Adaptive Mixed-Scale Feature Fusion Network for Blind AI-Generated Image Quality Assessment

Tianwei Zhou, Songbai Tan, Wei Zhou, Yu Luo, Yuan-Gen Wang, and Guanghui Yue

***Abstract*—With the increasing maturity of the text-to-image and image-to-image generative models, AI-generated images (AGIs) have shown great application potential in advertisement, entertainment, education, social media, etc. Although remarkable advancements have been achieved in generative models, very few efforts have been paid to design relevant quality assessment models. In this paper, we propose a novel blind image quality assessment (IQA) network, named AMFF-Net, for AGIs. AMFF-Net evaluates AGI quality from three dimensions, i.e., "visual quality", "authenticity", and "consistency". Specifically, inspired by the characteristics of the human visual system and motivated by the observation that "visual quality" and "authenticity" are characterized by both local and global aspects, AMFF-Net scales the image up and down and takes the scaled images and original-sized image as the inputs to obtain multi-scale features. After that, an Adaptive Feature Fusion (AFF) block is used to adaptively fuse the multi-scale features with learnable weights. In addition, considering the correlation between the image and prompt, AMFF-Net compares the semantic features from text encoder and image encoder to evaluate the text-to-image alignment. We carry out extensive experiments on three AGI quality assessment databases, and the experimental results show that our AMFF-Net obtains better performance than nine state-of-the-art blind IQA methods. The results of ablation experiments further demonstrate the effectiveness of the proposed multi-scale input strategy and AFF block.**

***Index Terms*—AI-generated images, blind image quality assessment, adaptive feature fusion, multi-scale feature.**

I. Introduction

With the advent of the Web3.0 era [1], artificial intelligence-generated Content (AIGC) is quietly leading a profound change, reshaping and even subverting the production and consumption mode of digital content. As a major branch of AIGC, AI-Generated Images (AGIs) have shown great application potential in various aspects of human life. The creation of AGIs involves inputting text prompts into the text-to-image generative model or inputting an image into the image-to-image generative model to facilitate the generation process [2]. However, due to technical limitations, large quality variance exists among different AGIs, requiring manual selection before use, which greatly limits the development of generative techniques. Therefore, to improve production efficiency, it is of great significance to design an automatic AGI image quality assessment (IQA) method [3], [4].

Recently, deep neural networks (DNNs), especially Convolutional Neural Networks (CNNs) and Transformers, have been widely employed in IQA tasks due to their outstanding capabilities in feature extraction and fitting [5]–[10]. Current works mainly focus on natural scene images (NSIs) that include synthesized distortions or authentic distortions. Most existing DNN-based IQA methods started with automatically mining quality-aware features through newly designed shallow network architectures [11] or classical network architectures used in image classification tasks [12]. To strengthen the representative capabilities of DNN features, designers also focused on the development of more comprehensive feature extraction or fusion approaches [13]–[16]. Later, for more accurate assessment results, some strategies are used in view of the distortion knowledge during the network design, such as setting auxiliary tasks for naturalness evaluation [17], combing local and global features [18], generating pseudo reference [19], integrating multi-level features [5], etc. In addition, some works also considered the characteristics of the human visual system (HVS) in the process of network design, e.g., establishing perception rule [20], including visual saliency prediction as the auxiliary task [21], modeling the attention and contrast sensitivity mechanisms [22]. Recently, some works also proposed to utilize rank learning [23], semi-supervised learning [24], and contrastive learning [25] for robust feature representation, striving for more accurate assessment results.

However, unlike camera-captured NSIs, AGIs are directly generated by AI generative models. Fig. 1 shows a simple comparison between NSIs and AGIs. Generally, NSIs usually suffer from distortions (e.g., compression, blurriness, noise, etc.), and the perceptual quality is rated from what is the level of visual experience. By contrast, the quality definition and representation of AGIs are different and usually rated in a multi-dimensional perspective, typically in the form of visual quality, authenticity, and consistency [2]. For an AGI, visual quality is similar to the perceptual quality of NSIs that is rated by analyzing the visual experience influenced by distortions, such as compression and color artifacts. Authenticity measures the realness degree to reality. Consistency is gauged by the alignment between image content and textual labels. Generally, the rated scores of these three dimensions are usually different. For instance, in the last subfigure of Fig. 1, the generated bald eagle is clear, without obvious distortions, and the image content conforms to the text description. But the image doesn't give enough details about the bald eagle with stiff hair and coarse structures, making it easy to identify as fake. Therefore, the rating scores of visual quality and consistency of this image are high, while the rating score of authenticity is low. In practical applications, apart from the visual quality affected by the distortion and authenticity affected by the artistic expression, users also concern about how well the generated AGI matches the task, i.e., content consistency. Since traditional NSI oriented IQA methods can only evaluate the image quality in the dimension of visual experience, they are not suitable for the

quality assessment task of AGIs. Therefore, it is necessary to design a comprehensive quality evaluation method to form more detailed understanding of AGIs.

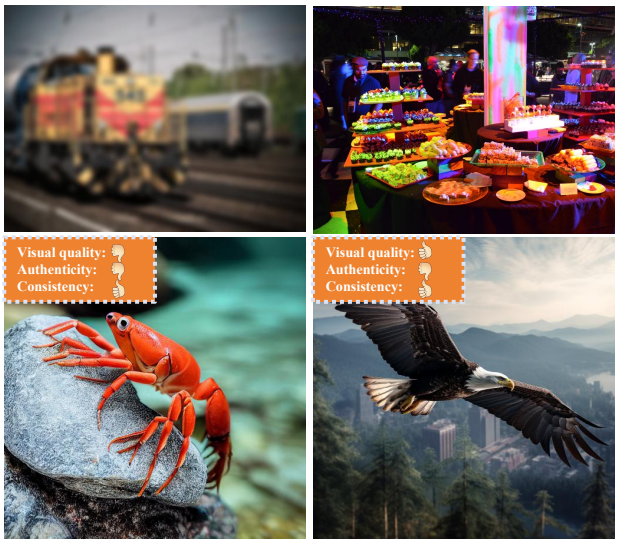

Fig. 1. Comparisons between NSIs (the first row, selected from KADID-10k [26] and KonIQ-10k [27]) and AGIs (the second row, selected from PKU-I2IQA [28]). The quality score of NSIs is mainly rated in the dimension of visual experience affected by distortions, while the quality scores of AGIs are rated in the dimensions of visual quality affected by distortions, authenticity affected by realness degree to reality, and consistency affected by the alignment between image content and textual labels.

As a new topic, the research on AGI quality assessment is still in infancy, with very limited progress. A widely used strategy for evaluating the quality of AGIs is calculating the distance between the NSIs groups and the AGIs groups [29], [30]. However, such a strategy must evaluate a group of images, which is unsuitable for the case of only one image. Therefore, more advanced methods should be specifically proposed. Following the general steps of IQA tasks, the quality assessment databases were first reported based on subjective experiments to promote the development of objective IQA methods [3], [28], [31]. These reported databases contain text prompts and multi-dimensional quality labels. Alongside these databases, some mainstream NSI oriented IQA methods were tested by respectively retraining them multiple times, each with a specific dimension of quality labels as the ground truth. As a result, most existing methods perform poorly, especially on the task of evaluating content consistency, as they only take the image as the input and cannot measure the mismatch degree between the text prompt and generated image. Recently, some works proposed to process the text prompt and image separately with different encoders and concatenate the extracted features from two encoders to generate the quality score [32]. However, directly concatenating features cannot effectively measure the semantic difference between the text prompt and image. For detailed alignment measurement between the text prompt and image, some researches tried to segment the text prompt into multiple morphemes, cut the image into multiple patches, and compute the alignment scores between morphemes and sub-images one by one [3]. However, image cutting is highly depended on the designer's experience and how to build the correspondence between morphemes and sub-images is unclear. To sum up, current studies mainly stay at benchmarking these databases with some mainstream IQA methods, lacking in-depth research on designing AGI quality assessment methods.

To move this field forward, this paper proposes a novel Adaptive Mixed-Scale Feature Fusion Network (AMFF-Net) for blind AGI quality assessment. Specifically, AMFF-Net adopts a multi-task framework and evaluates the quality of AGIs from three dimensions: distortion, authenticity, and content consistency. Considering that the subjective evaluation result of an image varies when the distance of the image plane from the observer changing, AMFF-Net scales the AGI up and down and feeds the scaled images and the original-sized image into the image encoder of the pre-trained Contrastive Language-Image Pre-Training (CLIP) model [33]. After that, the extracted multi-scale features are adaptively fused by an Adaptive Feature Fusion (AFF) block. The fused features contain information at different scales of the image, thereby being more effective for characterizing the distortion and authenticity of AGIs. For content consistency prediction, AMFF-Net uses the text encoder in the pre-trained CLIP to encode the text prompt and computes the similarity between the obtained textual features and the fused multi-scale features. The main contributions of this paper can be summarized as follows:

- A novel blind IQA method is proposed to comprehensively evaluate the quality of AGIs. Different from existing works that only measure the "visual quality" of an image, our method evaluates an AGI from the perspectives of "visual quality", "authenticity", and "consistency".
- Given that both local and global information should be considered during subjective rating of visual quality and authenticity, we propose to utilize a multi-scale input strategy to help the network capture image details at different levels of granularity.
- An AFF block is proposed to fuse multi-scale features. Different from current works that directly concatenate or add multi-scale features, the proposed block adaptively calculates the weights for different features, reducing the risk of information masking caused by concatenation and addition.
- Extensive experiments on three AGI quality assessment databases show that AMFF-Net achieves superior results compared to nine state-of-the-art blind IQA methods. Ablation experiments further demonstrate the effectiveness of the multi-scale input strategy and the AFF block.

II. Related Work

A. DNN-based Blind IQA Methods for NSIs

In the past decades, DNN-based blind IQA methods have attracted increasing attention from scholars, obtaining superior performance over traditional handcrafted feature-based methods on multiple tasks [16], [34]. In earlier studies, shallow CNNs were constructed to build the mapping between image patches and quality scores [11]. For instance, Bosse *et al*. [13] proposed a CNN-based blind IQA network, which consists of only

ten convolutional layers and five pooling layers for feature extraction and two fully connected layers for quality regression. Since the constructed shallow networks are usually with small input size, they have limited ability to characterize global distortions [35]. For more accurate predictions, latter works mainly utilized the pre-trained CNNs, Transformers, or the hybrid for the image classification tasks as the backbone of the IQA network. These networks usually have a relatively large input size, which is conducive to extract global information. Su *et al*. [20] utilized the pre-trained ResNet [36] as the backbone to extract local and global features and fused these features with a hyper network. Golestaneh *et al*. [18] extracted local information of the image via a pre-trained CNN and modeled them as a sequential input to a Transformer block for the non-local representation. Ke *et al*. [37] proposed a Transformer-based blind IQA network, in which the native resolution images with varying sizes are processed for a multi-scale image representation.

Generally, the aforementioned methods are purely data-driven and do not fully consider the characteristics of distortions or HVS, leaving much room for performance improvement. Recently, some researches proposed to build multi-task IQA frameworks for better performance, in which the auxiliary task is related to distortion understanding or HVS-inspired predictions. For instance, given that different distortions have diverse impacts on the perceptual quality, Wu *et al*. [38] designed a multi-task IQA network that takes image distortion type recognition as the auxiliary task. Yang *et al*. [21] set the visual saliency prediction as the auxiliary task in view of that different regions in an image receive non-uniform visual attention as they exhibit different qualities. Considering that synthesized distortions can be characterized and quantified by natural scene statistics (NSS), Yan *et al*. [39] set the NSS feature prediction as the auxiliary task to help the main task, i.e., quality prediction, learn a better mapping between the image and its quality score. Song *et al*. [17] proposed a knowledge-guided blind IQA framework by integrating domain knowledge from NSS and HVS. However, simply utilizing multi-task learning may have limited feature representations when dealing with images with complex distortions. Sun *et al*. [40] proposed a staircase structure to hierarchically fuse low-level and high-level features for better feature representation.

B. DNN-based Blind IQA Methods for AGIs

Compared to NSIs, the quality research of AGIs is still in its infancy, with only a few explorations. For a long time, researchers mainly utilized the *Inception Scor*e proposed by Salimans *et al*. [41] for AGI quality evaluation. Considering the difference between NSIs and AGIs, the *Frechet Inception Distance* [29] and *Kernel Inception Distance* [30] were later designed to measure the quality of AGIs by calculating the distance between the AGI groups and the NSI groups. Since these metrics only evaluate the quality of AGIs in a single dimension and are not suitable for evaluating a single AGI, more advanced metrics are highly required. Zhang *et al*. [42] made one of the pioneering discussions on evaluating AGI quality in a multi-dimensional manner and suggested that the quality of AGIs should be measured in the aspects of technical issues, artificial intelligence, unnaturalness, degree of difference, and aesthetics. Unfortunately, this work does not propose an AGI quality evaluation algorithm, and the strategy of how to incorporate AGI distortion representation into the evaluation algorithm is not clear. More recently, some efforts have been paid to propose specific IQA methods for AGIs by drawing inspirations from NSI-oriented IQA methods. For instance, Yuan *et al*. [43] took the AGI as the input of the IQA network and compared the differences among various images for a better feature representation using a contrastive regression framework. Later, they [32] also introduced a text-image encoder-based regression framework that respectively processes the text prompts and generated images with a text encoder and an image encoder, and concatenates the extracted features for quality prediction. To evaluate the consistency between the text prompt and generated image, some current works leverage the strong reasoning abilities of large Language models (LLMs) for evaluation [44]–[46]. For example, LLMScore [45] generates quality scores with multi-granularity compositionality, which transforms the image into image-level and object-level visual descriptions, leveraging LLMs to evaluate text-to-image models. However, these LLMs based methods have a large number of parameters and require abundant labeled AGIs databases for training. This motivates to design simpler methods. Li *et al*. [3] proposed a simply StairReward alignment model, which segments the prompt into morphemes and divides the image into stairs to predict the final score through their one-to-one alignment.

Although these AGI quality assessment methods perform relatively superior performance over traditional NSI-oriented IQA methods, they are still not fully qualified for practical applications due to the following limitations. First, most methods only predict the visual experience of an image. In practice, before use, users not only comprehensively check the AGI from the visual experience and authenticity, but also concern about how well the generated image matches the task, i.e., content consistency. Second, most methods only take the AGI as the input, which is insufficient to measure the text-to-image consistency. Although the measurement of alignment degree between text morphemes and image stairs can reflect the consistency to some extent, the prompt segmentation and image cutting are highly dependent on the designer's experience, limiting the generalization ability of such methods. Third, current methods ignore the correlation and interaction between features from different modalities, usually resulting in unsatisfactory performance.

C. CLIP-based IQA Methods

In 2021, Radford *et al*. [33] trained and released the CLIP model based on 400 million picture-text pairs to enable zero-shot prediction. Since then, many researchers have proposed CLIP-based models for many visual tasks, including image classification [47], object detection [48], and image retrieval [49]. Thanks to the robust capabilities of semantic extraction, CLIP has recently been applied to the IQA tasks. Wang *et al*. [50] explored the rich prior knowledge and visual-perception

in CLIP, and evaluated the image quality through prompt engineering. Pan *et al*. [51] introduced a two-stage IQA model, in which a CLIP-based text encoder is used for quality-aware feature extraction. Miyata [52] introduced a CLIP-based IQA model that can identify both the perceived quality rating of the image and the reason on which the rating is based. Zhang *et al*. [53] utilized the CLIP model to measure semantic affinity in the digital human quality assessment task. Although increasing attempts made in designing CLIP-based IQA methods for NSIs, very few works have been reported for AGIs.

III. Proposed Method

A. Motivation

In this paper, we propose a simple yet effective multi-task blind IQA network, named AMFF-Net, for AGIs. The design motivation behind our AMFF-Net are as follows: 1) Inspired by the fact that the perceivability of image details is highly related to the distance of the image plane from the observer, AMFF-Net scales the AGI up and down, and encodes the scaled images and original-size image to capture image details at different levels of granularity, closer to the subtle and holistic nature of human visual perception. 2) With the multi-scale image representations, an adaptive feature fusion block is used to adaptively incorporate image details at different resolutions, contributing to accurate quality and authenticity predictions. 3) AMFF-Net evaluates the AGI from multi-dimensional perspectives, i.e., the visual quality, authenticity, and content consistency, targeting at helping the users understand the quality of AGIs more comprehensively than current methods that only provide one-dimensional prediction. Also, computing the similarity between textual features and image features requiring no designer's experience is a good choice to evaluate the consistency between the prompt and generated image.

B. Overall Architecture

Fig. 2 illustrates the overall architecture of the proposed AMFF-Net. AMFF-Net takes the text prompt and multi-scale AGIs as the inputs and outputs multi-attribute quality scores, including content consistency S_C, visual quality S_V, and authenticity S_A:

$$(S_C, S_V, S_A) = \mathcal{F}_\theta(T, I^{1.5\times}, I^{1.0\times}, I^{0.5\times}), \tag{1}$$

where $\mathcal{F}_\theta(\cdot,\cdot,\cdot,\cdot)$ denotes the mapping between inputs and outputs, and T is the text prompt. $I^{1.5\times}$, $I^{1.0\times}$, and $I^{0.5\times}$ are the scaled AGIs with $1.5\times$, $1.0\times$, and $0.5\times$ resolutions of the original image.

Specifically, considering that both local and global details affect the subjective ratings of visual quality and authenticity, AMFF-Net inputs the scaled AGIs, i.e., $I^{1.5\times}$, $I^{1.0\times}$, and $I^{0.5\times}$, into an image encoder of the pre-trained CLIP model [33] to obtain multi-scale semantic representations, denoted as $F_I^{1.5s} \in \mathbb{R}^{1\times1024}$, $F_I^{1.0s} \in \mathbb{R}^{1\times1024}$, and $F_I^{0.5s} \in \mathbb{R}^{1\times1024}$. In this study, we select ResNet50 [36] as the image encoder as it has been widely validated to be effective for IQA tasks. In the default settings of CLIP, the image encoder can only accept inputs with the size of 224×224 due to the presence of positional embedding. In this study, we add some operations on the last layer of ResNet50 to make it adaptive to inputs with different sizes. Specifically, for the input with larger size than 224×224, we down-sample the feature of the ResNet50's last layer using adaptive maximum pooling. For the input with smaller size than 224×224, we up-sample the feature of the ResNet50's last layer using the bilinear interpolation. Both the down-sampling and up-sampling operations aim to make the feature size of the ResNet50's last layer be (2048, 7, 7) to match the positional embedding. The obtained semantic features of different scaled images are then fused by an AFF block to form the merged feature $F_I \in \mathbb{R}^{1\times1024}$. Subsequently, F_I is fed into two parallel Multi-layer Perceptions (MLPs), which consists of two fully connected layers, to separately predict the scores of visual quality and authenticity:

$$\begin{cases} S_V = \mathcal{M}_{\vartheta1}(F_I), \\ S_A = \mathcal{M}_{\vartheta2}(F_I), \end{cases} \tag{2}$$

where $\mathcal{M}_{\vartheta1}(\cdot)$ and $\mathcal{M}_{\vartheta2}(\cdot)$ are the MLP operations for visual quality score prediction and authenticity score prediction, respectively. The number of neural nodes in the MLP is $\{1024, 256, 1\}$. For content consistency score prediction, we first utilize the Transformer-based text encoder [54] of the pre-trained CLIP model to encode the text prompt, and subsequently compute the cosine similarity between the extracted textural feature $F_T \in \mathbb{R}^{1\times1024}$ with the merged image feature F_I to predict the content consistency score:

$$S_C = \frac{F_I \odot (F_T)^T}{\|F_I\|_2 \, \|F_T\|_2}, \tag{3}$$

where $\odot$ denotes the matrix-multiplication, and $(\cdot)^T$ stands for the matrix transpose operation.

Overall, AMFF-Net can extract multi-scale representations and adaptively aggregate them, providing rich semantic information for accurate visual quality and authenticity predictions. In addition, AMFF-Net considers the interaction between the prompt and image, which is conducive to accurate content consistency prediction.

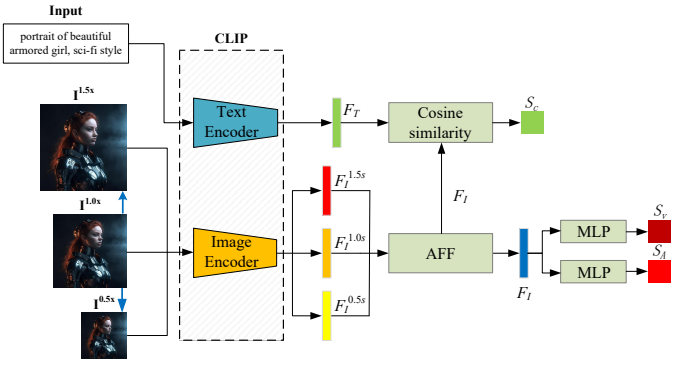

Fig. 2. Overview of the proposed AMFF-Net. It takes both the text prompt and three scaled AGIs as the inputs and outputs the consistency score S_C, visual quality score S_V, and authenticity score S_A. Here, the image and text encoders are selected from the pre-trained CLIP model. AFF and MLP denote the adaptive feature fusion block and multi-layer perception, respectively.

C. Adaptive Feature Fusion Block

By feeding three scaled AGIs into the image encoder, we can obtain multi-scale semantic representations, i.e., $F_I^{1.5s}$, $F_I^{1.0s}$, and $F_I^{0.5s}$. How to fuse these features is our next focus. In previous works, one widely used method is directly concatenating or adding these features. However, such a method is easy to cause information masking at different scales, which is unsuitable for the IQA task. In this study, we propose a novel feature fusion block, named AFF, to adaptively fuse these features.

Fig. 3 presents the architecture of the proposed AFF block. First, these multi-scale features are stacked together, resulting in $V_I \in \mathbb{R}^{3\times 1024}$. Then, V_I is processed by two linear layers and a Softmax function. Between two linear layers, a ReLU activation function is embedded. Next, the resultant feature is processed by a chunk operation, resulting in three distinct features, namely $A_I^{1.5s} \in \mathbb{R}^{1\times 1024}$, $A_I^{1.0s} \in \mathbb{R}^{1\times 1024}$, and $A_I^{0.5s} \in \mathbb{R}^{1\times 1024}$:

$$V_m = \Psi(\mathcal{U}(F_I^{0.5s}, F_I^{1.0s}, F_I^{1.5s})), \tag{4}$$

where $\Psi(\cdot)$ indicates the weights generation process, including the stack operation $\mathcal{U}$, linear layers, and a Softmax function. $A_I^{1.5}$, $A_I^{1.0}$, and $A_I^{0.5}$ represent the weights associated with the three image scales. With multi-scale semantic features and their scale-specific weights, we can obtain the fused feature $F_I \in \mathbb{R}^{1\times 1024}$ through the element-multiplication and element-addition operations:

$$F_I = A_I^{1.5s} \cdot F_I^{1.5s} + A_I^{1.0s} \cdot F_I^{1.0s} + A_I^{0.5s} \cdot F_I^{0.5s}. \tag{5}$$

This above process ensures an adaptive integration of multi-scale semantic features, boosting the feature representation.

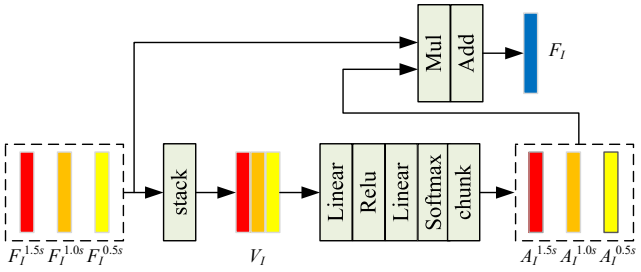

Fig. 3. Architecture presentation of the proposed AFF block.

D. Loss Function

As shown in Fig. 1, our AMFF-Net evaluates the quality in three dimensions. The overall loss function $\mathcal{L}$ for training AMFF-Net is a linear combination of three components:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_V + \mathcal{L}_A \tag{6}$$

where $\mathcal{L}_C$, $\mathcal{L}_V$, and $\mathcal{L}_A$ are the loss functions respectively used for three tasks, i.e., content consistency score prediction, visual quality score prediction, and authenticity score prediction, during network training. For the task of consistency score prediction, we utilize the fidelity loss function [55]:

$$\mathcal{L}_C = \frac{1}{N^2}\sum_{i,j\in N}\left(1-\sqrt{\mathcal{P}_{i,j}\widehat{\mathcal{P}}_{i,j}}-\sqrt{(1-\mathcal{P}_{i,j})(1-\widehat{\mathcal{P}}_{i,j})}\right), \tag{7}$$

where N is the image number in a mini-batch, and $\mathcal{P}_{i,j}$ is a binary function that compares the quality of the i-the and j-th images. If $Q_V^i \geq Q_V^j$, $\mathcal{P}_{i,j} = 1$; otherwise, $\mathcal{P}_{i,j} = 0$. Here, Q_V denotes the ground truth value of content consistency. $\widehat{\mathcal{P}}_{i,j}$ computes the probability of the j-th image predicted better than the j-th image using the Thurstone's model [56]. The reason why we choose the fidelity loss is that it can preserve information granularity of the cosine similarity between image and text features, making it suitable for our image-text consistency score prediction task. For the tasks of visual quality and authenticity score prediction, the mean square error is used as the loss function to train the network:

$$\mathcal{L}_V = \frac{1}{N}\sum_{i=1}^{N}(Q_V^n - S_V^n)^2, \tag{8}$$

$$\mathcal{L}_A = \frac{1}{N}\sum_{i=1}^{N}(Q_A^n - S_A^n)^2. \tag{9}$$

In Eq. (10) and Eq. (9), for a mini-batch with N images, Q_V^n and Q_A^n are the ground truth values of the n-th image's visual quality and authenticity, respectively.

IV. Experiments

A. Experiments Setup

1) Databases: In this study, we selected three public AGI quality assessment databases, including AGIQA-3K [3], AIGCIQA2023 [31], and PKU-I2IQA [28], for evaluating and comparing different blind IQA methods. Brief descriptions of these databases are given below.

- **AGIQA-3K**: It has 2,982 AGIs generated by 6 Text-to-Image generative models, including GLIDE [57], Stable Diffusion V1.5 [58], Stable Diffusion XL2.2 [58], AttnGAN [59], Midjourney [60], and DALLE2 [61]. It also provides 300 text prompts of different scenes and styles, and contains score labels of visual quality and consistency.
- **AIGCIQA2023**: It comprises 2,400 AGIs generated by six cutting-edge Text-to-Image generative models with 100 text prompts, including GLIDE [57], Lafite [62], Stable Diffusion [58], DALLE [61], Unidiffusion [63], and Controlnet [64]. For each image, it provides score labels of visual quality, authenticity, and consistency.
- **PKU-I2IQA**: It consists of 1,600 AGIs generated by two Image-to-Image models (Midjourney [60] and Stable Diffusion V1.5 [58]). It contains score labels of visual quality, authenticity, and consistency, and also provides 200 text prompts of different scenes and styles.

Fig. 4 shows some examples from the three AGI quality assessment databases. Following previous works, we randomly divide the AGIQA-3K into training set and test set in a ratio of 8:2. For AIGCIQA2023 and PKU-I2IQA, we conduct a 3:1 train-test split on the images generated by each generative

(a) (b) (c)

Fig. 4. Images from three AGI quality assessment databases: (a) AGIQA-3K [3], (b) AIGCIQA2023 [31], and (c) PKU-I2IQA [28].

model. The resolution of images in three databases is 512×512, and we resize each image into the size of 224×224 due to the limited computational source.

2) Evaluation Metrics: This study selects three widely used evaluation metrics in the IQA field to present quantitative results, including Spearman rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), and Kendall rank correlation coefficient (KRCC). In general, higher values (the maximum value is 1) of these metrics indicates better prediction performance. As suggested by VQEG [65], a four-parameter logistic function is used before computing PLCC:

$$\widetilde{s} = \frac{\kappa_1 - \kappa_2}{1 + exp(\kappa_4(s - \kappa_3))} + \kappa_2, \tag{10}$$

where $\widetilde{s}$ is the mapped value of a predicted score s. In Eq. (10), κ_i ($i \in \{1, 2, 3, 4\}$) can be obtained by comparing the predicted scores and their ground truths using the least square method.

3) Implementation Details: Our proposed AMFF-Net was implemented under the open source Pytorch repository. The server used in the experiments was powered by one NVIDIA Geforce GTX3090 GPU and two Intel XEON 6226R CPUs. During network training, the AdamW optimizer was used, and the batch size was set to 32. The model was trained end-to-end for 120 epochs in a two-stage manner. In the first stage, we froze the parameters of the text and image encoders in the CLIP and trained the remaining part of AMFF-Net in the first 20 epochs. The learning rates for AIGCIQA2023 and the other two databases were set to 1×10^{-3} and 5×10^{-4}. In the second stage, we unfroze the text and image encoders and fine-tuned the whole network with a learning rate of 5×10^{-6}. At the 80-th epoch, the learning rate was adjusted to 5×10^{-7}. In addition, we set an early stop strategy, and the training process was stopped if the performance did not improve after 20 epochs.

B. Performance Comparison

1) Prediction Ability Comparison: In this study, we compare our proposed AMFF-Net with nine state-of-the-art blind IQA methods, including ResNet50 [36], ViT-B/32 [66], MUSIQ [37], DB-CNN [67], HyperIQA [20], TReS [18], Re-IQA [68], StairIQA [40], and PSCR [43]. Among these methods, the first eight methods are designed for NSIs, while the last one is a method specifically designed for AGIs. All these methods, except PSCR, are re-trained on the three AGI quality assessment databases using their default settings. Since PSCR does not release the source code, we directly extract results from its original paper, in which only partial results on AGIQA-3K and AIGCIQA2023 are reported. Table I, Table II, and Table III tabulate the quantitative results of our proposed AMFF-Net and other methods on three databases. The results are the median values of 10 trials, in which a random train-test split is conducted according to the settings in Section IV-A1. For the convenience of comparison, the best results are marked in bold.

From the tables, we have the following observations. First, AMFF-Net outperforms these competing methods on AGIQA-3K and PKU-I2IQA databases. For example, AMFF-Net is ahead of the second best method (i.e., TReS) by approximately 2.367% and 18.018% in terms of SRCC when evaluating the visual quality and consistency on AGIQA-3K, respectively. It also achieves a performance gain by approximately 3.154%, 7.694%, and 3.639% than the second best method (i.e., HyperIQA) in terms of SRCC when evaluating visual quality, consistency, and authenticity on PKU-I2IQA. Second, our AMFF-Net ranks the second when evaluating visual quality and authenticity on AIGCIQA2023 and is slightly inferior to the best method (i.e., HyperIQA) by approximately 0.872% and 0.628%. A possible reason for this is that, the images (e.g., generated by the Lafite model [69]) in the AIGCIQA2023 possess too similar characteristics, while our method has limited capability to discriminate fine-grained differences between images. Despite this, it still exhibits superior performance than other methods in evaluating consistency and achieves the best average result of three dimensions, with 3.196% SRCC advantages over HyperIQA. Third, among all selected NSI oriented methods, HyperIQA performance best. A possible reason for this is that, HyperIQA evaluates the image quality in a content-aware manner using a hyper network, contributing to understand the semantic distortions, which get more attention during subjective rating of AGIs. Fourth, the specifically designed method (PSCR) for AGIs generally performs better than most traditional NSI-oriented IQA methods. This may be attributed to the fact that it compares two AGIs generated by the same text prompts during the network design. Nevertheless,

TABLE I
QUANTITATIVE RESULTS COMPARISON BETWEEN DIFFERENT BLIND IQA METHODS ON AGIQA-3K. FOR CONVENIENT VIEWING, WE ALSO PRESENT THE AVERAGE VALUE OF EACH EVALUATION METRIC ON TWO TASKS IN THE EIGHTH TO TENTH COLUMNS OF THE TABLE.

Method	AGIQA-3K [3]									FLOPs	#Params
	Quality			Consistency			Avg.				
	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC		
ResNet50 [36]	0.8252	0.8863	0.6396	0.6396	0.7878	0.4645	0.7324	0.8370	0.5521	**8.27G**	24.56M
ViT-B/32 [66]	0.8063	0.8687	0.6195	0.6171	0.7652	0.4438	0.7117	0.8170	0.5316	8.73G	87.61M
MUSIQ [37]	0.8349	0.8862	0.6496	0.6292	0.7839	0.4557	0.7321	0.8350	0.5526	124.77G	78.55M
DB-CNN [67]	0.8154	0.8747	0.6278	0.6329	0.7823	0.4567	0.7242	0.8285	0.5423	33.00G	**15.31M**
HyperIQA [20]	0.8400	0.8958	0.6565	0.6276	0.8087	0.4543	0.7338	0.8522	0.5554	8.67G	27.38M
TReS [18]	0.8367	0.8973	0.6531	0.6366	0.8134	0.4620	0.7366	0.8553	0.5575	16.77G	34.46M
Re-IQA [68]	0.8187	0.8799	0.6312	0.6373	0.7880	0.4606	0.7280	0.8339	0.5459	33.09G	40.29M
StairIQA [40]	0.8235	0.8864	0.6381	0.6348	0.8006	0.4600	0.7291	0.8435	0.5490	10.22G	33.01M
PSCR [43]	0.8498	0.9059	-	-	-	-	-	-	-	-	-
AMFF-Net(ours)	**0.8565**	**0.9050**	**0.6759**	**0.7513**	**0.8476**	**0.5663**	**0.8039**	**0.8763**	**0.6211**	41.48G	50.30M

TABLE II
QUANTITATIVE RESULTS COMPARISON BETWEEN DIFFERENT BLIND IQA METHODS ON AIGCIQA2023. FOR CONVENIENT VIEWING, WE ALSO PRESENT THE AVERAGE VALUE OF EACH EVALUATION METRIC ON THREE TASKS IN THE LAST THREE COLUMNS OF THE TABLE.

Method	AIGCIQA2023 [31]											
	Quality			Consistency			Authenticity			Avg.		
	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
ResNet50 [36]	0.8208	0.8408	0.6083	0.6997	0.6962	0.5038	0.7087	0.6964	0.5073	0.7431	0.7445	0.5398
ViT-B/32 [66]	0.7881	0.8140	0.5737	0.6404	0.6413	0.4557	0.6975	0.6920	0.4991	0.7087	0.7158	0.5095
MUSIQ [37]	0.8423	0.8596	0.6327	0.7620	0.7527	0.5591	0.7615	0.7509	0.5585	0.7886	0.7878	0.5835
DB-CNN [67]	0.8339	0.8577	0.6240	0.6837	0.6787	0.4915	0.7485	0.7436	0.5449	0.7554	0.7600	0.5535
HyperIQA [20]	**0.8483**	**0.8689**	**0.6389**	0.7541	0.7439	0.5517	0.7798	0.7718	**0.5766**	0.7940	**0.7949**	0.5890
TReS [18]	0.8436	0.8666	0.6357	0.7292	0.7266	0.5331	0.7661	0.7602	0.5655	0.7796	0.7845	0.5781
Re-IQA [68]	0.8144	0.8317	0.5980	0.6430	0.6355	0.4564	0.7224	0.7110	0.5214	0.7266	0.7261	0.5253
StairIQA [40]	0.8186	0.8450	0.6063	0.6641	0.6625	0.4748	0.7155	0.7131	0.5151	0.7328	0.7402	0.5321
PSCR [43]	0.8371	0.8588	-	0.7465	0.7379	-	**0.7828**	**0.7750**	-	0.7888	0.7906	-
AMFF-Net(ours)	0.8409	0.8537	0.6310	**0.7782**	**0.7638**	**0.5747**	0.7749	0.7643	0.5684	**0.7980**	0.7939	**0.5914**

TABLE III
QUANTITATIVE RESULTS COMPARISON BETWEEN DIFFERENT BLIND IQA METHODS ON PKU-I2IQA. FOR CONVENIENT VIEWING, WE ALSO PRESENT THE AVERAGE VALUE OF EACH EVALUATION METRIC ON THREE TASKS IN THE LAST THREE COLUMNS OF THE TABLE.

Method	PKUI2IQA [28]											
	Quality			Consistency			Authenticity			Avg.		
	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
ResNet50 [36]	0.6219	0.6328	0.4440	0.6601	0.6511	0.4792	0.4932	0.5211	0.3437	0.5917	0.6017	0.4223
ViT-B/32 [66]	0.5151	0.5197	0.3598	0.5507	0.5389	0.3862	0.4226	0.4485	0.2937	0.4962	0.5024	0.3465
MUSIQ [37]	0.6408	0.6410	0.4596	0.6379	0.6531	0.4599	0.6354	0.6505	0.4571	0.6380	0.6482	0.4589
DB-CNN [67]	0.5975	0.5970	0.4247	0.6083	0.5925	0.4345	0.5667	0.5820	0.4024	0.5908	0.5905	0.4205
HyperIQA [20]	0.6849	0.6955	0.4988	0.7239	0.7062	0.5378	0.6596	0.6902	0.4794	0.6894	0.6973	0.5053
TReS [18]	0.6374	0.6427	0.4572	0.6480	0.6456	0.4690	0.6003	0.6393	0.4311	0.6286	0.6425	0.4524
Re-IQA [68]	0.5996	0.6106	0.4248	0.5705	0.5690	0.4029	0.5509	0.5787	0.3884	0.5737	0.5861	0.4054
StairIQA [40]	0.5855	0.6038	0.4151	0.5739	0.5720	0.4048	0.5535	0.5879	0.3927	0.5709	0.5879	0.4042
AMFF-Net(ours)	**0.7065**	**0.7174**	**0.5169**	**0.7796**	**0.7708**	**0.5921**	**0.6836**	**0.7206**	**0.5002**	**0.7232**	**0.7363**	**0.5364**

it is inferior to our proposed AMFF-Net in most cases. To compare these methods more intuitively, Fig. 5 presents the scatter plots of different IQA methods on AGIQA-3K database. As seen, our AMFF-Net produces more consistent predictions with subjective ratings than competing methods.

We also compare our AMFF-Net with competing methods in terms of the floating-point operations (FLOPs) and the number of parameters (#Params). As shown in the last two columns of Table I, AMFF-Net has 41.48G of FLOPs and 50.30M of #Params, ranking eighth and seventh among ten competing methods, respectively. This indicates that, compared to its competitors, our AMFF-Net does not exhibit competitive advantage in these two aspects. One possible reason for this could be that AMFF-Net requires multi-scale images as input and needs to process features from three different scales in the inference stage. Despite this, our proposed AMFF-Net is more competent for the quality assessment tasks in terms of prediction accuracy.

2) Generalization Ability Comparison: Apart from prediction ability, generalization ability is another crucial factor for a blind IQA method. In this section, we further compare these selected methods in terms of generalization ability by conducting cross-validation experiments. Since the images in AGIQA-3K and AIGCIQA2023 are generated by Text-to-Image generative models, while those in PKU-I2IQA are generated by Image-to-Image generative models, we select AGIQA-3K and AIGCIQA2023 for the experiments. Specifically, each method is trained on one database and tested on the other database

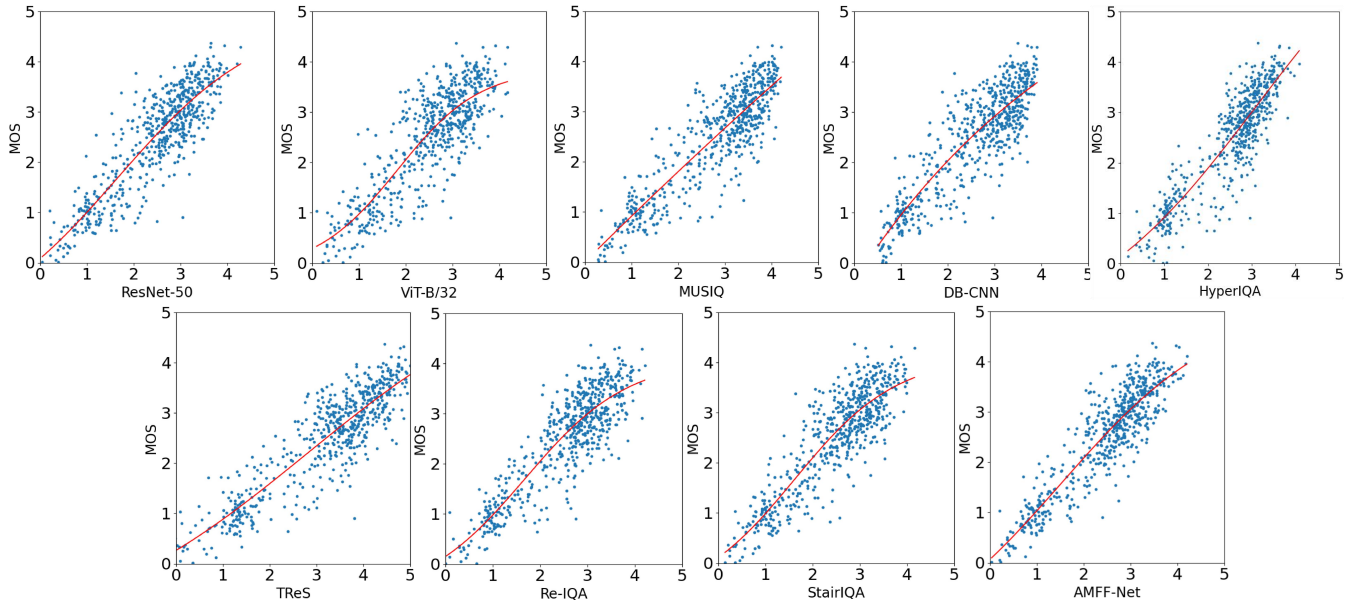

Fig. 5. Scatter plots of different IQA methods tested on the AGIQA-3K database. Due to the space limitation, we only show the predictions of visual quality.

TABLE IV
CROSS-VALIDATION RESULTS OF DIFFERENT BLIND IQA METHODS.

Method	AIGCIQA2023(Train) → AGIQA-3K(Test)						AGIQA-3K(Train) → AIGCIQA2023(Test)					
	Quality			Consistency			Quality			Consistency		
	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
RestNet50 [36]	0.576	0.612	0.397	0.473	0.523	0.326	0.599	0.609	0.413	0.432	0.435	0.305
ViT-B/32 [66]	0.509	0.571	0.350	0.434	0.496	0.302	0.517	0.529	0.357	0.381	0.390	0.259
MUSIQ [37]	0.635	0.673	0.445	0.394	0.437	0.265	0.650	0.643	0.444	0.525	0.515	0.359
DB-CNN [67]	0.627	0.688	0.442	0.390	0.435	0.263	0.654	0.664	0.455	0.470	0.460	0.317
HyperIQA [20]	**0.657**	0.692	0.443	0.418	0.465	0.280	0.669	**0.672**	0.472	0.464	0.431	0.314
TReS [18]	0.646	**0.702**	0.451	0.445	0.488	0.303	0.650	0.654	0.440	0.505	0.483	0.346
Re-IQA [68]	0.473	0.352	0.325	0.243	0.154	0.165	0.654	0.654	0.441	0.479	0.484	0.323
AMFF-Net(ours)	0.654	0.695	**0.459**	**0.554**	**0.624**	**0.385**	**0.678**	0.669	**0.474**	**0.546**	**0.549**	**0.381**

TABLE V
RESULTS OF ABLATION EXPERIMENTS ON THREE DATABASES. DUE TO THE SPACE LIMITATION, ONLY THE SRCC VALUES ARE PRESENTED HERE.

Method	AGIQA-3K [3]		AIGCIQA2023 [31]			PKU-I2IQA [28]		
	Quality	Consistency	Quality	Consistency	Authenticity	Quality	Consistency	Authenticity
w/o MSI	0.855	0.748	0.827	0.774	0.766	0.670	**0.781**	0.648
w/o AFF	0.852	0.739	0.835	0.768	0.774	0.694	**0.781**	0.680
AMFF-Net	**0.856**	**0.751**	**0.841**	**0.778**	**0.775**	**0.706**	0.780	**0.684**

when evaluating a specific quality of AGIs, i.e., visual quality and content consistency. Table IV presents the experimental results in the form of three evaluation metrics. Here, we do not consider StairIQA as it utilizes mixed databases for training, which is unfair for the cross-validation comparison. From the table, we can observe that our proposed AMFF-Net outperforms the competing blind IQA methods by a large margin, with higher SRCC, PLCC, and SRCC values in most cases. This indicates its higher generalization ability.

C. Ablation Experiments

1) Effectiveness of Each Component: In this study, the proposed AMFF-Net utilizes the multi-scale input strategy (MSI) and adaptive feature fusion (AFF) block for accurate quality prediction. Here, we conduct some ablation studies to investigate the effectiveness of the MSI strategy and the AFF block. The experimental settings are the same as the main experiment, and the median results of 10 trials are reported. Due to the space limitation, only the SRCC scores are given, as shown in Table V. The "w/o MSI" denotes the learned IQA model when using three original-sized images as the inputs. The "w/o AFF" is the learned IQA model when directly adding the multi-scale representations, instead of using the AFF block for adaptive fusion. From the table, it is clear that the absence of either the MSI strategy or the AFF block will degrade the prediction performance. For example, the removal of the MSI strategy and the AFF block leads to a SRCC drop of 0.713% and 1.665% respectively when evaluating visual quality on AIGCIQA2023. Similarly, AMFF-Net has a SRCC drop of 0.585% and 5.263% if we do not use the MSI strategy and

the AFF block separately when evaluating the authenticity on PKU-I2IQA. These results show that both the MSI strategy and the AFF block play a positive role in evaluating AGIs, contributing to achieving accurate predictions for AMFF-Net.

2) Effectiveness of the Similarity Metric: In Eq. (3), we choose cosine function to measure the similarity between text and image features. Here, we investigate its effectiveness on the AGI quality task by comparing it with two other similarity metrics: Euclidean distance and Manhattan distance. Fig. 6 shows the the SRCC results on three databases. It can be observed that cosine function achieves better results than Euclidean distance and Manhattan distance.

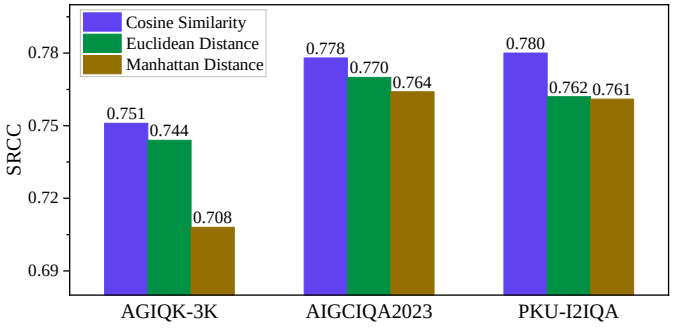

Fig. 6. SRCC results of three similarity metrics for consistency prediction on three databases.

D. Limitations and Future Works

Although our proposed AMFF-Net has demonstrated superiority over competing methods in quality assessment of AGIs, it still faces certain limitations. On the one hand, the prediction accuracy of AMFF-Net can be further improved. On the other hand, the computational complexity of AMFF-Net needs to be reduced. To update the proposed AMFF-Net, future works can be carried out from the following directions. Firstly, to improve the prediction accuracy, we will further strengthen the interaction between text features and image features, so that the model can get more discriminative features. Secondly, to reduce the computational complexity and accelerate inference speed, we will optimize the model structure by lightweight modules and design parallel computation schemes to simultaneously process different scaled inputs.

V. Conclusion

AGI quality assessment has recently emerged as a new and important topic, requiring comprehensive evaluation from multiple dimensions. In this paper, we propose a simple yet effective blind IQA network, termed AMFF-Net, for AGIs. Different from existing methods that only evaluate the images from the perspective of "visual quality", AMFF-Net utilizes a multi-task framework and aims to evaluate AGI quality from three dimensions, i.e., "visual quality", "authenticity", and "consistency". Specifically, considering that "visual quality" and "authenticity" are characterized by both local and global aspects, AMFF-Net utilizes a multi-scale input (MSI) strategy to capture image details at different levels of granularity. After that, an adaptive feature fusion (AFF) block is used to adaptively multi-scale features. To evaluate the content consistency, the similarity between semantic features of text prompts and AGI is computed. Through the cooperation of the MSI strategy and the AFF block, our AMFF-Net performs better than nine state-of-the-art blind IQA methods on three publicly available AGI quality assessment databases. In addition, ablations experiments demonstrate the effectiveness of the underlying concepts of the MSI strategy and the AFF block.

References

[1] W. Gan, Z. Ye, S. Wan, and P. S. Yu, "Web 3.0: The future of internet," *arXiv preprint arXiv:2304.06032*, 2023.

[2] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review," *Neural Networks*, vol. 144, pp. 187–209, 2021.

[3] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, and W. Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, accepted, in press, DOI: 10.1109/TCSVT.2023.3319020, 2023.

[4] S. Göring, R. R. R. Rao, R. Merten, and A. Raake, "Appeal and quality assessment for ai-generated images," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2023, pp. 115–118.

[5] X. Lan, M. Zhou, X. Xu, X. Wei, X. Liao, H. Pu, J. Luo, T. Xiang, B. Fang, and Z. Shang, "Multilevel feature fusion for end-to-end blind image quality assessment," *IEEE Transactions on Broadcasting*, vol. 69, no. 3, pp. 801–811, 2023.

[6] G. Yue, H. Wu, W. Yan, T. Zhou, H. Liu, and W. Zhou, "Subjective and objective quality assessment of multi-attribute retouched face images," *IEEE Transactions on Broadcasting*, accepted, in press, DOI: 10.1109/TBC.2024.3374043, 2024.

[7] B. Hu, T. Zhao, J. Zheng, Y. Zhang, L. Li, W. Li, and X. Gao, "Blind image quality assessment with coarse-grained perception construction and fine-grained interaction learning," *IEEE Transactions on Broadcasting*, accepted, in press, DOI:10.1109/TBC.2023.3342696, 2023.

[8] G. Yue, D. Cheng, H. Wu, Q. Jiang, and T. Wang, "Improving iqa performance based on deep mutual learning," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2182–2186.

[9] S. Lang, X. Liu, M. Zhou, J. Luo, H. Pu, X. Zhuang, J. Wang, X. Wei, T. Zhang, Y. Feng *et al.*, "A full-reference image quality assessment method via deep meta-learning and conformer," *IEEE Transactions on Broadcasting*, accepted, in press, DOI: 10.1109/TBC.2023.3308349, 2023.

[10] G. Yue, H. Wu, Q. Jiang, T. Zhou, W. Yan, and T. Wang, "Perceptual quality assessment of retouched face images," *IEEE Transactions on Multimedia*, accepted, in press, DOI: 10.1109/TMM.2023.3338412, 2023.

[11] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[12] B. Hu, G. Zhu, L. Li, J. Gan, W. Li, and X. Gao, "Blind image quality index with cross-domain interaction and cross-scale integration," *IEEE Transactions on Multimedia*, accepted, in press, DOI: 10.1109/TMM.2023.3303725, 2023.

[13] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[14] L. Zheng, Y. Luo, Z. Zhou, J. Ling, and G. Yue, "Cdinet: Content distortion interaction network for blind image quality assessment," *IEEE Transactions on Multimedia*, accepted, in press, DOI: 10.1109/TMM.2024.3360697, 2024.

[15] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[16] T. Zhou, S. Tan, B. Zhao, and G. Yue, "Multitask deep neural network with knowledge-guided attention for blind image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, accepted, in press, DOI: 10.1109/TCSVT.2024.3375344, 2024.

[17] T. Song, L. Li, J. Wu, Y. Yang, Y. Li, Y. Guo, and G. Shi, "Knowledge-guided blind image quality assessment with few training samples," *IEEE Transactions on Multimedia*, vol. 25, pp. 8145–8156, 2023.

[18] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1220–1230.

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

[20] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[21] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1383–1391.

[22] J. You and J. Korhonen, "Attention integrated hierarchical networks for no-reference image quality assessment," *Journal of Visual Communication and Image Representation*, vol. 82, p. 103399, 2022.

[23] F.-Z. Ou, Y.-G. Wang, J. Li, G. Zhu, and S. Kwong, "A novel rank learning based no-reference image quality assessment method," *IEEE Transactions on Multimedia*, vol. 24, pp. 4197–4211, 2022.

[24] G. Yue, D. Cheng, L. Li, T. Zhou, H. Liu, and T. Wang, "Semi-supervised authentically distorted image quality assessment with consistency-preserving dual-branch convolutional neural network," *IEEE Transactions on Multimedia*, vol. 25, pp. 6499–6511, 2023.

[25] Z. Yang, L. Li, Y. Yang, Y. Li, and W. Lin, "Multi-level transitional contrast learning for personalized image aesthetics assessment," *IEEE Transactions on Multimedia*, vol. 26, pp. 1944–1956, 2024.

[26] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[27] ——, "Koniq-10k: Towards an ecologically valid and large-scale iqa database," *arXiv preprint arXiv:1803.08489*, 2018.

[28] J. Yuan, X. Cao, C. Li, F. Yang, J. Lin, and X. Cao, "Pku-i2iqa: An image-to-image quality assessment database for ai generated images," *arXiv preprint arXiv:2311.15556*, 2023.

[29] 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," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[31] 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," *arXiv preprint arXiv:2307.00211*, 2023.

[32] J. Yuan, X. Cao, J. Che, Q. Wang, S. Liang, W. Ren, J. Lin, and X. Cao, "Tier: Text and image encoder-based regression for aigc image quality assessment," *arXiv preprint arXiv:2401.03854*, 2024.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[34] M. Zhou, L. Chen, X. Wei, X. Liao, Q. Mao, H. Wang, H. Pu, J. Luo, T. Xiang, and B. Fang, "Perception-oriented u-shaped transformer network for 360-degree no-reference image quality assessment," *IEEE Transactions on Broadcasting*, vol. 69, no. 2, pp. 396–405, 2023.

[35] G. Yue, C. Hou, W. Yan, L. K. Choi, T. Zhou, and Y. Hou, "Blind quality assessment for screen content images via convolutional neural network," *Digital Signal Processing*, vol. 91, pp. 21–30, 2019.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[37] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.

[38] Q. Wu, H. Li, K. N. Ngan, B. Zeng, and M. Gabbouj, "No reference image quality metric via distortion identification and multi-channel label transfer," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2014, pp. 530–533.

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

[40] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 6, pp. 1178–1192, 2023.

[41] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[42] Z. Zhang, C. Li, W. Sun, X. Liu, X. Min, and G. Zhai, "A perceptual quality assessment exploration for aigc images," *arXiv preprint arXiv:2303.12618*, 2023.

[43] J. Yuan, X. Cao, L. Cao, J. Lin, and X. Cao, "Pscr: Patches sampling-based contrastive regression for aigc image quality assessment," *arXiv preprint arXiv:2312.05897*, 2023.

[44] 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," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[45] Y. Lu, X. Yang, X. Li, X. E. Wang, and W. Y. Wang, "Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[46] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Human preference score: Better aligning text-to-image models with human preference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2096–2105.

[47] R. Abdelfattah, Q. Guo, X. Li, X. Wang, and S. Wang, "Cdul: Clip-driven unsupervised learning for multi-label image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1348–1357.

[48] Z. Teng, Y. Duan, Y. Liu, B. Zhang, and J. Fan, "Global to local: Clip-lstm-based object detection from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[49] A. Sain, A. K. Bhunia, P. N. Chowdhury, S. Koley, T. Xiang, and Y.-Z. Song, "Clip for all things zero-shot sketch-based image retrieval, fine-grained or not," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2765–2775.

[50] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.

[51] W. Pan, Z. Yang, D. Liu, C. Fang, Y. Zhang, and P. Dai, "Quality-aware clip for blind image quality assessment," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2023, pp. 396–408.

[52] T. Miyata, "Interpretable image quality assessment via clip with multiple antonym-prompt pairs," in *2023 IEEE 13th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*. IEEE, 2023, pp. 39–40.

[53] Z. Zhang, W. Sun, Y. Zhou, H. Wu, C. Li, X. Min, X. Liu, G. Zhai, and W. Lin, "Advancing zero-shot digital human quality assessment through text-prompted evaluation," *arXiv preprint arXiv:2307.02808*, 2023.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[55] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, and W.-Y. Ma, "Frank: a ranking method with fidelity loss," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 383–390.

[56] L. L. Thurstone, "A law of comparative judgment." *Psychological Review*, vol. 101, no. 2, p. 266, 1994.

[57] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

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

[59] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324.

[60] A. Borji, "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2," *arXiv preprint arXiv:2210.00586*, 2022.

[61] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[62] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "Towards language-free training for text-to-image generation,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 907–17 917.

[63] F. Bao, S. Nie, K. Xue, C. Li, S. Pu, Y. Wang, G. Yue, Y. Cao, H. Su, and J. Zhu, "One transformer fits all distributions in multi-modal diffusion at scale," *arXiv preprint arXiv:2303.06555*, 2023.

[64] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[65] J. Antkowiak, T. J. Baina, F. V. Baroncini, N. Chateau, F. FranceTelecom, A. C. F. Pessoa, F. S. Colonnese, I. L. Contin, J. Caviedes, and F. Philips, "Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000," *Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000*, 2000.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[67] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[68] A. Saha, S. Mishra, and A. C. Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5846–5855.

[69] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "Lafite: Towards language-free training for text-to-image generation. arxiv 2021," *arXiv preprint arXiv:2111.13792*, vol. 2, 2021.