IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 March 2023, accepted 13 April 2023, date of publication 17 April 2023, date of current version 24 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3267968

RESEARCH ARTICLE

Analysis of Appeal for Realistic AI-Generated Photos

STEVE GÖRING, RAKESH RAO RAMACHANDRA RAO, RASMUS MERTEN, AND ALEXANDER RAAKE, (Member, IEEE)
Audiovisual Technology Group, Technische Universität Ilmenau, 98693 Ilmenau, Germany

Corresponding author: Steve Göring (steve.goering@tu-ilmenau.de)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) through German Research Foundation under Grant DFG-437543412, and in part by the Open Access Publication Fund through Technische Universität Ilmenau.

ABSTRACT AI-generated images have gained in popularity in recent years due to improvements and developments in the field of artificial intelligence. This has led to several new AI generators, which may produce realistic, funny, and impressive images using a simple text prompt. DALL-E-2, Midjourney, and Craiyon are a few examples of the mentioned approaches. In general, it can be seen that the quality, realism, and appeal of the images vary depending on the used approach. Therefore, in this paper, we analyze to what extent such AI-generated images are realistic or of high appeal from a more photographic point of view and how users perceive them. To evaluate the appeal of several state-of-the-art AI generators, we develop a dataset consisting of 27 different text prompts, with some of them being based on the DrawBench prompts. Using these prompts we generated a total of 135 images with five different AI-Text-To-Image generators. These images in combination with real photos form the basis of our evaluation. The evaluation is based on an online subjective study and the results are compared with state-of-the-art image quality models and features. The results indicate that some of the included generators are able to produce realistic and highly appealing images. However, this depends on the approach and text prompt to a large extent. The dataset and evaluation of this paper are made publicly available for reproducibility, following an Open Science approach.

INDEX TERMS Image appeal, AI-generated images, image aesthetic.

I. INTRODUCTION

Recent developments in image processing have led to an increase in the popularity of AI-generated images. Such generators are, for example, DALLE 2,[1] Midjourney,[2] Stable Diffusion [35],[3] Glide [27], or Craiyon [6].[4] In general, they allow users to generate images automatically based on a given text prompt, which can be used to explore different word combinations and settings. As for example shown in [28], where small adjustments to given text prompts may have a large impact on the appeal generated of the image. Most of the generators provide a web-based interface or have their implementation open sourced and users can generate several images using a text prompt as input and select the best fitting image manually. Most of the generators share a similar common generic approach [40], such as Generative Adversarial Networks (GANs) in combination with natural language text processing and deep neural network-based upscaling methods.

Five examples of such generated images are shown in Figure 1. All images have been created with the same prompt "Purple flowers with yellow and a small bug". The first one (DALL-E-2) looks natural, however, the bug is not really visible. In turn, the second example (Glide) shows a yellow flower, in contrast to the required purple. The middle image (Craiyon) contains the flower and the bug in accordance with the text prompt, though, the image has a slightly artificial look, which may be due to the used upscaling method. The

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.

[1]https://openai.com/dall-e-2/
[2]https://www.midjourney.com/
[3]https://huggingface.co/spaces/stabilityai/stable-diffusion
[4]https://www.craiyon.com/

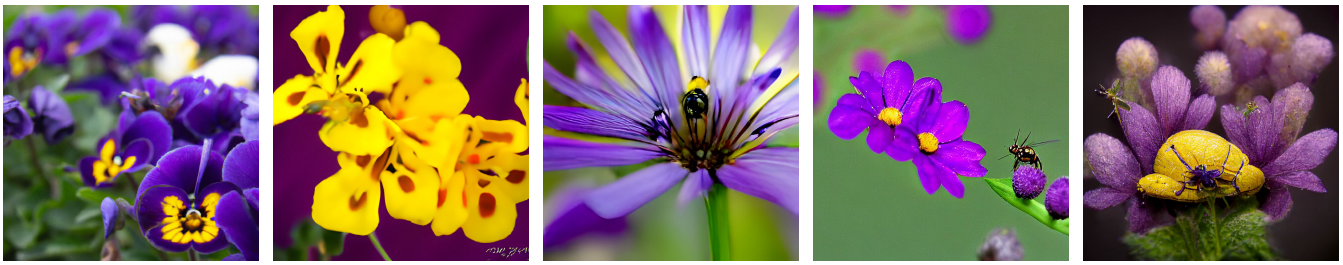

FIGURE 1. Examples of images generated with different AI approaches for the text prompt "Purple flowers with yellow and a small bug" (from left to right: DALL-E-2, Glide, Craiyon, Stable Diffusion and Midjourney).

next image (Stable Diffusion) looks noisy and artificial. The last image (Midjourney) gets the concepts well, however, the overall look of the generated image is more artistic similar to a painting. Here, it is visible, that such AI generators can produce images matching the provided text prompt, however, not all images may be realistic and of high visual quality or appeal. Clearly, the visual appeal depends on the used generator, hyperparameters, selection criterion, and the used text prompt.

Within this context, the question arises as to what extent these generators can be used to create photo-realistic and highly appealing images. To tackle this problem, we first needed suitable text prompts to create an appropriate dataset. Therefore, we started to explore different state-of-the-art collections for commonly used benchmark text prompts and use a sub-sampled set of prompts from the DrawBench [36] queries. Furthermore, we selected 11 real photos (created by the authors) in advance and created text prompts for them manually. Using the finally gathered 27 different text prompts, we generated images for five different text-to-image generators. These include DALL-E-2 (in the beta phase at the time of writing), Craiyon, Glide, Midjourney, and Stable Diffusion as text-to-image generators. For each of the generators, we created images for all the selected text prompts, which resulted in a total of $27 \times 5 = 135$ images, plus 11 real photos, therefore 146 are included in our dataset. As the objective is to evaluate the realism and the appeal of the images, we designed a subjective study using our self-developed tool AVrateVoyager [12]. Within the study, participants are asked to evaluate the image appeal, image realism, and how well the given image matches the provided text prompt. In addition to the subjective evaluation, we further evaluate various state-of-the-art no-reference metrics for image quality and appeal. Besides the no-reference metrics, different no-reference image quality or appeal raw features are extracted and evaluated in addition. The analysis also includes the training of machine learning models, namely random forest models and others, to check whether the mentioned metrics or the features can be used to derive prediction models for AI-generated image appeal. All results and images are publicly accessible,[5] as AVT-AI-Image-Dataset, following an Open Science approach, to enable reproducibility and further investigations. The results indicate that DALL-E-2 can produce highly appealing and realistic images, which also depends on the text prompts. On the other hand, Glide seems to produce the least realistic and worst appealing images. The subjective evaluation shows, that the participants are clearly able to distinguish between realistic and AI-generated images. The considered models and features were not capable of predicting image appeal thus highlighting the need for new and more sophisticated features and appeal prediction models. It is shown that the features can be used to classify and identify the used AI generators. However, these results are provided only as proof-of-concept and would need further investigation.

The article is organized as follows: In Section II a brief overview of state-of-the-art text-to-image generation and image appeal is provided. Afterward, in Section III, an overview of the dataset and the steps for its generation are outlined. This is followed by an analysis of the subjective and objective evaluation in Section IV. The paper finally concludes with a discussion and highlights open and future work.

II. TEXT-TO-IMAGE GENERATION AND APPEAL

In the following Section, we briefly describe some of the text-to-image generators and link them to image appeal.

The Craiyon generator [5], [6], previously known as DALL-E-mini, mainly focuses on providing a good text-to-image generator, while using limited processing power. In general, image generation in this case works as follows. First, the text prompt is encoded using the BART [20] encoder. Afterward, for each text token, a BART decoding step is performed, and then images are generated using the VQGAN encoder [8]. Several images are generated and CLIP [32] is used to select the best matching image. For the Craiyon model, a public web service is available, and also the implementation is open-sourced.

Similarly, the Glide image generation model is also open-source [27]. The evaluation of the Glide model includes photorealism and caption similarity and showed promising results. Glide follows a diffusion model approach, where several steps are performed to "de-noise" an input image to match a given text prompt. This approach is similar to the Stable Diffusion model [35].

[5]https://github.com/Telecommunication-Telemedia-Assessment/avt_ai_images

Furthermore, there are generators such as DALL-E-2 and Midjourney, where the internals are not published. However, in general, most other image generation models are, for example, based on GANs in combination with the CLIP model. For example, in [23] Liu et al. present a training-free method using an off-the-shelf GAN in combination with the CLIP model. The CLIP model is used to optimize the latent space of the GAN. The optimized GAN generates images, which can achieve a maximum semantic relevance score. In this case, the evaluation was done using the MS COCO Dataset [21]. The presented approach outperforms state-of-the-art text-to-image generators like DALL-E-2 or CogView considering their text prompt matching performance. The MS COCO Dataset [21] has been developed for scene understanding. It consists of images, object classifications, annotations, segmentations, and captions. With the release of image transformers [7], the application of such models has also increased for image-related problems, and some of the aforementioned models use a transformer architecture. Moreover, in [42] an autoregressive model is presented by Yu et al. It is based on transformer networks and on a decoder-encoder structure. By upscaling the model to 20B parameters, the network achieves very good performance on the MS COCO Database. The model outperforms state-of-the-art image-to-text generators, like DALL-E-2 and Imagen considering the Fréchet inception distance (FID), which is a metric used for the evaluation of generated images considering the matching to a reference image. Furthermore, Huang [16] analyzed and compared DALL-E-2 and two other text-to-image generators from former generations. Three metrics were used to evaluate the generated images: ''aesthetic'', ''comprehension and interpretation'', and ''creativity''. They found that DALL-E-2 can produce images with a high level of aesthetic quality, which outperforms the former generations. Other approaches may use lookups for databases. For example, Chen et al. [3] present a text-to-image generator model that uses a retrieval database to collect images for text prompts that are too specific or have not been covered by the training dataset considering their semantics. The model is able to retrieve (image, text) pairs, which are then used to gain knowledge about high-level semantics and low-level visual details. It is shown that this approach is able to outperform DALL-E-2 and Imagen in the MS COCO Benchmark considering FID scores. However, the MS COCO Benchmark [21] focuses more on labels, captions, and segmentation, and hence is not really suitable for image quality, appeal, and realism assessment.

Other approaches were developed to benchmark text-to-image generators. One such dataset is DrawBench [36]. The DrawBench text prompts have been used to evaluate the Imagen model [36] from Google. The model is not accessible, however, the text prompts are. In total 200 text prompts ranging from simple queries to contradictory ones are included. These text prompts are used for performance evaluation of text-to-image generators, where human annotators compare two models and are asked to decide which has generated the preferred image. However, in general, there is only a limited subjective evaluation of the image appeal of AI-generated images reported in the literature. For example, in [29], a description to create good prompts for Stable Diffusion is provided. The evaluation used crowdsourcing and was focused on aesthetically-pleasing images. This work is a good starting point, while it still focuses only on one image generator and keyword-based text prompts. To tackle this problem, in our dataset, we include several AI generators and use queries from the DrawBench dataset.

Some of the examples of studies related to AI-generated images and image appeal are the works conducted by Ling et al. [22] and Lei et al. [18]. Ling et al. [22] analyze image appeal of mobile games, thus computer generated images, with several dimensions, and show that, for example, the CPBD [26] contrast feature correlates best with the subjective appeal scores, with approximately 0.48 Pearson correlation. This work has been extended by Lei et al. in [18] to also include a no-reference deep learning-based prediction model, which predicts the four dimensions of the dataset, namely fineness, colorful, harmony, and overall appeal. The results indicate good prediction performance, however, only 10% of the overall dataset, approximately 100 images, are used for the evaluation. Furthermore, the model is not accessible, and thus cannot be included in the evaluation, and the context of this model was mobile game images. A general and accessible model to predict image appeal is the NIMA model [19]. It includes two prediction models, one for image quality and one for image appeal. The appeal model part has been trained with the AVA dataset [25], which only includes real photos.

III. OVERVIEW OF THE DATASET

As mentioned before, to generate an image, a prompt, which is a textual description of the image, is required. For this reason, 16 prompts have been selected out of a total of 200 from the DrawBench [36] list. The selection of these prompts was based on the criterion of covering a wide range of possible realistic images. In addition to the DrawBench prompts, 11 real photos are included in the dataset as well. For each of the real photos, a prompt has been manually defined, similar to an image caption. The overall idea is to evaluate how realistic the different generators are and whether subjects would recognize real photos within the dataset. In total, 27 different text prompts are considered in the dataset. The selection of prompts is summarized in Table 1. To fulfill the requirement that the generated images should be realistic or to guide the generator to produce more realistic photos, each of the prompts has been extended by '', real photo''. This is similar to the text prompt engineering shown in [28]. In case one generator provided more than one image, we manually selected the best-fitting and realistic-looking one.

For example, the prompts with the IDs p01, p15, and p16 originate from the DrawBench list, while p23 and p24 are captions of our own images (these images are referred to as ''own'' in the following) of the dataset. The prompt p15 is used as an anchor within the test because the scene is

TABLE 1. Text prompts used for the generation of the dataset.

ID	Prompt	Origin
p01	A red colored car	DrawBench
p02	A black colored dog	DrawBench
p03	A blue bird and a brown bear	DrawBench
p04	A pizza cooking an oven	DrawBench
p05	Two cars on the street	DrawBench
p06	Two cats and one dog sitting on the grass	DrawBench
p07	A triangular purple flower pot. A purple flower pot in the shape of a triangle	DrawBench
p08	A couple of glasses are sitting on a table	DrawBench
p09	A fisheye lens view of a turtle sitting in a forest	DrawBench
p10	A connection point by which firefighters can tap into a water supply	DrawBench
p11	A mechanical or electrical device for measuring time	DrawBench
p12	In late afternoon in January in New England, a man stands in the shadow of a maple tree	DrawBench
p13	A pear cut into seven pieces arranged in a ring	DrawBench
p14	A zebra to the right of a fire hydrant	DrawBench
p15	A photo of a confused grizzly bear in calculus class	DrawBench
p16	Hyper-realistic photo of an abandoned industrial site during a storm	DrawBench
p17	Pixel-art on windows	own
p18	Electronic Christmas trees	own
p19	Apollo 9 command module in San Diego	own
p20	Purple flowers with yellow and a small bug	own
p21	A seagull in the blue sky	own
p22	A train in the London subway	own
p23	A portrait of a mule	own
p24	A lonely bench in Iceland	own
p25	Group of people in front of a fountain illuminated from behind	own
p26	Black cows lying on the grass	own
p27	A box with tools for home office	own

FIGURE 2. The original real photo for p20, the AI-generated images are shown in Figure 1.

obviously not realistic, thus all participants should rate a matching generated image with this description as being not realistic. Similarly, the text prompt p04 has been selected, because it is contradicting, and therefore generated images cannot fulfill this description. Figure 1 shows the result for the text prompt p20 for all AI generators, while the real photo is shown in Figure 2 for comparison. As stated in the introduction, some of the generated images look realistic, however, some do not follow the given text prompt. In Table 2, an overview of the used AI-image generators is provided.

In total, five different text-to-image generators have been used to create the image dataset, namely, DALL-E-2 (in all plots and tables referred to as dall-e), Craiyon, Glide, Midjourney, and Stable Diffusion. We selected these generators because they are publicly available with provided web interfaces or open-source implementations and they represent a wide range of generators. In the case that the output resolution of the generated image was below 512×512 pixels for a specific generator, an additional up-scaling step has been performed. For up-scaling, the Real-ESRGAN [39], which is a state-of-the-art AI-based up-scaling algorithm, with the model "realesrgan-x4plus" was used. Furthermore, the TOPAZ Gigapixel up-scaler has been checked and as it showed similar visual results, the final up-scaling was implemented using Real-ESRGAN, enabling an automated procedure. It is important to mention, that also the used upscaling algorithm has an influence of the appeal for the generated images, however, also Midjourney and DALL-E-2 use deep neural network-based upscalers. Upscaling was required for Glide and Craiyon, and a manual check has been performed, whether the upscaling introduced more distortions considering the artificial look, which was not the case.

In addition to the listed image generators, Pixray[6] and others have been considered. However, for example, the results for several generation models of Pixray were considered too far away from being realistic by the authors in an informal inspection. Therefore this generator has been dropped from the list. Other generators have been dropped, too, such as NightCafe. It should be noted that especially for the paid or beta services only a limited amount of prompts could be evaluated, thus in some cases, several registrations were required.

Moreover, it is not necessarily possible to reproduce the generated images by the given prompts, due to the fact that some models change over time, or that e.g. a specific snapshot of the model is required (for example, in the case of the Google Colab notebook of the Stable Diffusion model, a seed of 1024 for the majority of images has been used, which cannot be ensured by other web-based generators, where no access to the model is available). In total 146 images are part of the dataset. Figure 3 gives an overview of all included images of the AVT-AI-Image-Dataset. Each row corresponds to a different AI generator, and the columns map to the text prompts p01 to p27, which are listed in Table 1. In general, it can be seen that most of the images follow a similar color scheme, and also the content seems to be similar to each other.

IV. EVALUATION OF QUALITY, APPEAL AND REALISM

In the following section, we describe the objective and subjective evaluation. The objective evaluation focuses on image quality and appeal, while the subjective evaluation covers image appeal and realism.

A. OBJECTIVE EVALUATION OF THE IMAGES

First of all, objective methods were used to analyze the given images and characterize the dataset considering the different AI generators. To evaluate the objective image quality and appeal of the generated images, several models could be used, such as, for example, Deimeq [9], the NIMA model [19] or NIQE [24]. The NIMA model offers two modes, one for image quality prediction and one for image appeal. Therefore,

[6]https://github.com/pixray/pixray

TABLE 2. AI image generators used in the evaluation.

Generator	Implementation	Comment	Code/URL/SRC
Craiyon	Web service	low resolution, requires up-scaling	[6], https://www.craiyon.com/
DALL-E-2	Web service (beta)	requires registration/paid service	https://openai.com/dall-e-2/
Glide	Python	low resolution, requires up-scaling	[27], https://github.com/openai/glide-text2im
Midjourney	Discord bot	requires registration/paid service	https://www.midjourney.com/
Stable Diffusion	Web service/Google Colab	registration to download model weights	[35], https://huggingface.co/spaces/stabilityai/stable-diffusion

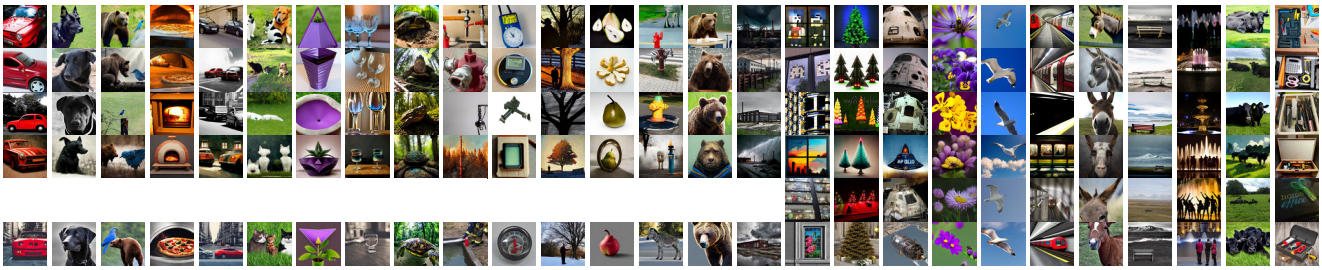

FIGURE 3. Overview of all images included in the dataset, rows: Craiyon, DALL-E-2, Glide, Midjourney, "own", Stable Diffusion; columns: text prompts p01 to p27.

NIMA was chosen for both objective evaluations. Deimeq is more a proof of concept model, which has been trained for higher resolution images and the generated images have a too low resolution. Furthermore, we included NIQE, because it is a natural scene-based statistic metric, and may be able to detect the artificial properties of the AI-generated images.

In Figure 4, boxplots for the estimated image quality and appeal values of NIMA and NIQE [24] are shown. First of all, it can be observed that there is no clear trend for the NIMA image appeal prediction to prefer any of the included generators. In turn, the "own" images (only 11) lead to a lower range of results than the majority of generators, this may be because the images do not cover all prompts and therefore do not span all possibilities of quality scores. Similarly, the NIMA quality model does not show a clear difference between the images. Here, it should be mentioned that all images have the same resolution and do not include any compression or other degradations, thus only content-related aspects could influence the prediction. The NIMA quality model is a deep neural network that has been trained using the TID2013 [31], which includes compression, noise, and other image degradations. The NIQE score indicates a wider range. Considering that NIQE includes measures of the statistics of the naturalness of the images, it can be concluded that Stable Diffusion seems to be the least natural generator in this regard. Even though it should be mentioned that the images of Midjourney seem more visually artistic when inspected visually, this is not reflected in the NIQE score.

Furthermore, other state-of-the-art no-reference image quality models have been included in the evaluation. For most of the models, the implementation "IQA-PyTorch" provided by Chen et al. [2] has been used.

In Figure 5, an overview of the results for the selected models included in "IQA-PyTorch" is given. The models MUSIQ [17], DBCNN [43], and MANIQA [41] are shown, and the results for further models are included in the data provided as open source with this paper. The overall trend of all models is similar, as it can be seen in Figure 5. As it can be seen from the plot, most of the images obtained from the different AI generators behave similarly. Only Stable Diffusion shows a higher value for the images for all three quality models, on the other side the MANIQ score of the Glide generator is the lowest. None of the mentioned image quality models is specifically trained or adjusted to AI-generated content, and furthermore does not include aspects such as image appeal or aesthetics. However, we evaluated these metrics to check whether there is a clear tendency for specific AI generators, and furthermore to evaluate the applicability of such quality metrics for this use case.

To this aim, we further evaluated the structure of the images from a photographic point of view. To evaluate whether the generated images follow common established photographic rules, the prediction model for the rule of thirds and image simplicity proposed by Göring et al. [10] is used. Other photo rules could be evaluated with a similar approach. The open source models provided by Göring et al. [10] handle the photo rule problem as a binary classification task for each of the two rules.[7] Here, a score of $= 1$ indicates that the rule is followed, and $= 0$ that the rule is not followed. The predictions for both rules are shown in Figure 6 for the AI-generated dataset. For the image simplicity rule, there is no trend visible in the evaluation, only that the values for DALL-E-2 are lower in comparison to the other generators. Furthermore, for the rule of thirds prediction, it can be seen that the majority of generators do not seem to follow this rule, in comparison to the "own" created images. Overall, the generators seem

[7]The implementation can be found here: https://github.com/Telecommunication-Telemedia-Assessment/sophoappeal_rule_prediction

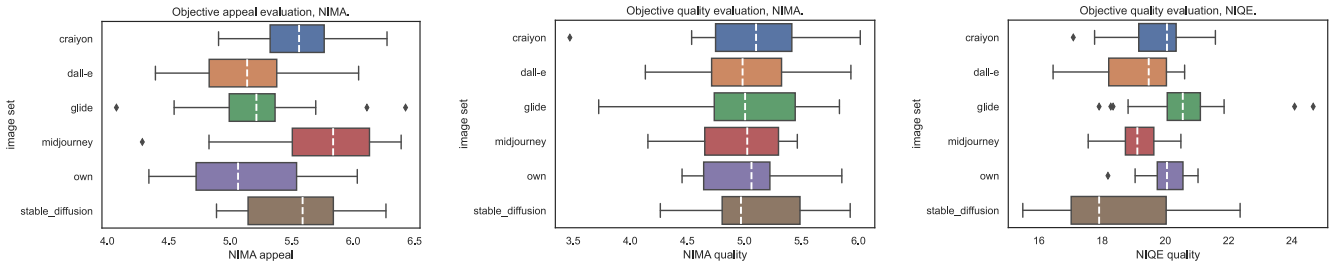

FIGURE 4. Objective image quality evaluation for the AI-image dataset (left to right: NIMA appeal, NIMA quality, and NIQE quality; NIMA higher scores are better; NIQE lower scores are better).

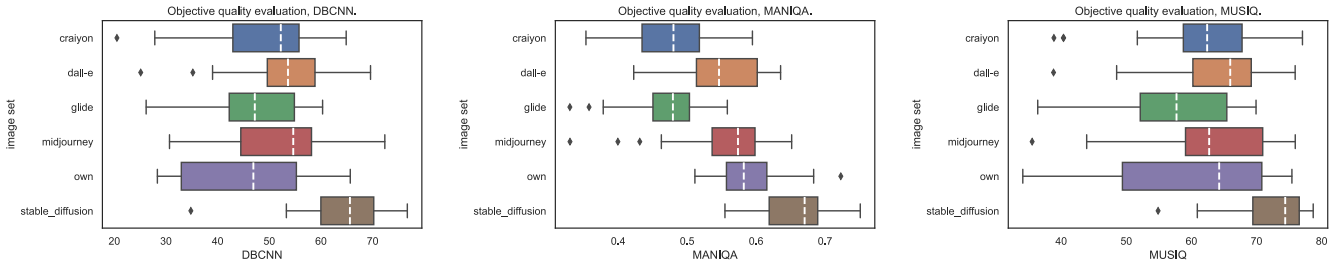

FIGURE 5. IQA-PyTorch selected quality models for the AI-image dataset (left to right: DBCNN, MANIQA, and MUSIQ quality; higher scores are better).

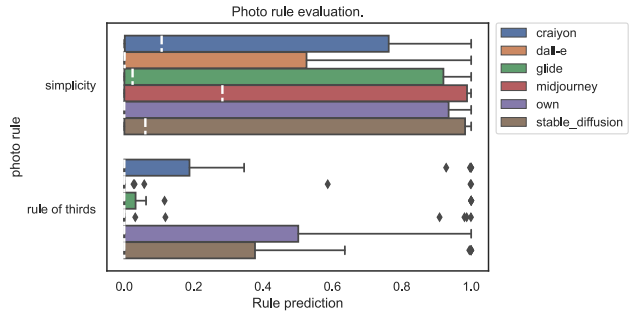

FIGURE 6. Image photo rule evaluation using [10].

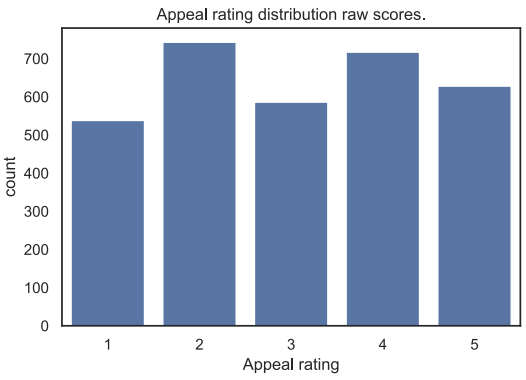

FIGURE 7. Rating distribution for image appeal.

to often produce simpler images, in contrast to images that follow the rule of thirds.

To sum up, the objective evaluation shows that there are no clear results that can be derived for the AI-generated images. For this reason, a more detailed subjective evaluation is required, which could be further used to improve existing models. It should also be noted that, in general, none of the mentioned metrics, models, and image analysis tools have been optimized or trained with AI-generated content, which also explains the results.

B. DESIGN OF THE ONLINE SUBJECTIVE TEST

To evaluate the appeal, realism, and matching of the given text prompt, a subjective test has been carried out. The test has been implemented using an online approach. For this reason, a slightly modified version of AVRate Voyager [12] has been used, which was used in other online studies before, see, e.g., [13] and [33]. The modifications involve adjusting the instructions for the particular purpose of this test, and furthermore an adaption of the rating scheme. In general, a participant was asked to rate the image appeal, image realism, and how well the shown image matched with the provided text prompt. In case a user did not rate all images, the ratings provided by the particular subject were excluded from further analysis. Based on this in total 146 images have been rated by each of the 22 participants.

The participants of the online tests were recruited within the context of the university and were unpaid. In the following, we will describe the evaluation of the appeal, realism, and matching of the text prompt in detail. For all ratings, the following procedure has been used. This approach ensures that only completed test runs are used for the evaluation.

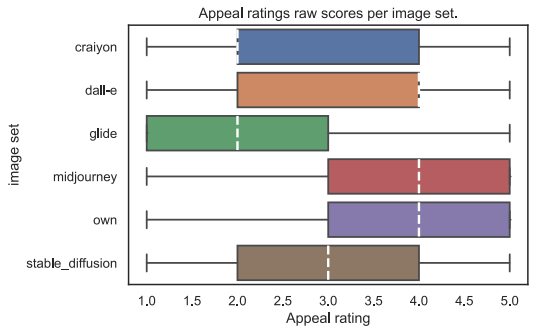

FIGURE 8. Image appeal rating per used AI generator; statistical testing (Kruskal-Wallis) showed that there are always significant differences between all paired combinations.

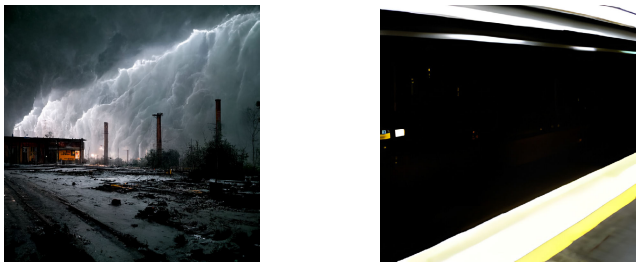

FIGURE 9. Best appealing image (left): Midjourney, p16, worst appealing image (right): Glide, p22.

C. IMAGE APPEAL

The first question of the subjective test covered the appeal of the shown image. An absolute category rating (ACR) scheme with five different values was used for rating the appeal, similar ACR scales have been also used in other appeal research, such as the AVA dataset [25]. The distribution of the ratings is summarized in Figure 7. In general, it is visible that there is a nearly uniform distribution of the ratings. Furthermore, in Figure 8, the distribution of raw scores for appeal for each individual AI generator is shown. Here, it can be seen that, for example, Midjourney seems to yield similar ratings as the "own" (real) images, while Stable Diffusion, DALL-E-2, and Craiyon have similar rating ranges, only the Glide Generator seems to create images that are not necessarily of high appeal. Here, the used upscaling algorithm and the already less realistic-looking images of the Glide generator are possibly influencing the overall appeal rating for Glide. To illustrate the overall range of the ratings, in Figure 9, the best and worst appealing images are shown. The worst appealing image has a mean appeal score of 1.32 (Glide p22) and the best has a score of 4.68 (Midjourney p16). The Glide generator seems to create an image that is not related to the text prompt, considering that for example there is not even a train visible in the generated image. On the other side, the Midjourney generator delivers a good-looking image of an abandoned industrial site during a storm (see p16). To visualize the diversity of the different generators, in Figure 10 for the prompt p22 all images are sorted by their corresponding mean appeal ratings. Interestingly, the "own" image has only a score of 3.59, while the image generated by Midjourney has a higher score of 4.5. Similarly, in Figure 11, all images produced by the different AI generators (there is no "own" image for p16) are shown and sorted according to their appeal scores. Overall, for the prompt p16 the AI generators create rather highly appealing images, which is reflected by the lowest score of 3.09 for Glide, and the best score of 4.68 for Midjourney.

TABLE 3. Correlation values of image appeal ratings compared to NIMA appeal prediction per image set.

Image set	Pearson	Spearman	Kendall
glide	0.58	0.39	0.31
dall-e	0.51	0.52	0.35
midjourney	0.47	0.44	0.31
own	0.38	0.25	0.20
craiyon	0.28	0.22	0.16
stable_diffusion	0.22	0.08	0.03

Furthermore, we performed a SOS-analysis [15] with the equation $SOS(x)^2 = a(-x^2 + 6 \cdot x - 5)$. The SOS-analysis is a method used to check the reliability of subjective tests proposed by Hoßfeld et al. [15]. The results of the SOS-analysis are shown in Figure 12. The estimated a is approximately 0.34, which is similar to reported values for the quality evaluation of cloud gaming in [15]. However, the a value for image appeal seems to be higher in general, as for example an a value of 0.27 is reported in [37] by Siahaan et al. for image appeal of real photos. Here, it can be argued that the subjective influence and thus inter-subject variation for the rating of image appeal is higher than, e.g., for image or video quality.

In the objective evaluation, there was only one prediction model for image appeal included, namely "NIMA appeal". In the following, we evaluate the performance of this model with the gathered image appeal ratings. We calculated the mean values of the appeal ratings for each image, and a scatterplot for comparison is shown in Figure 13.

The individual values for each image set are summarized in Table 3. All three calculated correlation coefficients (Pearson, Kendall, and Spearman) indicate only a small and weak correlation between the image appeal ratings and the objective model predictions. The main reason for this is that the prediction model itself is only trained on real-world content, and thus is not able to handle the artificial look and properties of the AI-generated images also including the AI-upscaling.

In addition, we compared the test results with the other objective image appeal/quality models. The results are listed in Table 4. Considering Person correlation, NIQE, BRISQUE, and MANIQA show the best absolute values. This trend is also visible in Kendall and Spearman correlations. However, none of the checked metrics shows a strong correlation to the subjective image appeal ratings. Considering their underlying principles based on image statistics, NIQE and BRISQUE may be good indicators for the artificial look of the images. The low performance of some of the models can be explained by the fact that these models

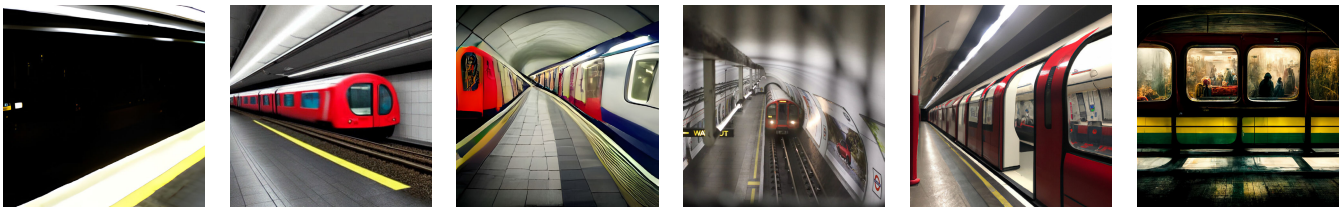

FIGURE 10. Image sorted by appeal values for p22 from worst (left) to best (right): Glide (1.32), Stable Diffusion (2.27), Craiyon (2.41), "own" (3.59), DALL-E-2 (3.64), and Midjourney (4.5).

FIGURE 11. Image sorted by appeal values for p16 from worst (left) to best (right): Glide (3.09), Craiyon (3.41), Stable Diffusion (4.09), DALL-E-2 (4.23), and Midjourney (4.68)– there is no image from "own" because p16 is a DrawBench prompt.

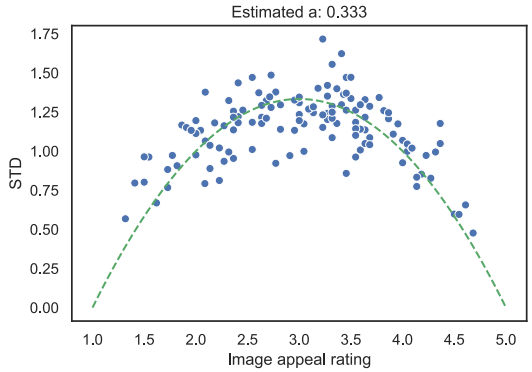

FIGURE 12. SOS analysis for image appeal ratings.

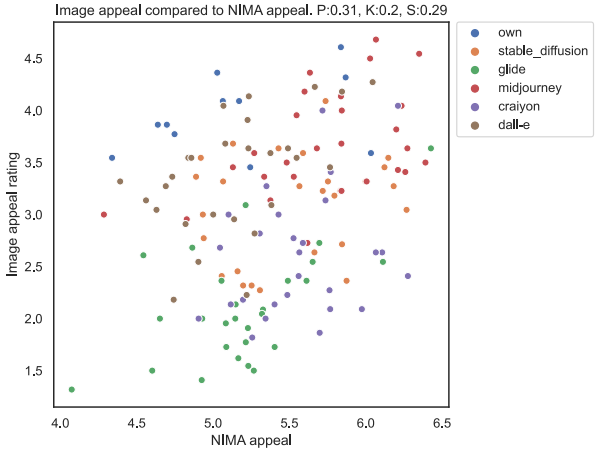

FIGURE 13. Image appeal ratings compared to NIMA appeal prediction.

are specifically trained to handle quality-related degradations, such as JPEG compression, which are not present

TABLE 4. Correlation values of image appeal ratings to other state-of-the-art image quality models.

Model	Pearson	Kendall	Spearman
NIQE	-0.43	-0.27	-0.40
BRISQUE	-0.38	-0.23	-0.33
ILNIQE	-0.35	-0.19	-0.28
NIMA_quality	-0.05	-0.05	-0.07
DBCNN	0.08	0.04	0.07
MUSIQ	0.15	0.09	0.15
NIMA_appeal	0.31	0.20	0.29
MANIQA	0.34	0.24	0.37
combined	0.47	0.29	0.44

in the generated images. To check, whether the individual models combined could reach a good prediction performance, we trained a random forest regressor (100 trees; from Scikit-learn [30] with default parameters) in a 10-fold-cross validation setting. The results are listed in Table 4 and marked as ***combined***. Here, the comparison of ***combined*** with the other models would not be fair, because the 10-fold-cross validation for ***combined*** may show better results than for the non-retrained models. However, the evaluation gives an idea of the extent to which the individual models can be used. Here, the overall performance does not indicate that the listed models are sufficient enough for a prediction model. Furthermore, it should also be noted that for a real prediction model, the used dataset should be larger. In addition to this, we checked whether it is possible to train a random forest classifier to estimate the used AI generator, following a similar 10-fold-cross validation approach.

The confusion matrix for the validation is shown in Figure 14. Stable Diffusion and Glide are best predicted with the trained model, even though there are also some cases of misclassification. The AI generator Craiyon seems to produce

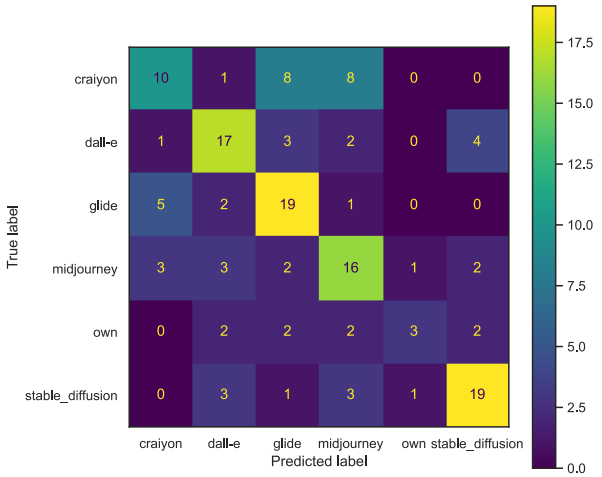

FIGURE 14. Confusion matrix for AI generator prediction.

TABLE 5. Image appeal and quality-related features with their corresponding sources.

Feature	Source
colorfulness	[14]
tone	[1]
saturation	[1]
contrast	[11] own
spatial information (SI)	[34]
CPBD	[26]
blur strength	[4]

TABLE 6. Correlation values of image appeal ratings to other state-of-the-art image appeal features.

Feature	Pearson	Kendall	Spearman
blur stength	-0.39	-0.26	-0.38
saturation	-0.03	-0.02	-0.03
colorfulness	0.02	-0.04	-0.06
si	0.04	0.00	0.01
tone	0.08	-0.00	-0.00
CPBD	0.14	0.12	0.18
contrast	0.25	0.15	0.22
combined features	0.28	0.17	0.27
features + models	0.49	0.31	0.45

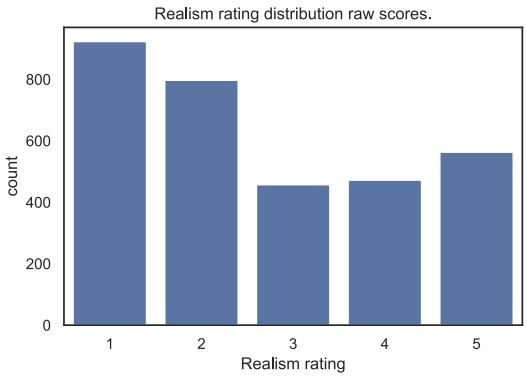

FIGURE 15. Rating distribution for image realism.

similar images such as Glide and Midjourney (according to the used models as features). The worst correctly classified parts are the ''own'' images, first of all, only 11 images are included, and some seem to be misclassified with DALL-E-2, Glide, or Midjourney.

In addition to the pure image quality or appeal models, we calculated several state-of-the-art features,[8] which are used in the context of image appeal. In total, we calculated 7 features, namely colorfulness, tone, saturation, contrast, spatial information, CPBD,[9] and blur strength.[10] All features with their corresponding references are listed in Table 5. The CPBD feature has been reported to show good performance for image appeal prediction in the case of mobile game images, according to Ling et al. [22].

Furthermore, we checked whether the extracted features are correlated to the image appeal ratings. In Table 6 the correlation values of the mentioned features are summarized. Similar to the objective models, no high correlation is visible in this case too. Blur strength and contrast have the strongest absolute correlation in terms of the Pearson correlation coefficient. Therefore, when an image has a higher contrast the appeal is higher, and in case the image is blurry then it has a lower appeal. Similar to the objective models, we trained a random forest model (same settings) to verify whether these features can be used to predict the image appeal of the AI-generated images. The results are listed as ***combined features*** in Table 5, and it is visible that the combination of the features only results in a weak correlation to the image appeal ratings. Also here, it is visible that the created images may include different aspects, which reduce the appeal, which is usually not included in the state-of-the-art features. In addition to this, we combined the features and the models to predict the image appeal shown as ***features + models*** using also a random forest model. The overall performance got improved due to the addition of the features, compare also Table 4, however overall the performance is still not high considering the estimated correlation values.

D. REALISM

Next to the pure appeal rating, the subjective evaluation also included for each participant to rate the realism of the given photo. We used a 1-5 ACR rating (1=not realistic, 5=realistic) scheme to evaluate the realism of the images.

The overview of the raw score distribution of the realism rating is shown in Figure 15. A general shift towards the lower end of the rating range is visible, thus there is a clear tendency toward the AI-images being perceived as not very realistic.

[8]Code: https://github.com/Telecommunication-Telemedia-Assessment/avt_ai_images

[9]Code: https://github.com/0x64746b/python-cpbd

[10]Code: scikit-image [38].

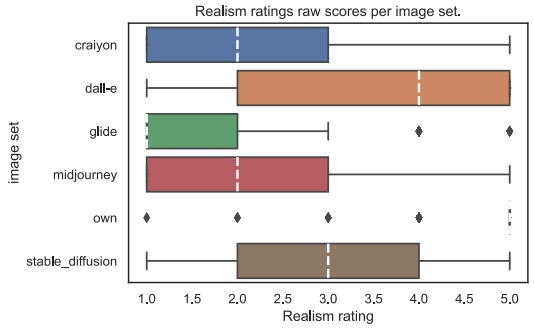

FIGURE 16. Image realism rating per used AI generator; statistical testing (Kruskal-Wallis) showed that there are always significant differences between all paired combinations.

Furthermore, we evaluated the realism rating considering the different AI generators, which is shown in Figure 16. It can be clearly stated, that the ''own'' (real) images are recognized as being realistic by the majority of the participants. The most realistic AI generator seems to be DALL-E-2, followed by Stable Diffusion. Midjourney creates images that are more similar to paintings and has therefore a lower rating. Craiyon has a similar rating range as compared to Midjourney while Glide generates the least realistic images.

The most and least realistic rated images are shown in Figure 17. The least realistic image, which was generated for p17 with the Glide AI generator, has a mean realism score of 1.14. The most realistic image is one that is a real photo from the ''own'' subset and it has a score of 4.86 for the prompt p21. The Glide generator fails for the text prompt p17, which is about pixel art on windows. There are no real windows visible and also no pixel art. To visualize the range of realism ratings the Figure 18 shows all images sorted by realism ratings for p17. The image from ''own'' is the most realistic for p17 and Glide is the least, the overall range for the realism ratings is from 1.14 to 4.77. The second best image according to realism rating is DALL-E-2 with a score of 4.55. Furthermore, in Figure 19 the images for the prompt p21 are shown sorted by their realism ratings. Also in this case, the ''own'' has the highest score (4.86) and is followed by DALL-E-2 with a score of 3.59. For both prompts there is a larger gap in the ratings from DALL-E-2 to the next image, which indicates that some of the other generators do produce lesser realistic photos.

Similar to the appeal ratings, we evaluated the correlations to the objective metrics considering the realism ratings of the images. The results are summarized in Table 7. It is visible that none of the included image quality metrics can be used to predict image realism. From the absolute values, only BRISQUE and MANIQA have a medium or low correlation considering Pearson Correlation Coefficient. To further evaluate whether the calculated image quality values can be used to predict image realism, we trained a random forest model similar to the one done for the image appeal case. We used the default values of scikit-learn [30], and performed a 10-fold-cross validation. The results of the trained model

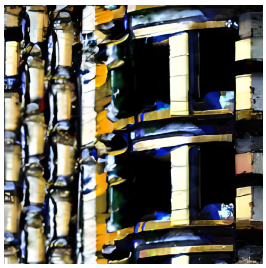

FIGURE 17. Most realistic image (left): ''own'', p21, least realistic image (right): Glide, p17.

TABLE 7. Correlation values of image realism ratings to other state-of-the-art image quality models.

Model	Pearson	Kendall	Spearman
BRISQUE	-0.31	-0.24	-0.36
ILNIQE	-0.25	-0.16	-0.23
NIMA_appeal	-0.20	-0.12	-0.17
NIQE	-0.17	-0.10	-0.16
DBCNN	-0.06	-0.01	-0.02
MUSIQ	0.02	0.05	0.08
NIMA_quality	0.03	0.01	0.01
MANIQA	0.23	0.16	0.24
combined	0.29	0.2	0.28

TABLE 8. Comparison of image appeal and image realism ratings for each individual generator/subset of the overall dataset; values are sorted by Pearson Correlation Coefficient and rounded to 2 decimal places; * indicates mean values.

Image set	Pearson	Kendall	Spearman	Appeal*	Realism*
midjourney	0.29	0.21	0.27	3.67	2.34
dall-e	0.33	0.27	0.32	3.38	3.41
''own''	0.49	0.37	0.56	3.96	4.63
stable_diffusion	0.69	0.53	0.69	3.09	2.78
glide	0.75	0.46	0.60	2.12	1.84
craiyon	0.87	0.62	0.80	2.61	2.20

are highlighted as ***combined*** in Table 7. It can be stated that there is only a low/medium correlation between the prediction with the realism ratings, therefore the used image quality models are not sufficient enough for the case of predicting realism. Similar to the appeal case, here also the more synthetic nature of the images is the main influence because none of the metrics has been trained or evaluated before with AI-generated images. To check whether there is a direct link between image appeal and image realism, we calculated various correlation coefficients for image appeal and image realism.

In Table 8 the comparison of realism and appeal ratings is summarized considering the different AI generators. It can be seen that generators that produce on average lower appealing images have a more direct linear functional connection between realism and image appeal. DALL-E-2 and Midjourney have a lower connection between realism and appeal, thus some images may look realistic and have a lower appeal than others. For the ''own'' (real) images also

FIGURE 18. Image sorted by realism values for p17 from worst (left) to best (right): Glide (1.14), Stable Diffusion (1.77), Craiyon (2.18), Midjourney (2.55), DALL-E-2 (4.55), and "own" (4.77).

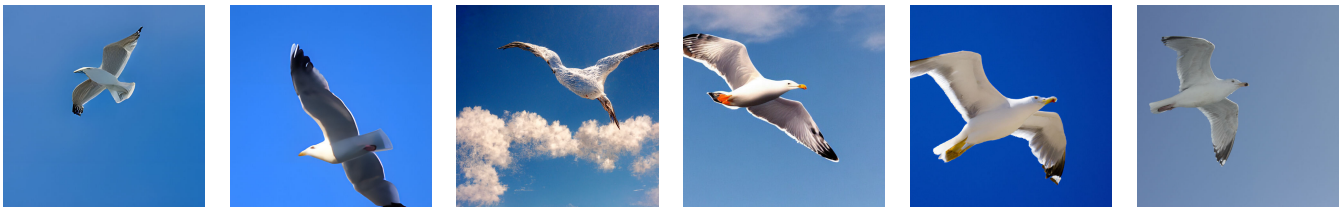

FIGURE 19. Image sorted by realism values for p21 from worst (left) to best (right): Craiyon (1.73), Glide (1.73), Midjourney (2.5), Stable Diffusion (2.82), DALL-E-2 (3.59), and "own" (4.86).

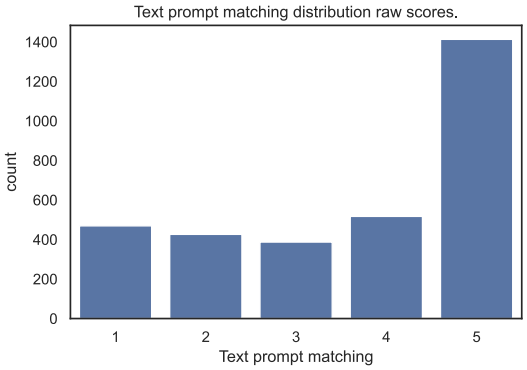

FIGURE 20. Rating distribution for text prompt matching.

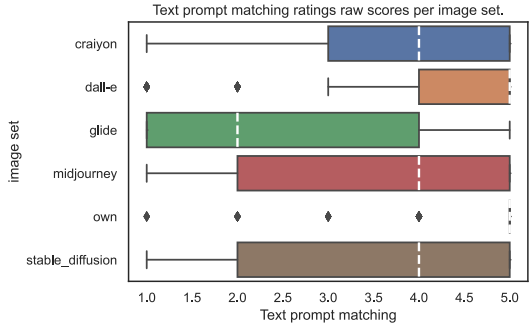

FIGURE 21. Text prompt matching per used AI generator; statistical testing (Kruskal-Wallis) showed that there are always significant differences between all paired combinations.

a medium correlation can be observed, here it should be considered that the overall mean appeal and realism ratings are the highest considering the subsets of the dataset. Furthermore, we checked for significant differences of the appeal and realism values. In the case of appeal, statistical testing

FIGURE 22. Best text prompt matching image (left): DALL-E-2, p02, least matching image (right): Glide, p10.

(Kruskal-Wallis) showed no statistical differences between (dall-e, midjourney), (dall-e, stable_diffusion), and (midjourney, "own"), all other pairs were statistically different considering appeal. We performed the same pairwise statistical testing for the realism ratings, here there was no statistical difference for (craiyon, midjourney), and (midjourney, stable_diffusion), and all other comparisons showed statistical differences.

E. TEXT PROMPT EVALUATION

In addition to image appeal and image realism, we asked the participants to rate how well the shown image matches the provided text prompt. We asked "Is the following text a good description of the image?", and the participants needed to rate using a 1-5 ACR rating scheme how good the text describes the images.

In Figure 20 the raw score distribution for the text prompt matching ratings is shown. It can be seen that the majority of images are matching their text prompts. However, some images seem to not match.

Therefore, we further estimated the matching considering the different AI generators. The results are shown in Figure 21. It is visible that the "own" (real) images are

FIGURE 23. Image sorted by text prompt matching values for p10 from worst (left) to best (right): Glide (1.09), Craiyon (2.18), Midjourney (2.23), Stable Diffusion (3.05), and DALL-E-2 (4.86).

FIGURE 24. Image sorted by text prompt matching values for p02 from worst (left) to best (right): Midjourney (4.32), Stable Diffusion (4.77), Craiyon (4.82), Glide (4.91), and DALL-E-2 (5.0).

mostly matching the shown text. Furthermore, DALL-E-2 seems to produce the best matching images, followed by Craiyon. Midjourney and Stable Diffusion seem to have a similarly good performance considering the prompt matching, while Glide is the worst of all considered AI generators. In Figure 22 examples for the best and least matching images to the corresponding text prompts are shown. Here, Glide with the text prompt p10 has the lowest matching value of 1.09, while DALL-E-2 with p02 has a score of 5.0. The text prompt p10 refers to a fire hydrant, which is hard to understand for the Glide generator due to the description of the text prompt, on the other hand, the text prompt p02 is simple because it just describes a black colored dog, which is very well represented by the generated image with DALL-E-2. To show the range of possible generated images, in Figure 23 for p10 and in Figure 24 for p02 are shown according to their text prompt matching scores. In both cases, DALL-E-2 has the highest score, as mentioned before p02 is a simple text prompt, therefore the text matching values are in the higher range (above 4.32), while for p10 nearly the full 1-5 range is covered, with DALL-E-2 being the only generated image above 4.

V. DISCUSSION

Starting with the observation that there are only a few studies considering the image appeal of AI-generated images available, we created a dataset covering popular AI-Text-To-Image generators. Using objective state-of-the-art image quality models and features, we analyze the differences between the AI generators. None of the models or features can be directly out-of-the-box used to estimate which AI generator is used or to detect whether an image is generated or not. To further evaluate the image appeal of the generated dataset, we design an online subjective test, which includes ratings for image appeal, image realism, and text prompt matching. The results indicate that there is only a low correlation with state-of-the-art objective models and features regarding image appeal and realism. Moreover, to check whether the used models and features are suitable for AI-generated images, we trained random forest regression and classification models. Here, it can be stated that image appeal and realism prediction is challenging for the models, considering that none of the mentioned features or state-of-the-art image appeal/quality models have been developed with the purpose of AI-generated images. Thus there is a need for specific models and features which include specific distortions of AI-generated content. Furthermore, it is possible to use the features and models combined to predict which AI generator has been used, however, the shown results may be limited due to the size of the dataset. In general, Midjourney and ''own'' (real photos) have the highest appeal ratings. And DALL-E-2 and ''own'' are the most realistic images. In addition, a direct link between image appeal and realism can be observed, here, for the Craiyon and Glide an appealing image must be also a realistic image, which is related to the fact that these two generators create a wider range of images considering their realism and appeal. Interestingly, the majority of participants were able to identify which images are real photos and which have been generated, which is visible in the realism evaluation for the ''own'' subset. The text prompt evaluation also shows, that the ''own'' images are well described and that DALL-E-2 produced the best matching images. Here, the language processing part of the AI generator is important, also because some of the used text prompts are per definition designed to challenge text processing engines.

VI. CONCLUSION

AI-based image generation gains popularity. To evaluate how good such generators are considering image appeal and

realism, we created a dataset with five different image generators and performed an evaluation. The evaluation was two-fold. In the first part, we evaluated the generated images using state-of-the-art image quality models or features and figured out that none of the existing models can be directly used for AI-generated images. Furthermore, in the second part, we designed a subjective evaluation test, considering the rating of image appeal, realism, and how well the image matches a given text prompt. The results indicate, that some generators are better for image appeal, and some for realism. However, in general, the "own" created images have been better rated and are clearly identified by the participants. Therefore it can be stated that the image generators still have limitations considering how realistic the images are. The text prompt evaluation showed that the majority of images are matching with the corresponding text prompt, while there are generators that are better able to match the text prompt. Considering that such AI-generators are improved continuously, the matching of text prompts will increase, however we also showed with the dataset that image appeal and realism should also be considered in the development. The dataset and the subjective ratings for image appeal, realism, and text prompt matching are publicly available and can be used for further evaluation. For example, the dataset could be used for the development of image appeal and quality prediction models, which also include the specific distortion aspects of AI-generated images. In future work, the dataset should be extended by more generators, more images, and more objective state-of-the-art features. Therefore, this dataset can be seen as a starting point for future projects. The gathered subjective annotations considering image appeal and realism, which are important aspects of AI-generated images, can be used to develop image classifiers or prediction models.

ACKNOWLEDGMENT

The authors would like to thank the participants of the conducted subjective test, and also would like to thank the AG Wissenschaftliches Rechnen, Technische Universität Ilmenau, for providing computing resources.

REFERENCES

[1] T. O. Aydın, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 1, pp. 31–42, Jan. 2015.

[2] C. Chen and J. Mo. (2022). *IQA-PyTorch: Pytorch Toolbox for Image Quality Assessment*. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch

[3] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-Imagen: Retrieval-augmented text-to-image generator," 2022, *arXiv:2209.14491*.

[4] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric," in *Proc. SPIE*, Feb. 2007, pp. 196–206.

[5] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Le, Luke, and R. Ghosh. (2022). *Dall-E Mini Explained*. [Online]. Available: https://wandb.ai/dalle-mini/dalle-mini/reports/DALL-E-Mini-Explained–Vmlldzo4NjIxODA

[6] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Lê Khac, L. Melas, and R. Ghosh. (Jul. 2021). *DallB7e Mini*. [Online]. Available: https://github.com/borisdayma/dalle-mini

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[8] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12873–12883.

[9] S. Goring and A. Raake, "Deimeq—A deep neural network based hybrid no-reference image quality model," in *Proc. 7th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Nov. 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8611703

[10] S. Göring and A. Raake, "Rule of thirds and simplicity for image aesthetics using deep neural networks," in *Proc. IEEE 23rd Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2021, pp. 1–6.

[11] S. Goring, R. R. R. Rao, B. Feiten, and A. Raake, "Modular framework and instances of pixel-based video quality models for UHD-1/4K," *IEEE Access*, vol. 9, pp. 31842–31864, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9355144

[12] S. Goring, R. R. R. Rao, S. Fremerey, and A. Raake, "AVrate Voyager: An open source online testing platform," in *Proc. IEEE 23rd Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2021, pp. 1–6.

[13] S. Göring, R. R. R. Rao, and A. Raake, "Quality assessment of higher resolution images and videos with remote testing," *Qual. User Exp.*, vol. 8, no. 1, Dec. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s41233-023-00055-6

[14] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Proc. SPIE*, vol. 5007, pp. 87–95, Jun. 2003.

[15] T. Hossfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Proc. 3rd Int. Workshop Quality Multimedia Exp.*, Sep. 2011, pp. 131–136.

[16] Z. Huang. (2022). *Analysis of Text-to-Image AI Generators*. Accessed: Apr. 19, 2023. [Online]. Available: https://digital.kenyon.edu/cgi/viewcontent.cgi?article=1033&context=dh_iphs_ai

[17] 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. (ICCV)*, Oct. 2021, pp. 5148–5157.

[18] Z. Lei, Y. Xie, S. Ling, A. Pastor, J. Wang, J. Dong, and P. Le Callet, "Multi-modal aesthetic assessment for mobile gaming image," in *Proc. IEEE 23rd Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2021, pp. 1–5.

[19] C. Lennan, H. Nguyen, and D. Tran. (2018). *Image Quality Assessment*. [Online]. Available: https://github.com/idealo/image-quality-assessment

[20] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[22] S. Ling, J. Wang, W. Huang, Y. Guo, L. Zhang, Y. Jing, and P. Le Callet, "A subjective study of multi-dimensional aesthetic assessment for mobile game image," in *Proc. 1st Workshop Quality Exper. (QoE) Vis. Multimedia Appl.*, Oct. 2020, pp. 47–53.

[23] X. Liu, C. Gong, L. Wu, S. Zhang, H. Su, and Q. Liu, "FuseDream: Training-free text-to-image generation with improved CLIP+GAN space optimization," 2021, *arXiv:2112.01573*.

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

[25] 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.

[26] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Sep. 2011.

[27] 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," 2021, *arXiv:2112.10741*.

[28] J. Oppenlaender, "A taxonomy of prompt modifiers for text-to-image generation," 2022, *arXiv:2204.13988*.

[29] N. Pavlichenko, F. Zhdanov, and D. Ustalov, "Best prompts for text-to-image models and how to find them," 2022, *arXiv:2209.11711*.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.
[31] N. Ponomarenko, O. Ieremeiev, V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. -C. J. Kuo, "A new color image database TID2013: Innovations and results," in *Proc. Adv. Concepts Intell. Vis. Syst., 15th Int. Conf. (ACIVS)*, Poznań, Poland: Springer, Oct. 2013, pp. 402–413.
[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[33] R. R. R. Rao, S. Goring, and A. Raake, "Towards high resolution video quality assessment in the crowd," in *Proc. 13th Int. Conf. Quality Multimedia Exp. (QoMEX)*, Jun. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9465425
[34] *Subjective Video Quality Assessment Methods for Multimedia Applications*, I. Recommendation document ITU-T P.910, 2022.
[35] 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. 2021, pp. 10684–10695.
[36] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022, *arXiv:2205.11487*.
[37] E. Siahaan, J. A. Redi, and A. Hanjalic, "Beauty is in the scale of the beholder: Comparison of methodologies for the subjective assessment of image aesthetic appeal," in *Proc. 6th Int. Workshop Quality Multimedia Exp. (QoMEX)*, Sep. 2014, pp. 245–250.
[38] S. Van Der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014, doi: 10.7717/peerj.453.
[39] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.
[40] M. Wolfe. (2022). *The Emerging World of AI Generated Images*. [Online]. Available: https://towardsdatascience.com/the-emerging-world-of-ai-generated-images-48228c697ee9
[41] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1191–1200.
[42] J. Yu, Y. Xu, J. Yu Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. Karagol Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling autoregressive models for content-rich text-to-image generation," 2022, *arXiv:2206.10789*.
[43] 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. 2020.

STEVE GÖRING received the B.Sc. and M.Sc. degrees in computer science from TU Ilmenau, in 2008 and 2013, respectively, and the Ph.D. degree with the focus of visual quality prediction using machine learning, in 2022. Before he started, in 2016, he was with the Audiovisual Technology Group and he was with the Big Data Analytics Group, Bauhaus-University Weimar. He is currently a Computer Scientist with the Audiovisual Technology Group, TU Ilmenau. His current research interests include data analysis problems for video quality models and video streams. His specializations are data analytics/machine learning, video quality, and distributed communication/information systems.

RAKESH RAO RAMACHANDRA RAO received the M.Sc. degree in communications engineering from RWTH Aachen, in 2017, with focus a on image content analysis and millimeter wave transmission systems. He has been an Electrical Engineer with Audiovisual Technology (AVT), TU Ilmenau, since 2017. His research interests include video quality analysis and modeling. He has been actively involved in standardization activities on QoE assessment methods in ITU-T Study Group 12. Before joining AVT, he was an intern with HEAD acoustics, where he worked on reference-based noise estimation. His specializations include video quality and image content analysis.

RASMUS MERTEN received the B.Sc. degree in mediatechnology from TU Ilmenau in cooperation with Fraunhofer IDMT, in January 2022, where he is currently pursuing the master's degree in mediatechnology. He has joined the Audiovisual Technology Group, in March 2022. He also focuses on the field of machine learning.

ALEXANDER RAAKE (Member, IEEE) received the Ph.D. degree (Dr.-Ing.) from the Electrical Engineering and Information Technology Faculty, Ruhr-Universitat Bochum, in 2005, with the book Speech Quality of VoIP. He has joined TU Ilmenau, in 2015, as a Full Professor, where he heads the Audiovisual Technology Group. From 2005 to 2015, he was a Senior Researcher, an Assistant Professor, and an Associate Professor with the An-Institut T-Laboratories, TU Berlin, a joint venture between Deutsche Telekom AG and TU Berlin, heading the Assessment of IP-Based Applications Group. From 2004 to 2005, he was a Postdoctoral Researcher with LIMSI-CNRS, Orsay, France. His research interests include audiovisual and multimedia technology, speech, audio, video signals, human audiovisual perception, and quality of experience. Since 1999, he has been involved in the ITU-T Study Group 12's standardization work on QoS and QoE assessment methods. He is a member of the Acoustical Society of America, AES, VDE/ITG, and DEGA.