Study of Subjective and Objective Naturalness Assessment of AI-Generated Images

Zijian Chen, *Graduate Student Member, IEEE*, Wei Sun, *Member, IEEE*, Haoning Wu, Zicheng Zhang, *Student Member, IEEE*, Jun Jia, Ru Huang, Xiongkuo Min, *Member, IEEE*, Guangtao Zhai, *Senior Member, IEEE*, and Wenjun Zhang, *Fellow, IEEE*

***Abstract*— The proliferation of Artificial Intelligence-Generated Images (AIGIs) has greatly expanded the Image Naturalness Assessment (INA) problem. Different from early definitions that mainly focus on tone-mapped images with limited distortions (*e.g.*, exposure, contrast, and color reproduction), INA on AI-generated images is especially challenging as it owns more diverse contents and could be affected by factors from multiple perspectives, including low-level technical distortions and high-level rationality distortions. In this paper, we take the first step to benchmark and assess the visual naturalness of AI-generated images. First, we construct the AI-Generated Image Naturalness (AGIN) dataset by conducting a large-scale subjective study to collect human opinions on the overall naturalness as well as perceptions from the technical quality and rationality perspectives. AGIN verifies several insights for the first time that naturalness is universally and disparately affected by both technical and rational distortions, while its manifestations vary with different generation tasks. Second, to automatically assess the naturalness of AIGIs that align with human opinions, we propose the Joint Objective Image Naturalness evaluaTor (JOINT). Specifically, JOINT imitates human reasoning in naturalness evaluation by jointly learning technical and rationality features with several specific designs to guide model behavior from respective perspectives. Experiments demonstrate that JOINT significantly outperforms existing methods for providing more subjectively consistent results on naturalness assessment. The dataset can be accessed at https://github.com/zijianchen98/AGIN.**

***Index Terms*— AI-generated images, image naturalness assessment, image quality assessment, dataset.**

Received 12 September 2024; revised 16 October 2024 and 20 November 2024; accepted 27 November 2024. Date of publication 29 November 2024; date of current version 7 April 2025. This work was supported in part by China Postdoctoral Science Foundation (CPSF) under Grant 2023TQ0212 and Grant 2023M742298; in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20231618; in part by Shanghai Pujiang Program under Grant 22PJ1407400; and in part by the National Natural Science Foundation of China under Grant 62271312, Grant 62301316, Grant 62101325, Grant 62101326, and Grant 62132006. This article was recommended by Associate Editor Q. Xu. *(Corresponding authors: Wei Sun; Xiongkuo Min; Guangtao Zhai.)*

Zijian Chen, Wei Sun, Zicheng Zhang, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang are with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zijian.chen@sjtu.edu.cn; sunguwei@sjtu.edu.cn; zzc1998@sjtu.edu.cn; jiajun0302@sjtu.edu.cn; minxiongkuo@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn; zhangwenjun@sjtu.edu.cn).

Haoning Wu is with the S-Laboratory, Nanyang Technological University (NTU), Singapore 639798 (e-mail: haoning001@e.ntu.edu.sg).

Ru Huang is with the School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China (e-mail: huangrabbit@ecust.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3509032

I. Introduction

RECENT advancements in generative models have sparked a new craze in Artificial Intelligence-Generated Images (AIGIs), which have gained significant progress across various applications, including text-to-image generation [1], [2], [3], image translation [4], [5], [6], [7], [8], image inpainting [9], [10], image colorization [11], [12], [13], and image editing [14], [15], [16]. However, even cutting-edge models occasionally generate irrational content or technical artifacts in the image, which we refer to as the image naturalness problem. Unlike natural scene images (NSIs) that are captured from real-world scenes, AI-driven image generation harnesses neural networks to learn synthesis rules from extensive image datasets [17]. Its instability and randomness of generation mode attach AIGIs with more diverse content, leading to varying degrees of naturalness, which often requires retouching and filtering before practical use so as to avoid misleading people and negative social repercussions. Consequently, objective models for evaluating the naturalness of AIGIs are urgently needed.

Conventionally, image naturalness is described as the degree of correspondence between a real-life scene and a photograph displayed on a device based on some technical criteria (*e.g.*, exposure, color reproduction, shooting artifacts) [18], [19], [20], which has been utilized for image quality assessment (IQA) to compare and guide the optimization of systems and algorithms [21], [22], [23]. Under this theory, the images with richer details (Fig. 1(c) and Fig. 1(d)) should have notably better naturalness than the blurred image in Fig. 1(a), which is opposite to the human opinion. More recently, the emergence of AIGIs broadened the definition of image naturalness to comprise more non-technical semantic-related factors (*e.g.*, existence and context), which are normally regarded as rationality factors [17], [24]. However, it is highly subjective and its mechanism of how rationality affects human perception in image naturalness perception is still ambiguous and may be multi-dimensionally coupled.

In this paper, we make the first attempt to investigate the naturalness assessment problem of AIGIs, a new field of quality assessment with increasing attention [24], [35], [39]. We collect the AI-Generated Image Naturalness (**AGIN**) dataset, the *first-of-its-kind* dataset to study this problem. Specifically, AGIN contains 6,049 images collected from five different generative tasks with 18 model variants to ensure

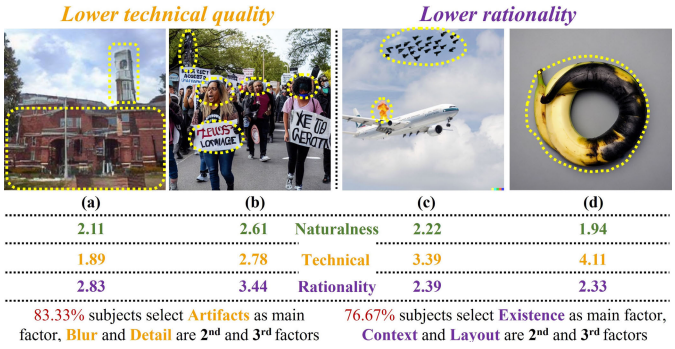


Fig. 1. An example from the proposed AGIN dataset. We highlight the regions with apparent technical or rationality distortions using dotted yellow circles. It is noticeable that although the background region is blurry and noisy in Fig.1(a), humans assign a higher naturalness score than Fig.1(d), because the banana in Fig.1(d) is more irrational. This indicates that traditional quality definition is insufficient to encompass the issues faced by AI-generated images, while multi-perspective settings can effectively avoid the perceptual bias on single absolute evaluation and provide more accurate judgments to serve as downstream supervisions.

diversity. A total of 907,350 human opinions for technical and rationality perspectives as well as their effects on the overall naturalness scores are collected from 30 participants. AGIN also provides several valuable observations for understanding human reasoning in visual naturalness. First, we find both technical distortions (*e.g.*, contrast, blur, and generative artifacts) and rationality distortions (*e.g.*, existence, color, and layout) can affect visual naturalness significantly. Second, we notice that most factors in the two perspectives are relevant, but have disparate impacts on the image naturalness, which can result in a biased naturalness assessment when applying traditional IQA models supervised by a single label. Furthermore, we observe that the overall naturalness score can be well-approximated by a weighted sum of the technical and rationality scores ($\text{MOS} = 0.145\text{MOS}_\text{T} + 0.769\text{MOS}_\text{R}$), which indicates that joint learning from technical and rationality perspectives can be a feasible way to predict naturalness.

Based on the AGIN dataset, we propose the Joint Objective Image Naturalness evaluaTor (**JOINT**), an objective naturalness assessment method that mimics human reasoning of image naturalness by jointly learning on both technical and rationality branches. Specifically, given the different characteristics of each branch, we elaborate several designs including perceptual artifacts-guided patch partition, deep feature regularization, and pre-training, to allocate each branch with corresponding learning interests. Two different supervision schemes including using the overall naturalness scores (JOINT) and the respective scores for each perspective (JOINT++) are applied to train the model. Subsequently, we use an effective subjectively-inspired weighting strategy that integrates the predictions of two branches to compute the overall naturalness score. With these designs, the proposed JOINT and JOINT++ not only reach better accuracy on the overall naturalness prediction but also provide more reliable results from technical and rationality perspectives.

Our contributions can be summarized as follows:

- We contribute the AGIN dataset (6,049 images), the first INA dataset for AIGIs that focuses on five prevalent generative tasks and contains 907,350 subjective opinions from technical quality and rationality perspectives as well as their effects on overall naturalness scores.
- We comprehensively analyze the human sensitivity and perceiving differences at various naturalness distortion types and further investigate the influence of technical and rationality factors on image naturalness.
- We propose the JOINT, an objective naturalness evaluator for AIGIs that models human perception of naturalness by brain-inspired joint learning from technical and rationality perspectives.
- Extensive experiments on AGIN and its subsets show that JOINT and JOINT++ outperform existing IQA methods by a large margin. Ablation experiments further demonstrate the effectiveness of the view decomposition strategy and other specific designs.

II. Related Work

A. Image Naturalness Assessment

Different definitions of image naturalness have been given. In the early days, naturalness was investigated by varying the colorfulness, saturation, and hue of color images of natural scenes at various lightness levels [18], [41]. Cadfk and Slavík [19] revealed that there exist high correlations between naturalness and image attributes, especially luminance and contrast. In [20], naturalness is defined as the degree of correspondence between an imaging device's captured photo and memories of real-life scenes, where human skin, grass, and sky were used as memory colors to express the perceived naturalness differences. Wang et al. [42] proposed that global naturalness is restrained by both high- and low-frequency information as well as illumination and reflectance, which are inherently related to the local details. Later in [23], naturalness is defined based on artifacts induced by some image processing methods (*e.g.*, halos, blur, and lost details) and on the individual feeling (*e.g.*, memory, opinion, and background). Recently, Le et al. [43] discussed the potential relationships between image aesthetics, quality, and naturalness. Empirical studies have demonstrated that while there exists a moderate correlation between image aesthetics and naturalness, these attributes represent distinct facets of image quality. In addition, there are some naturalness features have been proposed for image quality assessment (IQA) in few literature. Gu et al. [21] considered naturalness in tone-mapped images as the fitness of the standard deviation and the mean of pixel values to a Gaussian function and a Beta probability density function, which was computed based on statistics with 3,000 natural scene images (NSI). Liu et al. [44] characterized the naturalness variation through the distribution variations of the locally mean subtracted and contrast normalized (MSCN) coefficients and the products of pairs of the adjacent MSCN coefficients so as to quantify the image quality degradation. In another study, Yan et al. [22] integrated the naturalness degree prediction task into the deep learning-based IQA task to enhance the representation and generalization abilities of models.

In such an era dominated by generative models, AIGIs have emerged as another significant source of visual content. Most

TABLE I
COMPARISON OF THE PROPOSED AGIN DATASET WITH TRADITIONAL NATURAL SCENE AND AI-GENERATED IMAGE QUALITY DATASETS

DATASET	IMAGE SOURCE	# CONTENT	# IMAGE	# RATING	PERSPECTIVE	DISTORTION
LIVE (2006) [25]	Kodak test set	30	779	25,000	Quality	5 Artificial
TID2008 (2008) [26]	Kodak test set	25	1,700	56,100	Quality	17 Artificial
CSIQ (2009) [27]	Kodak test set	30	866	≈6,062	Quality	6 Authentic
TID2013 (2013) [28]	Kodak test set	25	3,000	524,340	Quality	24 Artificial
LIVEC (2015) [29]	Camera	1162	1162	≈350,000	Quality	15 Authentic
WED (2017) [30]	Internet	4,744	94,880	–	Quality	4 Artificial
MDID (2017) [31]	Internet	20	1,600	≈56,000	Quality	5 Artificial
KADID-10k (2019) [32]	Internet	81	10,125	303,750	Quality	25 Artificial
KonIQ-10k (2020) [33]	Multimedia	10,073	10,073	1,208,760	Quality	– Authentic
PIPAL (2020) [34]	Internet	250	29,000	≈1,130,000	Quality	40 Artificial
AGIQA-1k (2023) [35]	AI-generated	1,080	1,080	23,760	Quality	2 Generative
AIGCIQA2023 (2023) [36]	AI-generated	2,400	2,400	201,600	Quality, Authenticity, Correspondence	6 Generative
†Pick-a-Pic (2023) [37]	AI-generated	500,000	500,000	500,000	Preference	3 Generative
ImageRewardDB (2023) [38]	AI-generated	136,892	136,892	410,676	fidelity, alignment, overall	1 Generative
AGIQA-3k (2023) [39]	AI-generated	2,982	2,982	125,244	Quality, Alignment	6 Generative
PKU-AIGIQA-4K (2024) [40]	AI-generated	4,000	4,000	–	Quality, Authenticity, Correspondence	3 Generative
AGIN (Ours)	**AI-generated**	**6,049**	**6,049**	**907,350**	**Technical, Rationality, Naturalness**	**18 Generative**

† This dataset does not collect ratings in the form of MOS, as in mainstream IQA datasets, and thus cannot be used directly for training deep learning-based IQA models.

generative adversarial network (GAN)-based methods [4], [5], [6] even diffusion-based generative models [7], [9], [12], [15] are prone to introduce perceptible unnatural perturbations (*e.g.*, spurious details, disordered layout, or color mismatches) due to their instability and mode collapse issues. Unfortunately, prior naturalness assessment studies driven by image statistic distribution or handcrafted features have predominantly focused on NSIs, which fail in AIGIs, where more diverse contextual content variations with less significant intrinsic properties (*e.g.*, resolution, color space, and image format) exist. Consequently, there is an urgent need to define the naturalness in AIGIs, as its underlying influencing factors remain an open question.

B. IQA Datasets

In the past two decades, a variety of IQA datasets [25], [26], [27], [28], [29], [30], [31], [32], [33] have been established to support the development of objective IQA algorithms for either generic images or domain-specific purposes. Tab. I summarizes and compares these IQA datasets. Regarding generic IQA datasets, LIVE [25] was proposed with 779 artificially distorted images. As a successor, TID2013 [28] was released with more distortion types and images, compared with LIVE, to simulate practical situations (*e.g.*, image acquisition, image compression, watermarking, and registration) comprehensively. Later, the emerging deep learning methods raise new requirements on large-scale IQA databases. Subsequent datasets [29], [30], [31], [32], [33] selected source contents from the Internet or multimedia collections, achieving a significant advancement either in image quantity or simulated distortion types. For instance, KonIQ-10k [33] first created a large collection of images in-the-wild to depict a broad range of appropriate content and then sampled images via eight content indicators, which enforce a more uniform distribution.

More recently, the rapid development and popularity of generative models have spawned several IQA databases for AIGIs. As a pioneering work, Zhang et al. [35] constructed AGIQA-1k, a small dataset with 1,080 generated images for the task of IQA. In [36], Wang et al. proposed AIGCIQA2023 with 2,400 images generated by six text-to-image models, where three dimensions of ratings are collected including quality, authenticity, and correspondence. Meanwhile, Kirstain et al. [37] built the Pick-A-Pic, a large-scale dataset of text-to-image prompts and real users' preferences over 500,000 generated images. However, the images within the Pick-A-Pic dataset were generated only by three models, resulting in a lack of content variety. Additionally, it does not collect ratings in the form of mean opinion score (MOS), as did in mainstream IQA datasets, thus hindering its utility for training deep learning-based IQA models. A similar problem also occurs with ImageRewardDB [38] and PKU-AIGIQA-4K [40], where the number of adopted generation models and annotators per image is limited. In summary, the aforementioned works are either carried out within small-scale groups or merely collect coarse-grained, single-voice, and overall subjective opinions, lacking the exploration of underlying influencing factors and explainable evaluations on various image generation tasks, which may lead to the ambiguity of subjective quality opinions in AIGI quality assessment (AIGIQA).

C. Objective Quality Assessment on AIGIs

As a branch of IQA, the AI-generated image quality assessment task has garnered increasing interest among researchers. Early objective metrics such as Inception Score (IS) [45] measures perceptual quality by calculating the uniformity of AIGI group features from the output of Inception model. Distance-based methods such as Fréchet Inception Distance (FID) [46] and Kernel Inception Distance (KID) [47] as well as Precision-Recall [48] evaluate the discrepancy between distributions of AIGIs and NSIs to quantify the quality of AIGIs. Nevertheless, the above methods are all group-targeted and not suitable for assessing single image. Besides, the widely used CLIPScore [49] is already saturated in comparing state-of-the-art generative models with authentic images and can inflate for a model trained to optimize text-to-image alignment in the CLIP space [50]. Additionally, there have been a series of learning-based works for quality assessment on AIGIs [36],

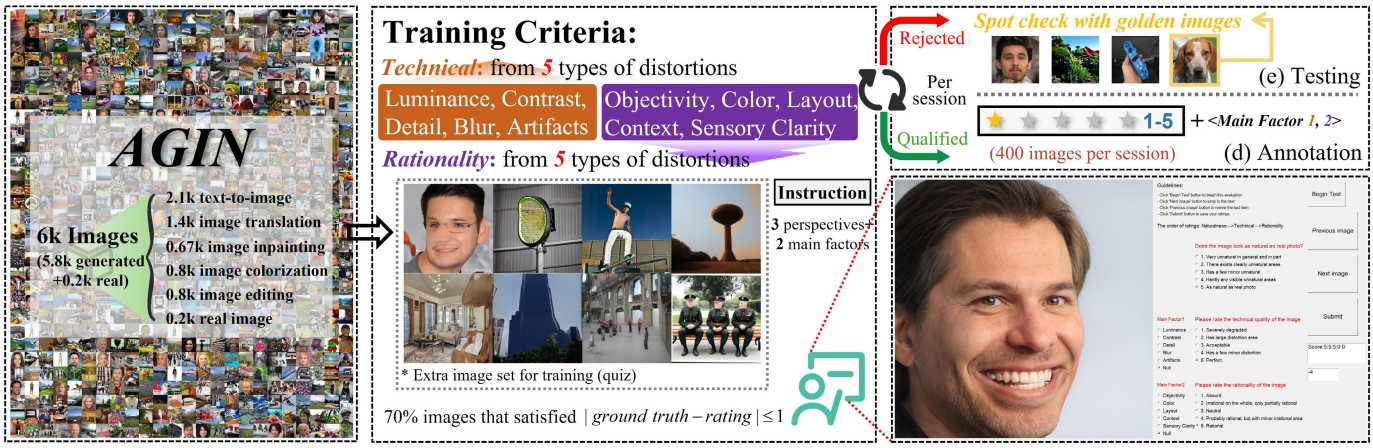


Fig. 2. Workflow of the subjective evaluation in AGIN. Source images were first collected from 5 generative tasks and real-world image datasets **(a)**, and then we conducted in-lab training with instructions **(b)**. After that, subjects were asked to rate the technical quality, rationality, and naturalness of AIGIs, and select the corresponding main factor through the radio buttons **(c, d)**, while the spot check was parallelly carried out **(e)** to control the annotation quality. Zoom in for better visualization.

[37], [38], [39], [40], [51], which mainly take perception, text-image alignment, and authenticity as the objective of AIGIQA. Li et al. [39] proposed to divide the prompt into multiple morphemes while cutting the whole picture into multiple stairs and giving the final score through their one-on-one alignment. In [38], Xu et al. used BLIP [52] to extract image and text features, then combined them with cross attention, and finally used an MLP to generate a scalar for human preference prediction. Inspired by the fact that the visual quality and authenticity are distance-sensitive, Zhou et al. [51] proposed an adaptive mixed-scale feature fusion network (AMFF-Net) for no-reference AIGIQA, which takes the scaled images and original-sized image as the inputs to obtain multi-scale features for better text-image feature alignment. However, most of these AIGIQA methods rely heavily on the correlation between textual and visual features, which is not conducive to other prompt-free image generation tasks (*e.g.*, image-to-image translation, editing, or colorization.). Besides, these models have not been specifically designed to address the unique distortion issues (*e.g.*, artifacts, layout, or existence) of AIGIs, remaining as an open problem. This motivates us to provide a more appropriate definition of image naturalness and propose an objective model to evaluate AIGIs from multiple perspectives.

III. AI-Generated Image Naturalness Dataset

To aid in the advancement of objective assessment models tailored for AI-generated images, we have painstakingly collected a novel dataset for AI-generated image naturalness assessment. Below, we elaborate on the construction procedures of the proposed AGIN dataset, shown in Fig. 2.

A. AIGI Collection

As an initial investigation, we choose five sources of AIGI from text-to-image, image translation, image inpainting, image colorization, and image editing tasks, which typically suffer from naturalness problems. We select 18 mainstream models including: **1)** five text-to-image models (*i.e.*, Stable Diffusion1.5 [3], [53], Stable Diffusion2.1 [54], Openjourney [55], Dreamlike [56], and Realistic Vision1.4 [57]) with over 400 prompts used for image generation, **2)** five image translation models (*i.e.*, RABIT [6], DiffuseIT [7], StyleCLIP [5], MATEBIT [8], and CoCosNet [4]), with various text-, image-, or mask-guided (*e.g.* edge map, semantic map) translation modes, **3)** two image inpainting models (*i.e.*, RePaint [9] and MAT [10]) that take mutilated images as inputs, **4)** three image colorization models (*i.e.*, PDNLA-Net [13], DDColor [11], and DISCO [12]) that colorize the grayscale images, and **5)** three image editing models (*i.e.*, InstructPix2Pix [15], DragGAN [14], and MagicBrush [16]) that perform layout or content editing on the image via text prompts or interactive anchor points. Note that we also added 200 extra real images into the AGIN dataset to help analyze the accuracy of the subjective experiment and objective algorithms, which stand for a high level of naturalness. As a result, 6,049 images were collected for the following subjective experiments.

B. Design of the Subjective Evaluation

1) Choice of Naturalness-Related Factors: We study the naturalness assessment problem from two perspectives, *i.e.*, the low-level technical quality-related perspective and the high-level rationality-related perspective. For the factors in technical perspective, we consider specific image attributes (*i.e. luminance* (**T-1**) and *contrast* (**T-2**)) that have high correlations with the naturalness of real images [19], [21], [22]. Besides, in-capture authentic distortions of NSI [58], such as reproduction of *details* (**T-3**) and *blur* (**T-4**), are also considered. We further include the *artifacts* (**T-5**) (content discontinuity or meaningless objects) introduced by the instability and mode collapse issues of generative models [24], [39], [59]. Since AIGIs possess richer content with diverse styles than real images, the visual naturalness of AIGIs is largely affected by rationality-related factors beyond technical distortions [17]. Such high-level factors are vaguely described as *the memory of the real-life* in previous research [20], [23], which are not suitable for qualitative and quantitative analysis. Here,

Fig. 3. Visualization of images with severe (1st row with red box) and minor effects (2nd row with green box) of each dimension.

we contribute five rationality-related dimensions to collect human feedback:

(R-1) *Existence*: Whether the scene or objects in the image exist or could exist in the real world.

(R-2) *Color*: Does the image follow the natural color rule with good color aesthetics [60]?

(R-3) *Layout*: Is the image layout logical?

(R-4) *Context*: Whether the objects in the image are related.

(R-5) *Sensory Clarity*: Perception of abstraction level. Whether the image content is easy to understand.

Fig. 3 presents examples with varying degrees of effect for each dimension, illustrating the manifestations of the naturalness problem across different dimensions.

2) Participants and Apparatus: To ensure the comprehensiveness and reliability of the evaluation, we recruited 30 participants (18 male, 12 female, age = 22.63.1) from campus, all with normal (corrected and no difficulty in color recognition) eyesight. We conduct the subjective studies in-lab to ensure that all subjects have a clear and consistent understanding of each factor. All images are displayed on a Lenovo 27-inch screen with a resolution of 2560 × 1440 and a viewing distance of about 70cm. Other settings, such as ambient brightness, lighting, and background, are configured according to the ITU-R BT.500 recommendation [61]. Note that we have addressed the ethical challenges involved in constructing such a dataset, by obtaining from each subject depicted in the dataset a signed and informed agreement, making it equipped with such legal and ethical characteristics.

3) Rating Strategy and Wording: We discuss the concrete rating form for subjective evaluation as follows. **a)** *Task-oriented absolute choice*. Since the wording of questions can significantly affect annotators' labeling behavior [50], we abandon the traditional endpoint labels from worst to best, which are too vague to describe the degree of naturalness. As a solution, we design specific labels for the evaluated three perspectives to reduce subjectivity, as shown in Fig. 2(c). **b)** *Pick up the main factor*. Most existing IQA datasets merely focus on the assessment of the overall score but neglect to explore the underlying factors. Therefore, we ask subjects to choose a primary factor that affects most for each perspective after rating the general scores [27], which enables us to investigate the correlation between each dimension and image naturalness. Fig. 2(c) shows the rating interface, which is composed of the left image display area and the right operation area with function buttons. To avoid interference between each perspective, we split the whole evaluation process into two phases that only after the naturalness evaluation is complete can participants move on to rate the technical quality and rationality, as well as to select their respective main factor.

4) Training, Testing, and Annotation: As shown in Fig. 2, before the annotation, we instructed all participants to have a clear and consistent understanding of all evaluated aspects and tested their eligibility via a 10-image pre-labeling. Their answer is compared with ground-truth ratings that were collected from five experts. Participants need to achieve at least 70% ratings that satisfy $|ground\ truth - rating| \leqslant 1$ to move on. During the formal annotation, we shuffled and divided all images into 15 sessions, each containing 400 images except for the 15th. It took a participant 22.7s on average to evaluate an image (= 15.3s of rating three perspectives + 7.4s of selecting the main factors). Consequently, a single session will require over two hours to complete. To reduce visual fatigue, there is a rest session with at least 15 minutes between two sessions. Moreover, we randomly inserted ten *golden images* (universally acknowledged with poor or good quality) into each session as an inspection to ensure the quality of annotation for in-process testing. Each participant was compensated $16 for each session according to the current ethical standard [62]. At last, a total of 6,049 × 30 × (3+2) = 907,350 ratings were collected. The mean opinion score ($\mathrm{MOS}_{a,i}$) for aspect $a \in \{\mathrm{T}, \mathrm{R}, \mathrm{N}\}$ of image i is obtained by averaging the raw opinions $\mathrm{OS}^k_{a,i}\big|^K_{k=0}$ from K subjects.

C. Annotation Quality Control

In addition to the pre-labeling and in-process check trials, we further assess the reliability of the rated scores by calculating the inter-annotator agreement metric, Krippendorff's α [63]. Specifically, Krippendorff's α for technical quality, rationality, and naturalness ratings are 0.32, 0.33, and 0.37, respectively, indicating appropriate variations among annotators. Furthermore, we use Spearman's rank-order correlation coefficient (SRCC) to calculate the correlation between a single participant and the overall MOS, thus judging whether an annotator is an outlier. As a result, we noticed two line clickers with over 37.5% of the same ratings, which have an extremely low SRCC (0.1851 and 0.2839). We removed all their ratings, improving Krippendorff's α of 0.07, 0.05, and 0.04 *w.r.t.* technical quality, rationality, and naturalness scores, respectively.

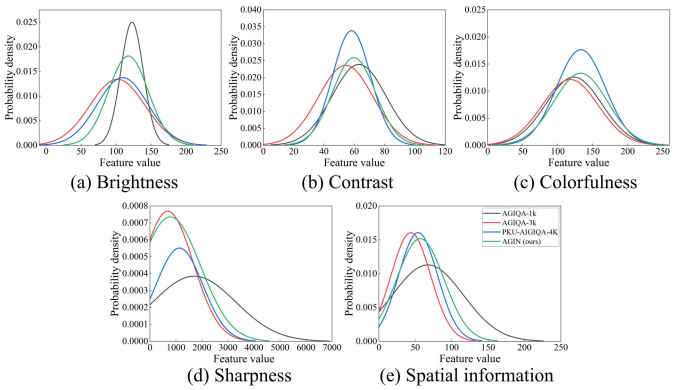


Fig. 4. Feature distribution comparisons among the four representative AI-generated image quality datasets: AGIQA-1k [35], AGIQA-3k [39], PKU-AIGIQA-4k [40], and the proposed AGIN.

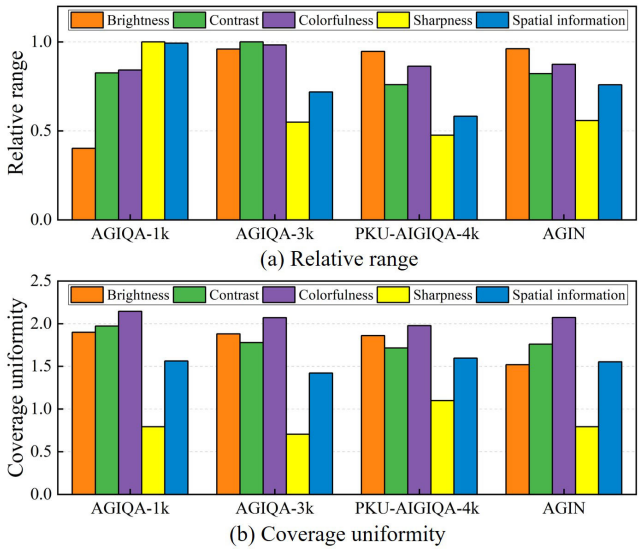


Fig. 5. Relative range R_i^k and coverage uniformity U_i^k comparisons of the selected five features computed on four representative AIGIQA datasets: AGIQA-1k [35], AGIQA-3k [39], PKU-AIGIQA-4K [40], and our AGIN.

D. Dataset Analysis

1) Image Attributes: Here we characterize the content diversity of the images in our dataset using five low-level features, including brightness, contrast, colorfulness, sharpness, and spatial information (SI). Three representative AIGIQA datasets, *i.e.*, AGIQA-1k [35], AGIQA-3k [39], and PKU-AIGIQA-4k [40], are selected for comparison. Fig. 4 shows the fitted normal distribution of each selected feature. Besides, we follow the procedure in [64] to quantify the relative range and coverage uniformity of these datasets over each feature space, which measure the inter- and intra-dataset differences, respectively. As shown in Fig. 4 and Fig. 5, our AGIN exhibits comparable coverage with the other datasets in terms of contrast and colorfulness. Regarding the sharpness and SI, AGIQA-1k shows a skew towards higher values than the other datasets, which is consistent with the observation that the images in AGIQA-1k are generated by only two diffusion-based text-to-image models. As for the brightness, our AGIN and AGIQA-3k are spread most widely, while AGIQA-1k adheres closer to middle values. In general, the proposed AGIN owns a preferable content diversity with appropriate range and uniformity.

TABLE II

CORRELATION BETWEEN DIFFERENT PERSPECTIVES AND OVERALL NATURALNESS IN AGIN. $a = 0.145$, $b = 0.769$

METRICS	MOS_T	MOS_R	$MOS_T + MOS_R$	$aMOS_T + bMOS_R$
SRCC↑	0.8647	0.9694	0.9672	**0.9777**
PLCC↑	0.8599	0.9639	0.9580	**0.9713**

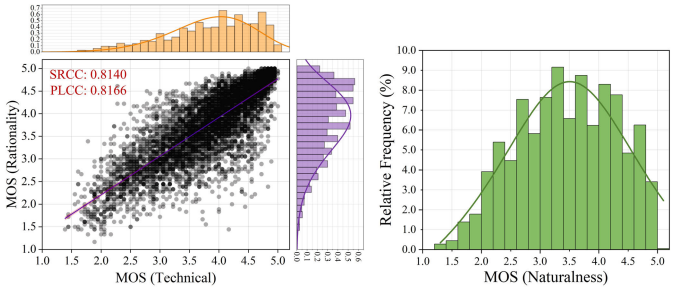


Fig. 6. Data properties of AGIN. (a) The correlation between technical and rationality perspectives, (b) distribution of the overall naturalness scores.

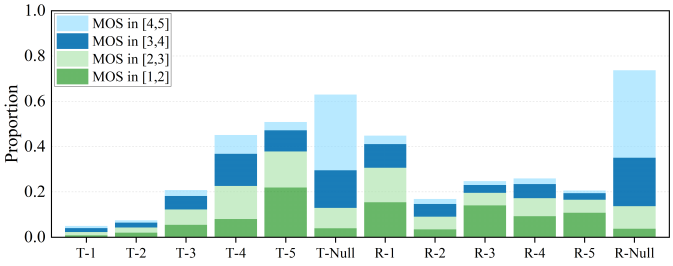


Fig. 7. The tendency of main factors chosen by participants across different ranges of MOS_N.

2) Statistics of the Subjective Ratings: We visualize the distribution of three evaluated perspectives and the inner correlation between the technical and rationality perspectives in Fig. 6. Tab. II lists Spearman's and Pearson's correlations between different perspectives and naturalness. We can observe a right-skewed distribution in all three dimensions, reflecting the overall performance of current generative models. Simultaneously, it is notable that technical quality and rationality affect naturalness unequally, *i.e.*, rationality has a greater impact on the overall naturalness (SRCC=0.9694) than technical perspective (SRCC=0.8647). Besides, a simple addition of these two perspectives has a higher correlation with the overall naturalness score than the technical itself. To seek the best fitting form, we apply a two-parameter approximation to explore the weights [65]. As a result, we find that the MOS_N can be well approximated as $0.145MOS_T + 0.769MOS_R$, which reaches 0.9777 in terms of SRCC. This indicates that employing mainstream IQA models, which follow an overall MOS regression strategy, could inadvertently lead to biased naturalness assessment.

3) Frequency of Different Naturalness Factors: Fig. 7 shows the proportion of each factor in different ranges of naturalness scores. The terms "T-Null" and "R-Null" signify the absence of either the absence of factors related

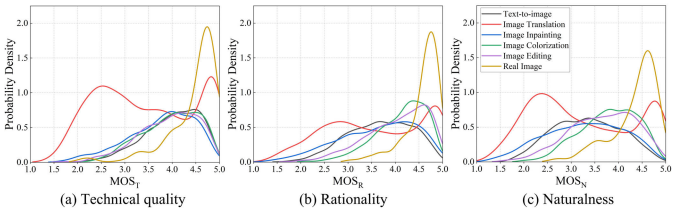


Fig. 8. MOS distribution comparisons among five image generation tasks.

to technical quality and rationality affecting the naturalness, or the difficulty subjects faced in identifying the primary factors. Specifically, we find that 'T-Null' and 'R-Null' are more prevalent in images with higher naturalness scores ($\text{MOS}_\text{N} \in [4, 5]$), indicating that images with a high degree of naturalness tend to exhibit relatively better technical quality and rationality. Moreover, we notice that humans are more sensitive to the artifacts (T-5) and blur (T-4) *w.r.t.* technical quality while focusing more on the existence (R-1) of the image contents *w.r.t.* rationality. Meanwhile, an notably high proportion of artifacts (T-5), existence (R-1), and layout (R-3) is found in cases of poor naturalness ($\text{MOS}_\text{N} \in [1, 2]$). This implies that severe artifact distortions can result in irrational contents and disorganized layouts, underscoring their significance as key factors of naturalness for consideration in AIGIs.

4) Statistics of Task-Wise Ratings: We can observe from Fig. 8 that real images have higher MOS in all three perspectives than AIGIs, highlighting the naturalness problems in current image generation tasks while emphasizing the necessity of addressing these issues. Besides, compared to the other generation tasks, the image translation task has a larger proportion in the low segment of all three perspectives, indicating that the change of content and style during image translation is prone to produce naturalness distortion. It is worth noting that image colorization owns the highest average MOS among all three perspectives, aligning with the observation depicted in Fig. 7, where color irrationality constitutes the smallest proportion of the primary factors impacting naturalness. Furthermore, we also investigate the occurrence frequency of each factor across models, shown in Fig. 9. Overall, these newly contributed dimensions describe the naturalness concerns of AIGIs, some of which have never been encountered in the conventional IQA domain, providing reliable intuitions for developing objective naturalness assessment models.

IV. Joint Objective Image Naturalness Evaluator

In this section, we introduce the proposed JOINT and JOINT++. We first discuss the design philosophy and motivation in Sec. IV-A. Then we elaborate on two key learning branches, *i.e.*, technical prior branch (Sec. IV-B) and rationality perceiving branch (Sec. IV-C). Lastly, we present the associated objective functions (Sec. IV-D) for model training as well as a subjective-inspired weighting strategy for naturalness assessment.

A. Motivation

To develop an objective INA model for evaluating the naturalness of AI-generated images, we commence by first gaining insights from the perceptual mechanism of human visual system. Studies in neurosciences [66], [67] suggest that humans possess two distinct visual systems, which follow two main pathways, *i.e.*, the dorsal stream and ventral stream, to handle low-level and high-level visual perception, respectively. The dorsal stream is involved in spatial information awareness from the visual cortex and is good at detecting and analyzing spatial distortions. The ventral stream is associated with object recognition and form representation, especially long-term stored representations (so-called *memory*), which are highly related to the rationality perception of an image. Meanwhile, we notice from the subjective studies in AGIN that overall naturalness opinions are affected by both low-level technical and high-level rationality perspectives.

Inspired by these findings, we propose the Joint Objective Image Naturalness evaluaTor (**JOINT**) to align model behavior with human perception process. Specifically, we decompose the AIGI into two views, namely the technical view (V_T) that focuses on technical quality perception, and rationality view (V_R) for vice versa, shown in Fig. 10. With the decomposed views as inputs, two technical (M_T) and rationality (M_R) branches evaluate different perspectives independently:

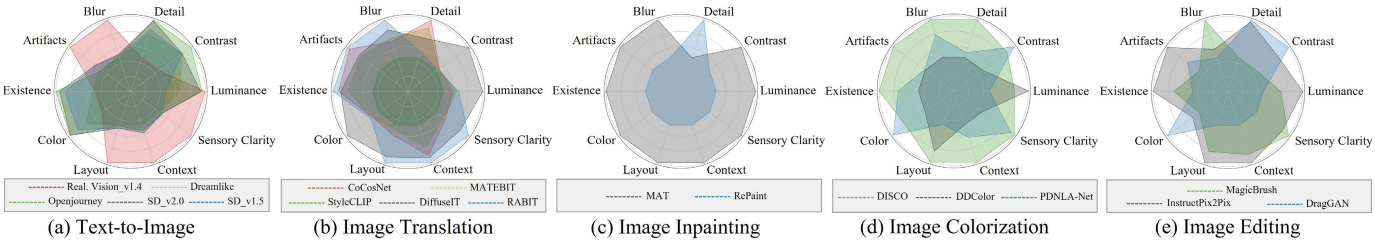
$$\hat{S}_\text{T} = \text{M}_\text{T}(V_\text{T}); \quad \hat{S}_\text{R} = \text{M}_\text{R}(V_\text{R}). \tag{1}$$

Although most perceptions related to the two perspectives can be separated, a small proportion of factors are related to both perspectives, such as *color* (an argument could also be made for it to fall under the technical perspective that is closely linked to low-level visual features including *luminance* and *contrast* [68].), or *artifacts* (which can greatly impact the semantic rationality of an image [24]). Henceforth, we circumvent separate these factors but instead employ inductive biases for each branch (*patch partition strategy* and *feature regularization*) to allocate these two branches with corresponding learning interests. Furthermore, these two branches are separately supervised, either both by the overall naturalness scores (denoted as **JOINT**) or by respective technical and rationality opinions exclusively in the AGIN (denoted as **JOINT++**), introduced as follows.

B. Technical Prior Branch

For technical prior branch, we explicitly guide the model to prioritize the technical distortions while minimizing the impact of semantic information. A straightforward way is to crop the image $\mathcal{I}$ into size-fixed patches and stitch them together ($\mathcal{I}_{rand}$) to disorganize most contents and layout while retaining technical distortions, thus destroying semantic information and rationality factors in images [69], [70]. However, different from most global technical distortions (noise or blur), generated artifacts could become indistinguishable by random patch partition. Thus, we propose to localize possible perceptual artifacts first and bypass these regions to keep their local distortion information.

Specifically, we use the segmentation model proposed in [71] as the artifacts locator, which detects and segments the artifact areas that are noticeable to humans, to guide the patch partition. Given an image $\mathcal{I}$ of size $\mathcal{I}_W \times \mathcal{I}_H$, the perceptual



(a) Text-to-Image (b) Image Translation (c) Image Inpainting (d) Image Colorization (e) Image Editing

Fig. 9. Radar chart of the frequency of occurrence at different generative tasks for ten types of naturalness distortions.

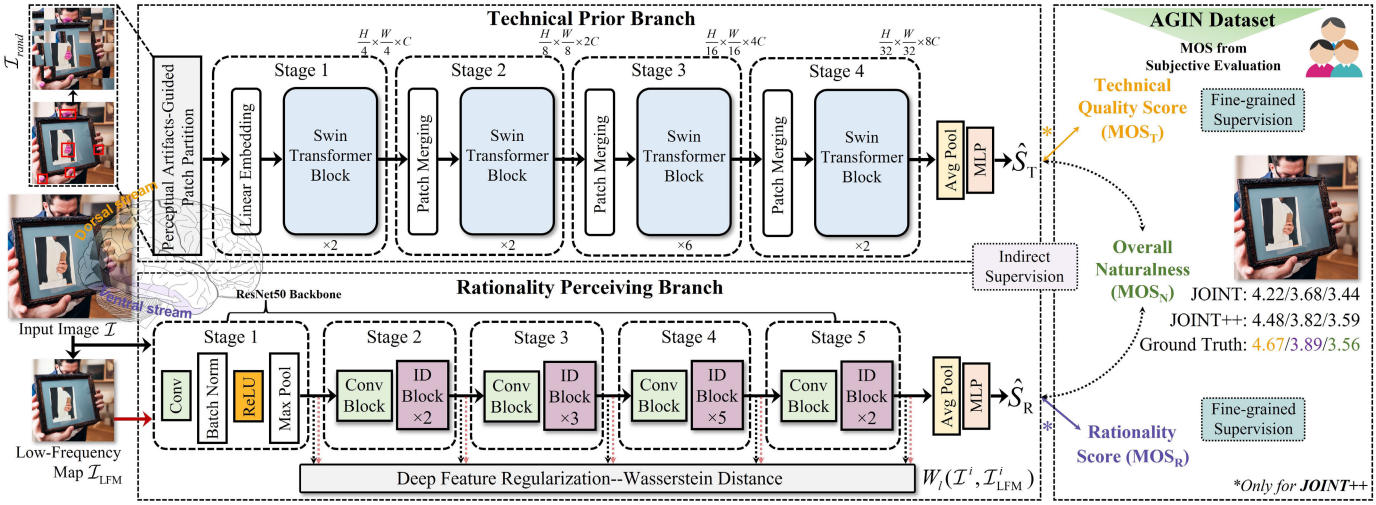


Fig. 10. Framework of the proposed JOINT and JOINT++, including the technical prior branch (Sec. IV-B) and rationality perceiving branch (Sec. IV-C) with indirect and fine-grained supervision strategies (Sec. IV-D).

artifacts-guided patch partition can be formulated as:

$$m_c, n_c = ALocator(\mathcal{I}), \quad c \in \{1, \ldots, C\} \tag{2}$$

$$\mathcal{I}_{rand} = RPart\left(\mathcal{I}_{j\in[1,\frac{\mathcal{I}_H}{N}]\setminus\{m_c\},k\in[1,\frac{\mathcal{I}_W}{N}]\setminus\{n_c\}}, N_8\right), \tag{3}$$

where $ALocator(\cdot)$ denotes the perceptual artifact locator that returns coordinates for the c-th artifacts in m-th horizontal grid and n-th vertical grid. C is the total number of detected artifact regions. The size of divided patch $\mathcal{I}_{j,k}$ is $N \times N$. Specifically, j and k are range from 1 to $\frac{\mathcal{I}_H}{N}$ and $\frac{\mathcal{I}_W}{N}$, respectively, while subject to the condition that $(j,k) \neq (m_c, n_c)$. $RPart(\cdot, N_8)$ indicates a random partition within the 8-connected neighborhoods of a patch, which destructs the local semantics of the image while preserving the global semantics (Fig. 11). Afterward, the re-permuted patches will pass through a Swin-T [72] backbone, and the extracted features are then flattened and sent to a multilayer perceptron (MLP) for technical quality score regression.

Fig. 11. Visualization of perceptual artifacts-guided patch partition. The first row shows the original images from AGIN. The second row exhibits the prediction results of artifacts (*violet areas*). The third row shows the resulting $\mathcal{I}_{rand}$ after perceptual artifacts-guided patch partition.

C. Rationality Perceiving Branch

Since the high-level semantic information in rationality concerns is likewise of interest to the image aesthetic assessment (IAA), we thereby pre-train this branch with IAA dataset AVA [73] first and introduce a deep feature regularization to mitigate the impact of technical perspective. In particular, to maintain the principal content of the image while reducing the impact of partial technical distortions (*e.g.*, noise, blur, or detail loss), we apply the piece-wise smooth algorithm [74] to obtain the low-frequency map (LFM) of image $\mathcal{I}_{\text{LFM}}$ (Fig. 12). We can observe that the $\mathcal{I}_{\text{LFM}}$ filters out some technical quality-related attributes but still preserves the semantic information of the original image.

Moreover, existing research [75], [76] suggests that the distribution differences of deep features among different stages are related to technical distortions. Henceforth, we employ the one-dimensional form of Wasserstein distance (WSD) as the penalty constraint $\mathcal{L}_{\text{WSD}}$ to eliminate the technical interference in rationality branch (M_R) by reducing the feature distribution

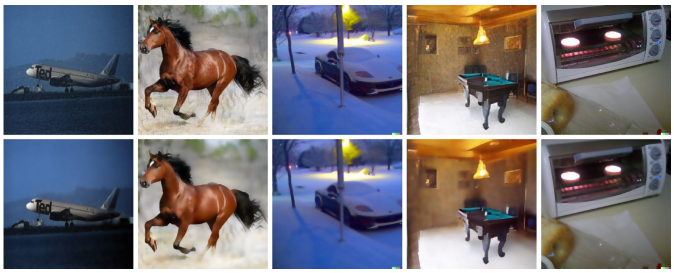


Fig. 12. Comparison of the original image (first row) and its corresponding low-frequency map (second row).

divergence between $\mathcal{I}$ and $\mathcal{I}_{\text{LFM}}$:

$$\mathcal{L}_{\text{WSD}} = W_l\left(\mathcal{I}, \mathcal{I}_{\text{LFM}}\right) + \sum_{i=1}^{N} W_l\left(\mathcal{I}^i, \mathcal{I}^i_{\text{LFM}}\right), \tag{4}$$

where $\mathcal{I}^i$ and $\mathcal{I}^i_{\text{LFM}}$ denote the extracted features of $\mathcal{I}$ and $\mathcal{I}_{\text{LFM}}$ at the i-th stage. $W_l(\cdot,\cdot)$ is the Wasserstein distance with l-norm. More specifically, as shown in Fig. 10, we take the original image $\mathcal{I}$ and its LFM $\mathcal{I}_{\text{LFM}}$ as inputs and use the ResNet50 [77] to extract their deep features at five stages, which has been proved to be useful for quality assessment [78]. Later, we reshape the $\mathcal{I}$, $\mathcal{I}_{\text{LFM}}$, $\mathcal{I}^i$, and $\mathcal{I}^i_{\text{LFM}}$ to 1D vectors to calculate the $\mathcal{L}_{\text{WSD}}$. Beyond this, the regression tactic in the rationality branch is the same as technical branch.

D. Learning Objectives

To utilize the relative quality among images while keeping the absolute prediction accuracy, we apply the weighted sum of monotonicity and the commonly used mean squared error (MSE) loss. Specifically, SRCC loss is adopted to boost the prediction monotonicity of models [86], which is defined in the form of Pearson's linear correlation coefficient (PLCC) between ranks:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M}\sum\nolimits_{n=1}^{M}\left\|y_n - \hat{y}_n\right\|_2^2, \tag{5}$$

$$\mathcal{L}_{\text{SRCC}} = 1 - \frac{\sum_n (v_n - \bar{v})(p_n - \bar{p})}{\sqrt{\sum_n (v_n - \bar{v})^2 \sum_n (p_n - \bar{p})^2}}, \tag{6}$$

$$\mathcal{L}_{\text{C}} = \mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{SRCC}}, \tag{7}$$

where v_n and p_n denote the rank of the ground truth y_n and the rank of predicted score $\hat{y}_n$, respectively. Since the naturalness is affected by both technical quality and rationality perspectives (Sec. III-D), we first optimize two branches using the overall naturalness MOS_{N} in an indirect supervision way ($\mathcal{L}_{\text{IS}}$):

$$\mathcal{L}_{\text{IS}} = \mathcal{L}_{\text{C}}\left(\hat{S}_{\text{T}}, \text{MOS}_{\text{N}}\right) + \mathcal{L}_{\text{C}}\left(\hat{S}_{\text{R}}, \text{MOS}_{\text{N}}\right) + \beta\mathcal{L}_{\text{WSD}}, \tag{8}$$

where β is a hyperparameter to control the strength of $\mathcal{L}_{\text{WSD}}$. $\hat{S}_{\text{T}}$ and $\hat{S}_{\text{R}}$ denote the predicted score of the technical quality and rationality branches, respectively. Besides, based on the AGIN dataset, we also propose a fine-grained version ($\mathcal{L}_{\text{FS}}$) using the corresponding perspective opinions for both branches:

$$\mathcal{L}_{\text{FS}} = \mathcal{L}_{\text{C}}\left(\hat{S}_{\text{T}}, \text{MOS}_{\text{T}}\right) + \mathcal{L}_{\text{C}}\left(\hat{S}_{\text{R}}, \text{MOS}_{\text{R}}\right), \tag{9}$$

and the proposed JOINT++ is trained by a weighted combination of the above two losses ($\mathcal{L}_{\text{IS}}$ and $\mathcal{L}_{\text{FS}}$) to obtain more accurate predictions for both branches:

$$\mathcal{L}_{\text{JOINT++}} = \mathcal{L}_{\text{FS}} + \lambda_{\text{IS}}\mathcal{L}_{\text{IS}} \tag{10}$$

Finally, we adopt a simple but effective fusion strategy to compute the overall naturalness score ($\hat{S}_{\text{N}}$) from two perspectives: $\hat{S}_{\text{N}} = 0.145\hat{S}_{\text{T}} + 0.769\hat{S}_{\text{R}}$, according to the subjective studies in AGIN.

V. Experiments

A. Experimental Setup

1) Implementation Details: In the technical branch, we crop patch at size 32×32, and Swin-T [72] is used as backbone. We use the ResNet50 backbone [77] pre-trained with AVA dataset [73] in the rationality branch. α and β in $\mathcal{L}_{\text{IS}}$ are set as 1 and 0.5, respectively. λ_{IS} in Eq. 10 is set as 0.5. We train our model for 30 epochs using the Adam optimizer [87] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is 2×10^{-5}, and the batch size is set to 32. In the rationality branch, all images are calculated at size 224×224 so as to satisfy the input requirement of ResNet50. Besides, for the regularization term $\mathcal{L}_{\text{WSD}}$, we set $l = 2$ at Eq. 4, as in [76], making the quality measure more sensitive to outliers, while the N in $\mathcal{L}_{\text{WSD}}$ is 5, corresponding to the five stages in Resnet50 backbone. Other baselines are initialized using their respective settings. Before training, we randomly split the AGIN dataset into the training, validation, and test sets with a ratio of 7:1:2 for 5 times, and report the averaged results. All experiments are conducted on a single NVIDIA RTX 4090 24G GPU.

2) Evaluation Criteria: We adopt two criteria to evaluate the accuracy of quality predictions: SRCC and PLCC, which measure the prediction monotonicity and linearity, respectively. Before calculating PLCC, we mapped the model's predictions to the MOSs using a four-parameter logistic nonlinearity [86].

3) Competing Methods: We select 14 representative methods for comparison, including two classical no reference (NR) IQA methods: BRISQUE [88] and NIQE [89], five deep NR IQA methods, DBCNN [90], HyperIQA [91], MUSIQ [92], UNIQUE [93], and MANIQA [94], four image aesthetic assessment (IAA) methods: PAIAA [95], TANet [83], Delegate Transformer (Dele-T) [60], and SAAN [84], and three contrastive language-image pre-training (CLIP) model-based IQA methods, CLIP-IQA [96], CLIP-IQA$^+$ [96], and LIQE [97].

B. Exploring the Necessity of AGIN Dataset

We first conduct experiments to verify whether the existing IQA and IAA datasets can solve the problem of AI-generated image naturalness assessment, *i.e.*, the necessity of AGIN dataset. Specifically, we test a large number of models (both traditional IQA and IAA models, along with the visual-language prior-based IQA models) that were trained on datasets from their respective domains, such as TID2013 [28] for synthetic distortions, KonIQ-10k [33] for in-the-wild authentic distortions, PIPAL [34] for GAN-based image restoration distortions, and AVA [73] for aesthetic analysis.

TABLE III

VALIDATING THE NECESSITY OF AGIN DATASET. ALL BASELINES ARE TRAINED USING DATASETS FROM THEIR RESPECTIVE DOMAINS. THE 1ST, 2ND, AND 3RD BEST SCORES ARE DENOTED IN **RED**, **BLUE**, AND **BLACK**, RESPECTIVELY

Methods	Type	Training Datasets	Technical		Rationality		Naturalness	
			SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BRISQUE (TIP, 2012)	*Classical IQA*	KonIQ-10k [33]	0.3544	0.3602	0.1268	0.1299	0.1618	0.1660
NIQE (SPL, 2013)	*(handcraft features)*	KonIQ-10k [33]	0.1843	0.1484	0.0377	0.0235	0.0707	0.0445
*DBCNN (TCSVT, 2018)	*Deep IQA (deep features)*	TID2013 [28]	0.2664	0.3138	0.0888	0.1199	0.1209	0.1376
– – *same as above* – –		LIVE Challenge [29]	0.4132	0.4903	0.1518	0.2082	0.1993	0.2422
		KonIQ-10k [33]	0.4951	0.5252	0.2275	0.2492	0.2786	0.2956
HyperIQA (CVPR, 2020)		KonIQ-10k [33]	0.4953	0.5541	0.2839	0.3211	0.3332	0.3725
* MUSIQ (ICCV, 2021)		PaQ-2-PiQ [79]	0.4329	0.4709	0.2061	0.2399	0.2443	0.2799
– – *same as above* – –		KonIQ-10k [33]	0.4817	0.5262	0.2512	0.2847	0.2951	0.3271
		SPAQ [80]	0.4324	0.5166	0.2193	0.2741	0.2561	0.3085
UNIQUE (TIP, 2021)		ImageNet [81]	0.5178	0.5756	0.2912	0.3324	0.3339	0.3735
*MANIQA (CVPRW, 2022)		KADID-10k [32]	0.4154	0.4214	0.2733	0.2655	0.3003	0.3001
– – *same as above* – –		KonIQ-10k [33]	**0.5771**	**0.5902**	**0.3453**	**0.3594**	**0.3937**	**0.4034**
		PIPAL2022 [34]	0.4014	0.4407	0.1985	0.2354	0.2341	0.2597
PAIAA (TIP, 2020)	*Deep IAA (deep features)*	PsychoFlickr [82]	0.1363	0.1234	0.2445	0.2587	0.2261	0.2298
TANet (IJCAI, 2022)		TAD66k [83]	0.1894	0.2015	0.2774	0.2803	0.2530	0.2619
Dele-T (ICCV, 2023)		AVA [73]	0.2549	0.3142	0.2348	0.2278	0.2232	0.2583
SAAN (CVPR, 2023)		BAID [84]	0.0515	0.0359	0.1477	0.1380	0.1413	0.1456
CLIP-IQA (AAAI, 2023)	*CLIP-based model (visual language prior)*	WIT-400M	0.2114	0.3275	0.0348	0.0827	0.0167	0.1109
CLIP-IQA$^+$ (AAAI, 2023)		WIT-400M+KonIQ-10k [33]	0.4959	0.5595	0.2613	0.3189	0.3078	0.3550
LIQE (CVPR, 2023)		*hybrid* [25], [27], [29], [32], [33], [85]	0.4928	0.5428	0.2457	0.2765	0.2974	0.3244
JOINT (Ours)	*Deep INA (deep features)*	$\text{AGIN}_{\text{train}}$	**0.8173**	**0.8235**	**0.7564**	**0.7711**	**0.7986**	**0.8028**
JOINT++ (Ours)			**0.8351**	**0.8429**	**0.8033**	**0.8127**	**0.8264**	**0.8362**

We calculate the SRCC and PLCC between the model's predicted outputs and the MOS across three perspectives in the AGIN dataset. We can obtain the following observations from Tab. III. First, either *state-of-the-art* IQA models or deep IAA models yield inferior results on all three perspectives. Among them, the second-best approach, MANIQA [94] trained on KonIQ-10k [33], performs lower than our JOINT++ in naturalness evaluation by 0.4327 (-109.91%) and 0.4328 (-107.29%) in terms of SRCC and PLCC. Meanwhile, the evaluated IAA models achieve average SRCC scores of only 0.1580, 0.2261, and 0.2109 in terms of technical, rationality, and naturalness perspectives, respectively. These results indicate that the distortion issues on AIGIs are significantly different from those on NSIs (neither a complete technical distortion nor an aesthetic problem) and show the inapplicability of current IQA and IAA algorithms in naturalness evaluation, thus highlighting the necessity of our AGIN dataset for developing future objective naturalness metrics. Second, evaluating AIGIs from technical and rationality perspectives exhibits notable differences. We notice that IQA methods perform relatively better in evaluating technical quality, whereas IAA methods are more proficient at assessing rationality, which underscores the reasonability and effectiveness of our distinct exploration of each perspective in the AGIN dataset. Third, the evaluation of image naturalness is distinct from the image quality assessment and image aesthetic assessment tasks. Since mainstream IQA and IAA approaches trained with current assessment datasets fail to provide subjectively consistent results for image naturalness (>105% lower in SRCC and PLCC), we speculate that this is due to the variance of image content sources and disparities in task objectives, illustrating the necessity of defining naturalness for AIGIs and exploring its underlying influencing factors.

TABLE IV

PERFORMANCE COMPARISONS ON THE AGIN. WE RETRAINED ALL MODELS USING THE SCORE OF EACH CORRESPONDING PERSPECTIVE

Methods	Technical		Rationality		Naturalness	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
BRISQUE	0.4867	0.4909	0.3608	0.3684	0.3745	0.4067
NIQE	0.4235	0.4279	0.3144	0.3211	0.3358	0.3378
DBCNN	0.7623	0.7661	0.6834	0.6838	0.7057	0.7128
HyperIQA	0.7752	0.7806	**0.7196**	**0.7292**	0.7365	**0.7509**
MUSIQ	0.7286	0.7355	0.6974	0.7013	0.7066	0.7103
UNIQUE	0.7358	0.7434	0.6583	0.6685	0.6772	0.6789
MANIQA	**0.7763**	**0.7817**	0.7192	0.7217	**0.7385**	0.7343
PAIAA	0.4763	0.4833	0.4532	0.4596	0.4483	0.4528
TANet	0.5367	0.5587	0.4731	0.4762	0.4782	0.4535
Dele-T	0.5882	0.6134	0.5037	0.4942	0.4805	0.4961
SAAN	0.4299	0.4380	0.4009	0.4160	0.4196	0.4184
JOINT (Ours)	**0.8173**	**0.8235**	**0.7564**	**0.7711**	**0.7986**	**0.8028**
JOINT++ (Ours)	**0.8351**	**0.8429**	**0.8033**	**0.8127**	**0.8264**	**0.8362**

C. Evaluation on the AGIN

1) Quantitative Studies: In Tab. IV, we evaluate the baseline algorithms by retraining and testing them on the AGIN dataset. It can be observed that the two classical IQA methods, BRISQUE and NIQE, perform significantly worse than deep IQA methods, and the proposed JOINT++ still achieves the best performance in terms of technical quality, rationality, and overall naturalness perspectives. Surprisingly, all IAA methods reach subpar performance compared to deep IQA models and perform on average 44.74%/45.56% *w.r.t.* SRCC/PLCC lower than the JOINT++ in naturalness evaluation. Their ineffectiveness can be attributed to a lack of consideration for technical factors and an attention bias in understanding the semantics of the content itself. It is worth noting that most IAA models concentrate more on global information (*e.g.*, semantic, composition, and color) than local artifacts that could overwhelmingly affect the image naturalness, which further demonstrates the differences between the definition

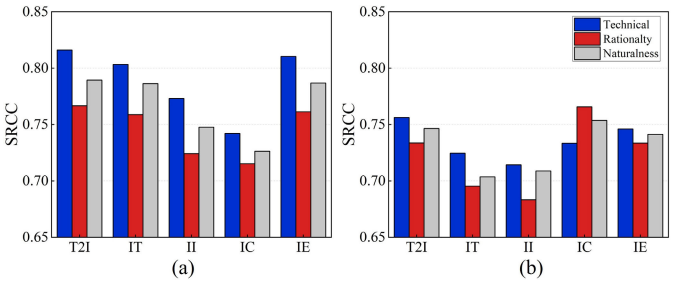


Fig. 13. Performance comparison on different generation tasks. T2I, IT, II, IC, and IE are the abbreviations of text-to-image, image translation, image inpainting, image colorization, and image editing tasks, respectively. We divided the AGIN dataset into five subsets according to the generation tasks and trained JOINT++ using the other four subsets while testing on the single target subset (a). Besides, we conducted train-test at a ratio of 8:2 within each subset (b).

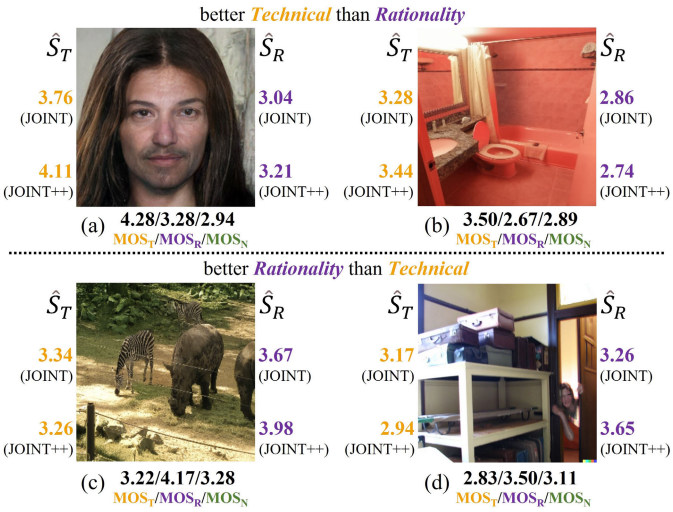


Fig. 14. Qualitative studies of JOINT/JOINT++. Visualizations of images in the AGIN where technical and rationality predictions are diverged.

of naturalness and aesthetics. Moreover, all IQA and IAA approaches solely consider a single perspective with highly coupled factors during image naturalness reasoning, thereby rendering them incapable of providing reliable results.

2) Performance on Different Generation Tasks: We further investigate the performance of JOINT++ on different generation tasks by conducting training and testing within the task-oriented subsets of AGIN. As shown in Fig. 13(a), when training with the other four subsets and testing independently on the T2I, IT, and IE subsets, JOINT++ performs similarly to when tested on the combinational subset of the whole AGIN. Nevertheless, when testing on the II and IC subsets, the performance dropped steeply. We conjecture that this corresponds to the variance of distortion types in each subset, *i.e.*, T2I, IT, and IE tasks involve richer forms of naturalness distortion than IC (mostly irrational color). Additionally, most of the images in II subset are derived from human face datasets that inherently differ from the other subsets in contents. Fig. 13(b) shows a marked decrease in terms of SRCC when conducting train-test evaluations within a single subset, primarily attributed to the reduced number of training samples. On the whole, performance on the technical perspective is slightly better than the rationality perspective except for the IC subset, since color

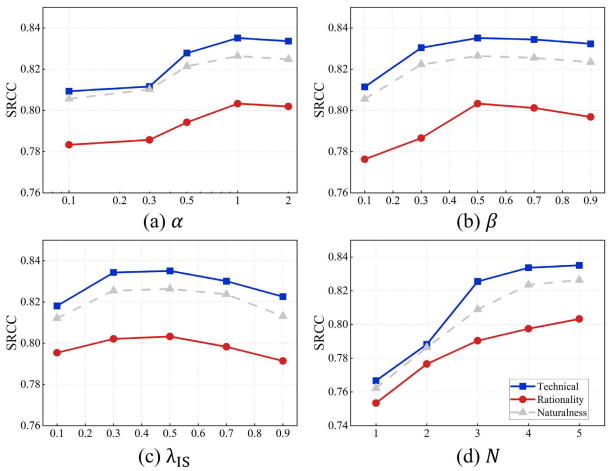


Fig. 15. SRCC performance of different parameters.

disharmony belongs more to a high-level rationality perception with aesthetics concerns.

3) Qualitative Studies: In Fig. 14, we visualize four images with diverged predicted technical and rationality scores. The two images with better technical scores (Fig. 14(a)&(b)) have clear contents yet suffer from irrational existence (a person with feminine characteristics has bushy beard hairs) and color (a nearly red bathroom). On the contrary, the two with better rationality scores (Fig. 14(c)&(d)) possess unambiguous semantics but with blurs and artifacts (incompletely formed zebra in the distance and a woman behind the door). These observations align with human perception of the two perspectives, proving the effectiveness of our joint learning strategy that can provide disentangled quality predictions. Furthermore, it should be noted from the perspective of real-world applications that rationality or irrationality may be strongly dependent on cultural context or diverse biometric differences. Therefore, the models trained on AGIN could be biased in some scenarios, which motivates us to continuously refine the dataset and conduct cross-cultural comparative research in the future.

D. Discussion on Model Parameter Selections

We further explore the selection of the parameters N in Eq. 4, α in Eq. 7, β in Eq. 8, and λ_{IS} in Eq. 10 by recording the SRCC results when different values are employed. A large value of α, β, and λ_{IS} indicates that more counterparts, *i.e.*, $\mathcal{L}_{\text{SRCC}}$, $\mathcal{L}_{\text{WSD}}$, and $\mathcal{L}_{\text{IS}}$, are considered in the loss function, respectively. N represents the granularity of deep features that participate in computing the Wasserstein distance. As shown in Fig. 15(a), the SRCC value experiences a vast improvement when α increases from 0.3 to 0.5, indicating that adding a proper proportion of monotonicity prediction terms within the loss function can boost model performance. Besides, optimal model performance is achieved when both balanced factors β and λ_{IS} are set to 0.5 (Fig. 15(b)&(c)). Fig. 15(d) shows that introducing more hierarchical features into the calculation of WSD can effectively enhance the ability of the rationality perceiving branch to suppress quality-sensitive information.

TABLE V
ABLATION STUDY OF SPECIFIC DESIGNS

Perspective/	Technical		Rationality		Naturalness	
Variants/Metric	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
w/o Localization	0.811	0.816	0.755	0.769	0.782	0.794
w/o Regularization	0.814	0.820	0.729	0.738	0.758	0.766
w/o Multi-perspective	0.768	0.781	0.703	0.712	0.727	0.733
JOINT (Ours)	**0.817**	**0.824**	**0.756**	**0.771**	**0.799**	**0.803**

TABLE VI
ABLATION STUDY OF CORRELATIONS BETWEEN PERSPECTIVES AND THE EFFECT OF SUBJECTIVE FUSION STRATEGY (DENOTED AS $\oplus$)

Variants			Technical		Rationality		Naturalness	
$\hat{S}_{\mathrm{T}}$	$\hat{S}_{\mathrm{R}}$	$\oplus$	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
✓			**0.817**	**0.824**	0.720	0.724	0.725	0.744
	✓		0.687	0.699	**0.756**	**0.771**	0.767	0.763
✓	✓		0.753	0.768	0.732	0.744	0.759	0.755
✓	✓	✓	0.711	0.723	0.746	0.762	**0.799**	**0.803**

TABLE VII
PERFORMANCE ON THE AGIN BY VARYING OF DIFFERENT BACKBONES IN THE TECHNICAL PRIOR BRANCH

Backbone	Size	Technical		Rationality		Naturalness	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
ResNet	18	0.753	0.761	0.763	0.777	0.758	0.765
	50	0.792	0.805	0.784	0.796	0.791	0.804
ViT	Tiny	0.815	0.818	0.794	0.806	0.807	0.813
	Base	0.841	0.847	0.803	0.813	0.828	0.838
ConvNext	Tiny	0.827	0.833	0.798	0.805	0.819	0.828
	Base	0.832	0.839	0.803	0.809	0.824	0.833
Swin Transformer	Tiny	0.835	0.843	0.803	0.813	0.826	0.836
	Base	**0.842**	**0.851**	**0.805**	**0.814**	**0.830**	**0.841**

E. Ablation Studies

1) Effects of Specific Designs: As Tab. V shows, we verify the effects of three special designs in the proposed JOINT by keeping other parts the same. First, JOINT performs superior to its variant *w/o Localization* that randomly shuffles the patches and destructs all image semantics, proving the importance of preserving the local perceptual artifact distortions for technical quality score regression. Second, a 0.027 improvement of SRCC in the rationality perspective is observed when equipped with deep feature regularization, demonstrating its effectiveness in reducing the technical influences in rationality prediction. Third, JOINT is notably better than the variant *w/o Multi-perspective* (+6.38%/+7.54%/+9.90% in terms of SRCC for three perspectives) that directly takes the original images as inputs of both branches, suggesting that explicit view decomposition encourages better naturalness-aware features to be learned.

2) Effects of Subjectively-Inspired Weighting: We discuss the naturalness score fusion strategy in Tab. VI. It is worth noting that predicted scores from any single branch can not adequately represent the naturalness, and directly taking $\hat{S}_{\mathrm{T}} + \hat{S}_{\mathrm{R}}$ as overall naturalness score without weights is also less accurate (−5.01%/-5.98% in terms of SRCC/PLCC) than the proposed subjectively-inspired weighting strategy. These results further support the subjective observations found in the AGIN.

3) Performances With Different Backbones: Since deploying different backbones within the rationality perceiving branch can affect the calculation of WSD, we merely probe the impact of different network architectures and model sizes on the technical prior branch of JOINT++. Four representative backbones, *i.e.*, ResNet [77], ViT [98], ConvNext [99], and Swin Transformer [72], are included. The results are listed in Tab. VII, from which we make two valuable observations. First, equipped with more sophisticated backbone networks deliver better performance. We can observe that the performance improves as the FLOPs of the backbones increase. ResNet50, ViT-Tiny, ConvNext-Tiny, and Swin Transformer-Tiny are with FLOPs around 4.5×10^9 (image resolution at 224×224), while their *base* versions are about 15.4×10^9. Second, replacing the backbone network of technical branch has a minimal impact on the prediction accuracy of the rationality branch (< 0.01 decay in SRCC except for ResNet), since they do not share parameters and only partial parameters are affected during the overall loss optimization process.

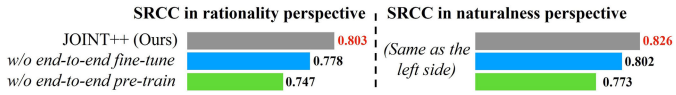


Fig. 16. SRCC performance of different parameters, compared with our method (JOINT++).

TABLE VIII
ABLATION STUDY OF LEARNING OBJECTIVES

Loss Functions		Technical		Rationality		Naturalness	
$\mathcal{L}_{\mathrm{IS}}$	$\mathcal{L}_{\mathrm{FS}}$	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
✓		0.817	0.824	0.756	0.771	0.799	0.803
	✓	0.831	0.834	0.786	0.794	0.819	0.828
✓	✓	**0.835**	**0.843**	**0.803**	**0.813**	**0.826**	**0.836**

4) Effects of Pre-Train&Fine-Tune Scheme: In Fig. 16, we evaluate the effects of the pre-train&fine-tune scheme applied in the rationality perceiving branch compared to direct training on the AGIN dataset (*w/o end-to-end pre-train*) and merely linear regression on pre-trained features (*w/o end-to-end finetune*). The large-scale pre-training on the image aesthetics dataset, AVA [73], contributes to the performance by about 7.5% and 6.9% in terms of rationality and naturalness perspectives, respectively. The end-to-end fine-tuning also leads to nearly 3.2% and 2.9% improvements. Both processes undoubtedly promote the model to integrate general high-level semantics knowledge with rationality-concerned visual information, leading to more accurate and reliable predictions.

5) Effects of Different Learning Objectives: In Tab. VIII, we further validate whether the extra objective ($\mathcal{L}_{\mathrm{FS}}$) can improve the subjective consistency of overall naturalness prediction. By combining $\mathcal{L}_{\mathrm{IS}}$ with $\mathcal{L}_{\mathrm{FS}}$, JOINT++ achieves +0.48%, +2.16%, and +0.85% performance gain in SRCC for three perspectives, respectively. In addition, even directly supervised by the overall naturalness MOS labels can also achieve comparable performance, suggesting the feasibility of explicitly modeling human perception of naturalness into an approximated sum of technical quality and rationality perspectives.

F. Can Re-Permuted Patches Preserve Semantics?

Prior studies [69], [70] show that technical quality perception should consider global semantics to better measure

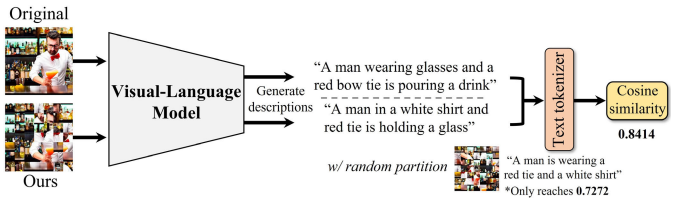


Fig. 17. An example of measuring the degree of semantic information retention.

TABLE IX
THE SRCC RESULTS OF CROSS-DATASET EVALUATION

Training Set	AGIQA-3K (Quality)		AGIN (Technical Quality)	
Testing Set	AGIN	AGIQA-3k	AGIQA-3k	AIGCIQA2023
HyperIQA	0.722	0.829	0.817	0.823
UNIQUE	0.736	0.833	0.819	0.831
JOINT (T-branch)	0.727	0.824	0.806	0.816
JOINT (R-branch)	0.675	0.766	0.757	0.749

TABLE X
THE SRCC AND PLCC RESULTS OF GENERALIZATION EVALUATION

Method	KADID-10k		KonIQ-10k		PIPAL	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
HyperIQA	0.8396	0.8505	0.9057	0.9183	0.3768	0.3852
UNIQUE	0.8704	0.8751	0.8954	0.9007	0.4311	0.4367
JOINT (T-branch)	0.8573	0.8628	0.8775	0.8854	0.3733	0.3856
JOINT (R-branch)	0.8153	0.8179	0.8363	0.8466	0.3662	0.3583

distortion levels. Here, we discuss this unclear question: can re-permuted images in the technical prior branch retain semantic information that can still be recognized by deep neural networks? Specifically, we evaluate its ability by conducting experiments: we first utilize the vision-language model to generate descriptions for the original and re-permuted images, respectively, and then tokenize them to calculate the cosine similarity. Fig. 17 shows an example of the procedures. It can be observed that our perceptual artifacts-guided patch partition strategy reaches 0.8414 cosine similarity, while random partition is about 10.3% lower. As for the whole AGIN, the average similarity score of our strategy is 0.7323 (on average 14.7% higher than *w/ random partition*). This indicates that it can preserve weak global semantics to distinguish textures from noises while guiding the model to focus more on low-level technical distortions.

G. Cross-Dataset & Generalization Evaluation

We further test the applicability of the AGIN dataset and the generalization ability of JOINT, shown in Tab. IX and Tab. X. Two representative IQA methods, *i.e.*, HyperIQA [91] and UNIQUE [93], are employed for comparison. Concretely, we select another two AIGIQA datasets, AGIQA-3K [39] and AIGCIQA2023 [36], to evaluate the performance of JOINT under the cross-dataset setting. Since these two datasets do not satisfy the training requirements of JOINT++, we test on their common dimension, namely the quality dimension and crop the rationality perceiving branch of our JOINT, leaving the technical branch to be trained (T-branch). Conversely, R-branch denotes that only the rationality branch to be trained. It is noteworthy that the performance of T-branch is slightly worse than the others since it is only composed of a Swin-T backbone followed by a simple regressor without other special designs. The performance in R-branch is significantly worse than the others, since they own different learning preferences. Moreover, we observe that models trained on AGIN show superior applicability when testing on other AIGIQA datasets (avg. SRCC $>$ 0.818 in same evaluation dimension). To evaluate the generalization ability of JOINT, we conduct experiments on three natural scene IQA datasets, *i.e.*, KADID-10k [32], KonIQ-10k [33], and PIPAL [34]. Specifically, PIPAL gathers and annotates images enhanced by various GAN-based image restoration algorithms, which do not follow the natural image distribution and exhibit quite different from the other datasets. It is clear from Tab. X that our JOINT (T-branch) is on par with UNIQUE and HyperIQA on KADID-10k and KonIQ-10k that achieves on average 0.8674 and 0.8741 in terms of SRCC and PLCC respectively. While the performance of JOINT (R-branch) is reduced due to its special design, which reduces the consideration of low-level distortions. We also notice that none of the tested methods presents promising results on PIPAL, suggesting the inapplicability of models under such particular algorithm-dependent distortions.

VI. CONCLUSION AND FUTURE WORK

In this paper, we contribute the AGIN dataset and the first subjective evaluation aimed at exploring the impact of technical and rationality perspectives on the naturalness of AIGIs. Besides, we propose JOINT, an objective naturalness evaluator that achieves higher alignment with human opinions against existing IQA and IAA approaches. Our work benefits the community by 1) presenting AGIN, which enables research on benchmarking and evaluating the naturalness of AIGIs by multi-dimensional human ratings; 2) encouraging new research on the naturalness assessment of AIGIs via analysis of technical and rationality features; 3) promoting the development of better naturalness assessment algorithms for AIGIs or other forms of AI-generated multimedia.

New possibilities of basic visual naturalness modeling advancement can be tried in a number of major directions. First, more fine-grained single-dimensional naturalness perturbations are needed for future advanced naturalness evaluation. Currently, naturalness distortions basically accompany multiple types of technical or rationality distortions, reducing the requirement of fine-grained perception for models. With the development of image generation models, the quality of synthesized images continues to improve, albeit showing unnaturalness only in specific small aspects, posing more stringent demands on the model's sensitivity to naturalness distortions. In addition, image naturalness assessment can be also used to distinguish whether the image is generated by AI. Further extensions can be made to evaluate the naturalness of face images (portraits), which stand for an important branch of image assessment, as well as assess the naturalness of AI-generated videos at frame-level.

REFERENCES

[1] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.

[4] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5143–5153.

[5] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2085–2094.

[6] F. Zhan et al., "Bi-level feature alignment for versatile image translation and manipulation," in *Proc. 17th Eur. Conf. Comput. Vis.*, Cham, Switzerland. Springer, 2022, pp. 224–241.

[7] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Feb. 2023.

[8] C. Jiang, F. Gao, B. Ma, Y. Lin, N. Wang, and G. Xu, "Masked and adaptive transformer for exemplar based image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22418–22427.

[9] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. van Gool, "RePaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11461–11471.

[10] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "MAT: Mask-aware transformer for large hole image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10758–10768.

[11] X. Kang, T. Yang, W. Ouyang, P. Ren, L. Li, and X. Xie, "DDColor: Towards photo-realistic image colorization via dual decoders," in *Proc. IEEE/CVF Int. Conf. Comp. Vis.*, Oct. 2023, pp. 328–338.

[12] M. Xia, W. Hu, T.-T. Wong, and J. Wang, "Disentangled image colorization via global anchors," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–13, Dec. 2022.

[13] H. Wang, D. Zhai, X. Liu, J. Jiang, and W. Gao, "Unsupervised deep exemplar colorization via pyramid dual non-local attention," *IEEE Trans. Image Process.*, vol. 32, pp. 4114–4127, 2023.

[14] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your GAN: Interactive point-based manipulation on the generative image manifold," in *Proc. ACM SIGGRAPH Conf.*, Jul. 2023, pp. 1–11.

[15] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18392–18402.

[16] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "Magicbrush: A manually annotated dataset for instruction-guided image editing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 31428–31449.

[17] Z. Lu et al., "Seeing is not always believing: Benchmarking human and model perception of AI-generated images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 25435–25447.

[18] H. de Ridder, F. J. Blommaert, and E. A. Fedorovskaya, "Naturalness and image quality: Chroma and hue variation in color images of natural scenes," *Proc. SPIE*, vol. 2411, pp. 51–61, Apr. 1995.

[19] M. Cadfk and P. Slavik, "The naturalness of reproduced high dynamic range images," in *Proc. 9th Int. Conf. Inf. Visualisation (IV)*, Jul. 2005, pp. 920–925.

[20] S. Y. Choi, M. Luo, M. Pointer, and P. Rhodes, "Investigation of large display color image appearance—III: Modeling image naturalness," *J. Imag. Sci. Technol.*, vol. 53, no. 3, p. 31104, May 2009.

[21] K. Gu et al., "Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 432–443, Mar. 2016.

[22] B. Yan, B. Bare, and W. Tan, "Naturalness-aware deep no-reference image quality assessment," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2603–2615, Oct. 2019.

[23] Q.-T. Le, P. Ladret, H.-T. Nguyen, and A. Caplier, "Study of naturalness in tone-mapped images," *Comput. Vis. Image Understand.*, vol. 196, Jul. 2020, Art. no. 102971.

[24] Y. Liang et al., "Rich human feedback for text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 19401–19411.

[25] H. Sheikh. (2005). *Live Image Quality Assessment Database Release 2*. [Online]. Available: http://live. ece. utexas. edu/research/quality

[26] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Mod. RadioElectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[27] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.

[28] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[29] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2015.

[30] K. Ma et al., "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.

[31] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, Jan. 2017.

[32] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.

[33] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.

[34] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong, "PIPAL: A large-scale image quality assessment dataset for perceptual image restoration," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 633–651.

[35] Z. Zhang, C. Li, W. Sun, X. Liu, X. Min, and G. Zhai, "A perceptual quality assessment exploration for AIGC images," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2023, pp. 440–445.

[36] J. Wang, H. Duan, J. Liu, S. Chen, X. Min, and G. Zhai, "Aigciqa2023: A large-scale image quality assessment database for ai generated images: From the perspectives of quality, authenticity and correspondence," in *Proc. CAAI Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2023, pp. 46–57.

[37] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 36652–36663.

[38] J. Xu et al., "Imagereward: Learning and evaluating human preferences for text-to-image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, Oct. 2023, pp. 15903–15935.

[39] C. Li et al., "AGIQA-3K: An open database for AI-generated image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6833–6846, Aug. 2024.

[40] J. Yuan et al., "PKU-AIGIQA-4K: A perceptual quality assessment database for both text-to-image and image-to-image AI-generated images," 2024, *arXiv:2404.18409*.

[41] H. de Ridder, "Naturalness and image quality: Saturation and lightness variation in color images of natural scenes," *J. Imag. Sci. Technol.*, vol. 40, no. 6, pp. 487–493, Nov. 1996.

[42] S. Wang, J. Zheng, H. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.

[43] Q.-T. Le, P. Ladret, H.-T. Nguyen, and A. Caplier, "Computational analysis of correlations between image aesthetic and image naturalness in the relation with image quality," *J. Imag.*, vol. 8, no. 6, p. 166, Jun. 2022.

[44] Y. Liu et al., "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 929–943, Apr. 2019.

[45] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, Jan. 2016, pp. 2234–2242.

[46] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6629–6640.

[47] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," 2018, *arXiv:1801.01401*.

[48] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 3927–3936.

[49] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," 2021, *arXiv:2104.08718*.

[50] M. Otani et al., "Toward verifiable and reproducible human evaluation for text-to-image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14277–14286.

[51] T. Zhou, S. Tan, W. Zhou, Y. Luo, Y.-G. Wang, and G. Yue, "Adaptive mixed-scale feature fusion network for blind AI-generated image quality assessment," *IEEE Trans. Broadcast.*, vol. 70, no. 3, pp. 833–843, Sep. 2024.

[52] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.

[53] *Stable Diffusion V1.5*. Accessed: Nov. 30, 2023. [Online]. Available: https://huggingface.co/runwayml/stable-diffusion-v1-5

[54] *Stable Diffusion V2.1*. Accessed: Nov. 20, 2023. [Online]. Available: https://huggingface.co/stabilityai/stable-diffusion-2-1

[55] *Midjourney*. Accessed: Nov. 20, 2023. [Online]. Available: https://www.midjourney.com

[56] *Dreamlike.art*. Accessed: Nov. 20, 2023. [Online]. Available: https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0

[57] *Realistic-Vision-1.4*. Accessed: Nov. 20, 2023. [Online]. Available: https://huggingface.co/SG161222/Realistic_Vision_V1.4

[58] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, "Predicting the quality of images compressed after distortion in two steps," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5757–5770, Dec. 2019.

[59] Z. Chen et al., "GAIA: Rethinking action quality assessment for AI-generated videos," 2024, *arXiv:2406.06087*.

[60] S. He, A. Ming, Y. Li, J. Sun, S. Zheng, and H. Ma, "Thinking image color aesthetics assessment: Models, datasets and benchmarks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21838–21847.

[61] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500-11, Int. Telecommun. Union, Geneva, Switzerland, 2002.

[62] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar, "Responsible research with crowds: Pay crowdworkers at least minimum wage," *Commun. ACM*, vol. 61, no. 3, pp. 39–41, Feb. 2018.

[63] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Commun. Methods Measures*, vol. 1, no. 1, pp. 77–89, Apr. 2007.

[64] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.

[65] H. Wu et al., "Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 1045–1054.

[66] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends Neurosci.*, vol. 15, no. 1, pp. 20–25, Jan. 1992.

[67] J. Norman, "Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches," *Behav. Brain Sci.*, vol. 25, no. 1, pp. 73–96, Feb. 2002.

[68] C. A. Párraga, G. Brelstaff, T. Troscianko, and I. R. Moorehead, "Color and luminance information in natural scenes," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 15, no. 3, pp. 563–569, 1998.

[69] H. Wu et al., "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland. Springer, 2022, pp. 538–554.

[70] H. Wu et al., "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 20144–20154.

[71] L. Zhang et al., "Perceptual artifacts localization for image synthesis tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7579–7590.

[72] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[73] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2408–2415.

[74] L. Bar, N. Sochen, and N. Kiryati, "Semi-blind image restoration via Mumford–Shah regularization," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 483–493, Feb. 2006.

[75] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[76] X. Liao, B. Chen, H. Zhu, S. Wang, M. Zhou, and S. Kwong, "DeepWSD: Projecting degradations in perceptual space to Wasserstein distance in deep feature space," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 970–978.

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[78] Z. Chen et al., "BAND-2k: Banding artifact noticeable database for banding detection and quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6347–6362, Jul. 2024.

[79] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 29, Jun. 2020, pp. 3575–3585.

[80] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3677–3686.

[81] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, Dec. 2015.

[82] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 213–222.

[83] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," in *Proc. 31st Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2022, pp. 942–948.

[84] R. Yi, H. Tian, Z. Gu, Y.-K. Lai, and P. L. Rosin, "Towards artistic image aesthetics assessment: A large-scale dataset and a new method," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22388–22397.

[85] A. Ciancio, A. L. N. T. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, Jan. 2011.

[86] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5944–5958, Sep. 2022.

[87] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[88] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[89] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[90] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2018.

[91] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3667–3676.

[92] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 5148–5157.

[93] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Trans. Image Process.*, vol. 30, pp. 3474–3486, 2021.

[94] S. Yang et al., "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1191–1200.

[95] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multi-task learning for generic and personalized image aesthetics assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 3898–3910, 2020.

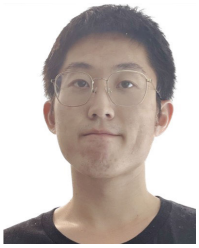
[96] J. Wang, K. Chan, and C. C. Loy, "Exploring CLIP for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 2555–2563.

[97] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14071–14081.

[98] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021.
[99] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11976–11986.

Zijian Chen (Graduate Student Member, IEEE) received the B.E. degree from Wenzhou University, Wenzhou, China, in 2020, and the M.E. degree from East China University of Science and Technology, Shanghai, China, in 2023. He is currently pursuing the Ph.D. degree with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai. His research interests include image quality assessment, perceptual signal processing, and machine learning.

Wei Sun (Member, IEEE) received the B.E. degree from East China University of Science and Technology, Shanghai, China, in 2016, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, in 2023. He is currently a Post-Doctoral Fellow with Shanghai Jiao Tong University. His research interests include image quality assessment, perceptual signal processing, and mobile video processing.



Haoning Wu received the B.S. degree from the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree with the S-Laboratory, School of Computer Science and Engineering, Nanyang Technological University, Singapore, supervised by Prof. Weisi Lin. His research interests include video quality assessment, including improving its robustness, efficiency, and interpretability and connecting it with related tasks.

Zicheng Zhang (Student Member, IEEE) received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Electronic Information and Electrical Engineering. His research interests include image quality assessment, video quality assessment, and 3D visual quality assessment. He is a reviewer for IEEE TRANSACTIONS ON MULTIMEDIA and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Jun Jia received the B.S. degree in computer science and technology from Hunan University, Changsha, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include computer vision and image processing.

Ru Huang received the B.S. degree from Nanjing University, Nanjing, China, in 1999, and the Ph.D. degree in circuits and systems from Shanghai Jiao Tong University, Shanghai, China, in 2008. From March 2015 to March 2016, he was a Visiting Scholar with the University of Wisconsin–Madison, Madison, WI, USA. He is currently an Associate Professor of electronics and communication engineering with East China University of Science and Technology, Shanghai. His research interests include wireless sensor networks, complex networks, and deep learning.

Xiongkuo Min (Member, IEEE) received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018. From June 2018 to September 2021, he was a Post-Doctoral Researcher with Shanghai Jiao Tong University. From January 2016 to January 2017, he was a Visiting Student with the University of Waterloo. From January 2019 to January 2021, he was a Visiting Scholar with The University of Texas at Austin and the University of Macau. He is currently a tenure-track Associate Professor with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. His research interests include image/video/audio quality, quality of experience, multimedia, image/video processing, and computer vision. He received the Best Paper Runner-Up Award of IEEE TRANSACTIONS ON MULTIMEDIA in 2021, the Best Student Paper Award of IEEE International Conference on Multimedia and Expo (ICME) in 2016, the Best Paper Award of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) in 2022, and several first place awards of grand challenges held at IEEE ICME and ICIP.

Guangtao Zhai (Senior Member, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen–Nürnberg, Germany. He is currently a Research Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. His research interests include multimedia signal processing and perceptual signal processing. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012.

Wenjun Zhang (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1984, 1987, and 1989, respectively. After three years working as an Engineer with Philips, Nürnberg, Germany, he went back to his Alma Mater in 1993 and became a Full Professor of electronic engineering at Shanghai Jiao Tong University in 1995. He is also the Chief Scientist of Chinese Digital TV Engineering Research Centre (NERC-DTV), an industry/government consortium in DTV technology research and standardization, and the Director of the Cooperative Media Network Innovation Centre (CMIC), an excellence research cluster affirmed by Chinese Government. He was one of the main contributors to Chinese DTTB Standard (DTMB) issued in 2006. He holds 238 patents and published more than 130 papers in international journals and conferences. His research interests include video coding and wireless transmission, multimedia semantic analysis, and broadcast/broadband network convergence.