neural networks using natural pre-images," *International Journal of Computer Vision*, vol. 120, pp. 233–255, 2016.
[54] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[55] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.
[56] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
[57] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.
[58] W. Peebles and S. Xie, "Scalable diffusion models with transformers," *arXiv preprint arXiv:2212.09748*, 2022.
[59] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.
[60] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
[61] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[62] C. Oh, H. Hwang, H.-y. Lee, Y. Lim, G. Jung, J. Jung, H. Choi, and K. Song, "Blackvip: Black-box visual prompting for robust transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 224–24 235.
[63] G. Han, J. Choi, H. Lee, and J. Kim, "Reinforcement learning-based black-box model inversion attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 504–20 513.
[64] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[65] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
[66] H. Zhu, L. Li, J. Wu, S. Zhao, G. Ding, and G. Shi, "Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization," *TCYB*, 2020.
[67] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

Shuai He (Member, IEEE) is a postdoctoral researcher in Computer Science at Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interests include image processing and image aesthetics assessment.

Shuntian Zheng is currently pursuing a PhD at the School of Computer Science at the University of Warwick in the UK. His main research interests include human perception based on wireless signals and image quality perception tasks.

Anlong Ming (Member, IEEE) received Ph.D. degree in Beijing University of Posts and Telecommunications in 2008. He is currently a professor with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include computer vision and robot vision.

Yanni Wang is currently a Ph.D candidate majoring in Computer Science in Beijing University of Posts and Telecommunications (BUPT). Her research interests include image processing and image aesthetics assessment.

Huadong Ma (Fellow, IEEE) received the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Science (CAS), in 1995. He is a professor at the School of Computer Science, Beijing University of Posts and Telecommunications, China. His research interests include multimedia networks and systems, Internet things and sensor networks.