IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 June 2023, accepted 3 July 2023, date of publication 5 July 2023, date of current version 13 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3292588

RESEARCH ARTICLE

Image Appeal Revisited: Analysis, New Dataset, and Prediction Models

STEVE GÖRING 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.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics committee of Technische Universität Ilmenau.

ABSTRACT There are more and more photographic images uploaded to social media platforms such as Instagram, Flickr, or Facebook on a daily basis. At the same time, attention and consumption for such images is high, with image views and liking as one of the success factors for users and driving forces for social media algorithms. Here, "liking" can be assumed to be driven by image appeal and further factors such as who is posting the images and what they may show and reveal about the posting person. It is therefore of high research interest to evaluate the appeal of such images in the context of social media platforms. Such an appeal evaluation may help to improve image quality or could be used as an additional filter criterion to select good images. To analyze image appeal, various datasets have been established over the past years. However, not all datasets contain high-resolution images, are up to date, or include additional data, such as meta-data or social-media-type data such as likes and views. We created our own dataset "AVT-ImageAppeal-Dataset", which includes images from different photo-sharing platforms. The dataset also includes a subset of other state-of-the-art datasets and is extended by social-media-type data, meta-data, and additional images. In this paper, we describe the dataset and a series of laboratory- and crowd-tests we conducted to evaluate image appeal. These tests indicate that there is only a small influence when likes and views are included in the presentation of the images in comparison to when these are not shown, and also the appeal ratings are only a little correlated to likes and views. Furthermore, it is shown that lab and crowd tests are highly similar considering the collected appeal ratings. In addition to the dataset, we also describe various machine learning models for the prediction of image appeal, using only the photo itself as input. The models have a similar or slightly better performance than state-of-the-art models. The evaluation indicates that there is still an improvement in image appeal prediction and furthermore, other aspects, such as the presentation context could be evaluated.

INDEX TERMS Image appeal, image aesthetic, image popularity, machine learning, image dataset.

I. INTRODUCTION

More and more images are uploaded on a daily basis to social media or are shared across the world. Flickr, Instagram, Facebook, and Whatsapp are just a few selected platforms to share, upload and consume images. The amount of uploaded images makes it harder for a user to decide whether they like an image or not, and even more difficult, whether an image is of high appeal or not, especially considering that thousands of images are uploaded to such photo-sharing platforms.[1] Moreover, typical photo-sharing platforms use internal methods to arrange and score images according to their popularity or visual appearance. Here, it can be assumed that image viewing and liking are based on different

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

[1] see https://www.flickr.com/photos/franckmichel/6855169886/

factors, for example, whether a picture is visible, i.e., shown sufficiently often to users, who has been posting the image, and whether the viewer is following that person, how many others have already liked or viewed the image, and by the properties of the image in regarding the aesthetic appeal, including, e.g., the technical/photographic knowledge [1], [2], [3], [4]. In this paper, the focus is on aesthetic appeal, that is, how appeal is being judged by human viewers, and how it can be predicted using algorithms for automatic appeal estimation.

Several methods to predict image appeal or related aspects have been developed [4], [5], [6], [7], [8], [9]. Within this context, it is also crucial to have valid and large amounts of data to train and validate such models. Here, for example, the AVA [10] or the AADB [11] datasets with a focus on image appeal have been published and are used as the gold standard in the field. Moreover, other datasets such as KonIQ-10k [12], [13], or LIVE In the Wild Image Quality Challenge Database [14] are available and related to image appeal, even though they rather focus on quality-related aspects. Prediction models, such as the NIMA model [5], [15][2] are publicly available and can be used to predict image appeal with promising results. The NIMA model has two modes, one for image appeal, which is in the following referred to as *nima_a* and one for image quality prediction, referred as $nima_q$. However, there are several open points in this regard. For example, the AVA dataset covers only lower-resolution images, and additional influence factors such as likes are not included in the evaluation of the dataset. The results of the NIMA appeal *nima_a* model indicate that there is a medium-high linear correlation of the model with subjective annotations, thus the overall image-appeal-prediction problem seems to be challenging.

In the following, we describe our own dataset, namely **AVT-ImageAppeal-Dataset**, which uses parts of already existing datasets and is extended by images from other sources. Included in the dataset are extracted data considering image segmentation, image depth estimation, and other state-of-the-art features, which can later be used for the appeal analysis. The dataset forms the basis for several subjective viewing tests conducted as part of this research. For example, we describe lab and crowd-tests considering the images alone, to verify the usage of crowd-tests. Furthermore, we extended the tests by including like and view statistics for a subset of the data, to evaluate the influence on the overall appeal rating. The results indicate that lab and crowd tests are highly similar to each other, indicating that such tests can be carried out as crowd tests with only slight adjustments. Furthermore, we analyzed the impact of like and view statistics shown along with the images and found that there is no influence on the appeal ratings. As a further step, the ratings of the subjective tests are used to evaluate state-of-the-art models as well as prediction models developed ourselves as part of this research. To this aim, we trained several models, e.g., using signal features or deep learning-based features for regression and classification of image appeal. It is shown that the models have a similar or better performance as compared to state-of-the-art models. In general, the classification task is easier for prediction, in contrast to the regression approach. We further re-trained parts of already existing deep neural networks using transfer learning for the regression formulation, and it is shown that such models perform well, however, bigger datasets are required for a proper evaluation.

The dataset, subjective annotations, models, and evaluation results are publicly available[3] to enable reproducibility and allow additional research.

The article is organized as follows. In the subsequent Section II, a brief overview of state-of-the-art image appeal assessment is provided. In the subsequent Section III, a description and exploration of the **AVT-ImageAppeal-Dataset** are presented. In Section IV the subjective tests are described and analyzed. The state-of-the-art and own prediction models are covered in Section V. The paper ends with a discussion in Section VI and a conclusion with an outlook on future work in Section VII.

II. OVERVIEW OF IMAGE APPEAL

In the following, a brief overview of image appeal and aesthetics will be provided. After a brief definition of image appeal and a conceptual model, we highlight three different aspects. Firstly, a brief overview of commonly established photo rules is given. Following such rules will usually help to improve the appeal of an image. Secondly, state-of-the-art publicly shared datasets are described, which are used in various follow-up works. Lastly, an overview of several prediction models for image appeal is provided. Here, the focus is on models which have been developed during the last years and are shared for researchers as open source.

A. DEFINITION AND CONCEPTUAL MODEL OF APPEAL

Appeal comes from the Latin "appelare" (to address), combining "ad" (to, towards) and "pellere" to drive.[4] In its widespread use, the noun appeal generally relates to "the quality of being attractive or interesting", and hence relates to the properties of the picture itself, including the depicted content. An aesthetic picture is defined as "giving or designed to give pleasure through beauty",[5] and thus relates to a set of features of the picture which are inherent to the depicted subject or have been chosen to please the spectator. The term beauty describes "a combination of qualities, such as shape, color, form, that please the aesthetic senses, especially the sight".[6] In the following, beauty is considered to result from the overall set of aesthetic features of a picture. It is noted that according to different accounts, beauty and the

[2] https://github.com/idealo/image-quality-assessment

[3] https://github.com/Telecommunication-Telemedia-Assessment/sophoappeal

[4] https://en.oxforddictionaries.com/definition/appeal

[5] https://en.oxforddictionaries.com/definition/aesthetic

[6] https://en.oxforddictionaries.com/definition/beauty

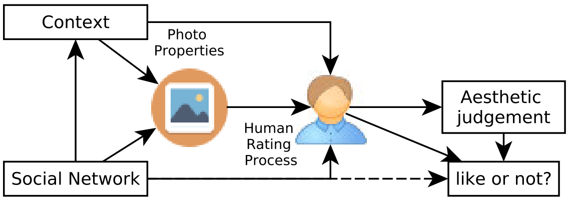


FIGURE 1. How humans rate aesthetic and decide liking [4], based on Leder et al.'s model [1].

underlying aesthetic properties of an object are not related to whether or not the object is desirable, however, may make an object desirable. It is further noted that a piece of art is not necessarily beautiful, and "good art" may appeal to mental processes or political or moral reasoning rather than leading to pleasure derived from beauty (see e.g. [16]). According to Zanwill [16] referring to Kant's "judgment of taste" in terms of "beauty and ugliness" [17], aesthetic judgment is both subjective and universal. Subjective here means that it is based on a feeling that may be related to e.g. pleasure or displeasure. Universal here means that the underlying properties based on which an aesthetic judgment is made can be generalized across subjects or groups of subjects to some extent. The characteristic of being universal is related to that of being "objective" (generalization across subjects). Note that the universal appreciation of an image by subjects in terms of its aesthetic judgment can be thought to reflect how well it represents the type of object it is (see e.g. Sartwell referring to Greek philosophers Plato and Plotinus, see [18]). In this proposal, aesthetic judgment is referred to by the term *aesthetic appeal*, in reference to the positive end of the "beautiful – ugly" dichotomy.

The term quality has been defined as the "result of appraisal of the perceived composition of an entity with respect to its desired composition", and thus results from a comparison of the picture's properties to respective internal references based on the spectator's world knowledge [19], [20], [21]. In the context of multimedia technology and signal processing, the term *quality* is typically used as a measure of how technical factors affect perceived quality, as a result of some technical "processing" [21], e.g. due to the camera optics, the camera-internal processing, image coding or rendering on a given screen. Hence, in this proposal the term *image quality* is used in relation to properties such as image noise, sharpness, or image coding artifacts.

The process of the construction of the aesthetic appeal judgment by an observer has already received considerable attention in the literature [22], [23].

A model describing the different components of aesthetic appeal judgments was established by Leder et al. [1]. This model was designed in the particular context of modern art and is based on the principle that aesthetic appeal appreciation is closely related to the pleasure of understanding an artwork. The understanding of the work is not limited to recognizing the scene, but also the technical properties, the historical context, previous artwork, presentation (e.g. in a social network), and more. In Figure 1 a summarized and extended (for the final liking decision) view of Leder et al.'s model of aesthetics ratings is presented. Three main factors (that influence each other) are important for human aesthetics judgment: the photo, the context, and the social influences [1]. Finally, based on the aesthetics rating and social impact of a shared photo the human will decide if the photo will be liked or not. Liking prediction is therefore related to aesthetic prediction combined with image appeal (e.g. technical properties of the image) and social network properties of the user (e.g. usage pattern of the user, users community, ...).

B. PHOTOGRAPHY RULES

Overall, there exist several rules of thumb based on practical experiences for photography [24]. Examples of such rules are the rule of thirds, simplicity, balancing elements, leading lines, framing, symmetry and patterns, background/foreground, cropping, and more. In the following, we briefly describe a few of these rules.

The rule of thirds consists of dividing the space of the picture into a grid of 3 × 3 rectangles [24]. The intersections between the lines correspond to the strength points of the pictures, where the subject of the picture should be located. This is motivated by research conducted by John Thomas Smith who found that the ratio of about two-thirds to one-third harmonizes the proportions [25]. It is also mentioned in [24] that it is important to not overuse this specific rule while composing a photo, e.g., placing the subject of the photo too strictly on these grid lines.

Another rule of thumb is image simplicity or the rule of simplicity. According to Krages [24] this rule is used to highlight the subject of interest in an image. It can be done using background blurriness or a neutral background and helps the viewer to more easily process and understand the image. Moreover, it may also emphasize the reason the specific subject was chosen by the photographer.

Leading lines and their link with aesthetics are motivated by their effect on the sight and how they drive visual attention within the picture [26], [27]. This rule could be used to create a depth effect by using strong leading lines from the background to the foreground, to drive the attention of the subject to a subject of importance, to create a "visual journey" within the picture, and more.

The degree of fulfilment of such rules can automatically be assessed using approaches from computer vision. For example, leading lines could automatically be analysed using scan-path prediction models [28]. It is also possible to use the Rule of Thirds combined with leading lines to predict important parts of the image [29]. Furthermore, for the rule of thumb or image simplicity for photography, Mai et al. proposed in [30] and [31], two automated systems to detect whether these two rules are fulfilled by an image or not. In both cases, the system is based on saliency prediction models. Here, saliency refers to the contribution of the image information to the visual attention of viewers (cf., e.g., [32]). Overall, several saliency models are used to estimate the subject of interest for a given image. Afterwards, the estimated saliency maps are processed and fed into several machine learning algorithms, e.g., Support Vector Machine, Adaboost or the K-Nearest Neighbour method. For the rule of thirds classification, an accuracy score of around 0.8 is reported, whereas for image simplicity the accuracy value is around 0.89. In [33] a similar approach using saliency maps and machine learning is proposed, with accuracy values of around 0.79. In addition to traditional signal-based approaches, Göring and Raake [9] proposes to use deep neural networks (DNNs) to predict whether a photo follows the rule of thirds or the rule of simplicity. Both rule prediction problems are modelled separately and handled as binary classification problems. The overall evaluation shows promising results, especially considering that the validation data is independent of the training data. In general, several DNNs are evaluated, and for the considered datasets, the ResNet152 DNN was found to be best for the rule of thirds prediction and DenseNet121 best for simplicity, with an accuracy of around 0.84 and 0.94, respectively.

C. IMAGE APPEAL AND AESTHETIC DATASETS

In general, to develop prediction models using machine learning or deep neural networks, it is required to have annotated datasets. An important aspect of such datasets is that the annotations are of high validity and reliability. In the field of computer vision, or Quality of Experience (e.g. considering video or image quality [34], [35]), such annotations can be gathered from participants in the controlled lab or in their own environment using crowdsourcing tests. For the case of image- and video-quality assessment, using the software framework AVrate Voyager [36], it has been shown in [34] and [35] that the lab and crowd results are highly comparable, even for the case of higher resolutions and overall quality.

In the context of image appeal, crowd tests are used, too. Examples are the research studies described in [37] and [38] or by [39] for the prediction of popularity of images.

Furthermore, the AVA dataset [10] is an open-source dataset that includes aesthetic, semantic, and style annotations. The style attributes include the rule of thirds, however, image simplicity is not part of the annotations. The annotations are used in [40] in a two-stage approach, where in the first step style attributes are predicted, and in the second step the overall aesthetic score. In total, the AVA dataset consists of 250k images, which have been annotated in an online platform by various users in several photographic challenges. In general, the images have a lower resolution, which is in the range of 90px to 800px for width and height.

The evaluation of Wang et al. [40] is based on binary classification of high and low-appealing photos. With a similar classification, Kong et al. [11] describe an automated approach for photo aesthetic ranking. Here, a different dataset has been used, the AADB database [11] including aesthetic ratings and attributes. The AADB dataset includes 10k annotated images. The resolutions of the images are in the range of 219px to 1024px for height and width, respectively, therefore only lower-resolution images are included.

The A&A dataset (Aesthetics and visual Attention) [41] includes 200 images with annotations and saliency maps, however, the dataset is not public and only available upon request, which for the authors of the present paper never resulted in getting access to the dataset.

Other related datasets for image appeal are, for example, YFCC100m [42] or KonIQ-10k [12], [13]. The YFCC100m dataset does not include subjective ratings and is a generic dataset with multimedia content from Flickr which can be used to assemble specific collections. For example, the KonIQ-10k [12], [13] dataset is based on YFCC100m [42] and includes 10k images. The focus of the KonIQ-10k dataset [12], [13] is quality evaluation, however, in general, no quality degradations are included, therefore some appeal and liking aspects are assumed to be included in the overall rating. In general, the KonIQ-10k dataset does not include high-resolution images, however, due to the fact that the image IDs are still available in combination with the YFCC100m dataset, the original URLS can be estimated and high-resolution versions of the images can be downloaded from Flickr. A similar dataset is the "LIVE In the Wild Image Quality Challenge Database [14]", which contains approximately 1200 images and corresponding quality ratings.

D. IMAGE APPEAL PREDICTION MODELS

In general, the problem to predict image appeal or aesthetics can be handled as a classification or regression problem, depending on the given ratings and the overall aggregation. This is similar to video quality prediction, as described in [43]. In general, assuming discrete ratings from participants for each individual image, the majority could be estimated, or using thresholds, discrete appeal classes could be derived. In the case of a regression problem formulation, the mean scores of all ratings for a given image could be used as a prediction target.

For example, in the case of handling the aesthetic appeal prediction as a binary classification, results are reported to have an accuracy of around 0.8 in [44]. In [11] a cross-dataset evaluation achieved accuracy scores of 0.15-0.31. This also shows that a unification for aesthetic datasets or models is required, which is presented in [45]. Abdenebaoui et al. use deep learning to predict image quality, semantical quality, and description of photographic rules for the datasets AVA and AADB, respectively.

Photographic rules are also the focus of prediction systems. For example, Dhar et al. [46] describe a system that uses classification of whether an image follows certain rules or not for the overall prediction of the "interestingness" of an image. Moreover, individual features to estimate image appeal are also presented in other studies, e.g., by Marchesotti et al. [47] where several signal and photo rule image features (low- and mid-level features) are used to estimate aesthetic quality. However, generic image appeal includes different further aspects, as it is analyzed by Machajdik and Hanbury [48], or Lebreton et al. [2] considering user's knowledge, Lebreton et al. regarding technical knowledge of the participants or by Göring et al. [4]. For example, in [4], several feature groups such as low-, mid-, and high-level features including social network aspects of photo-sharing platforms are used to estimate the so-called likeability of photos, which is related to image appeal (cf. Sec. II-A). Because DNNs tend to have high prediction accuracy for image classification tasks [49], they are also used for image appeal prediction [6], [7], [8]. For example, in [15] a deep neural network (NIMA) for image quality *nima_q* and image aesthetic *nima_a* prediction is presented. The NIMA model uses a baseline DNN, which is fine-tuned with additional layers for the new classification or regression task. The training for the appeal prediction part was performed with the AVA dataset [10]. NIMA predicts, for each discrete rating, a probability which is then used to calculate the overall image appeal. Similar is the approach presented in [50], which also uses DNN fine-tuning.

Image appeal can be estimated for traditional real photos, or also for art, or computer-generated content. For example, Ling et al. [51] analyze image appeal of smartphone game screenshots with several dimensions, and they show that, e.g., the CPBD contrast feature correlates best with approximately 0.48 Pearson correlation. This work has been extended by Lei et al. in [52] to also include a no-reference deep learning-based prediction model, which predicts the four dimensions of the dataset, namely fineness, colorfulness, 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 presented in this paper.

Paintings are analyzed in the research conducted by Amirshahi et al. [53]. Overall, a dataset with 281 images of paintings is used. 49 participants were asked to rate the appeal using a discrete rating scheme (1-4 rating). The analysis indicates that color has the highest influence on image appeal. Furthermore, a prediction model trained and evaluated with a 75-25 train-validation split has been proposed, showing promising results [53]. In addition, Bo et al. [54] provide a review for image appeal research covering generative art and applications for design. The overall focus of the paper is on fractal art, abstract paintings, and how to use computational aesthetics measures to improve them.

In general, image aesthetics or appeal assessment can be handled with several approaches. For example, the research conducted by Xu et al. [55] identifies two main points in the field of aesthetic assessment. Namely, the majority of approaches model image aesthetics as a classification or regression problem and ignore context information thus using only the image itself for the prediction. They train a model with context information and predict distributions of ratings using the AVA dataset with state-of-the-art performance. Similarly, Hou et al. [56] are handling the image aesthetics problem as a prediction of discrete rating distributions considering the region of interest and objects of the images. Furthermore, also structural aspects are important for image appeal, as, e.g., the photographic rules indicate. In [57] the aspect ratio and spatial layout of images for the overall appeal are analyzed. The overall model uses two stages, first, a feature-graph representation is extracted and then a graph neural model covers the overall aesthetic prediction. The model is trained using the AVA dataset.

To sum up, there are several prediction models for image appeal or natural, artistic, or generated images. The majority of models are using the AVA or AADB dataset for training and validation. Furthermore, most research handles the idea to model image appeal as pure regression or classification and shows promising results. There are still open points, e.g., not all models and data are shared and accessible, and most evaluation experiments are just performed with lower-resolution images or with a small number of images.

III. AVT-ImageAppeal-DATASET

To evaluate image appeal, we created our own dataset, namely the **AVT-ImageAppeal-Dataset**. In general, this dataset consists of images from several sources, also to evaluate different aspects of, for example, commonly established datasets for image appeal, or relevant for different photo-sharing platforms. Accordingly, the dataset includes the number of likes, number of views, and meta-data for the majority of images. In the following, a brief overview of the steps, which have been involved in the creation of the **AVT-ImageAppeal-Dataset**, are described. Firstly, we checked various photo-sharing platforms. Afterward, state-of-the-art image appeal datasets are described with the focus of including parts of them in the **AVT-ImageAppeal-Dataset**. This is followed by an in-depth description of sampling and a characterization of the dataset.

A. PHOTO-SHARING PLATFORMS

In Table 1 an overview of several popular image-sharing platforms is given. Each platform is analyzed considering its image license, contained meta-data, and social data. Not all platforms are suitable because one main requirement is that the photos (and additional data) can be shared later. The most promising photo-sharing platforms are Flickr, Pixabay, and 500px when considering CC0-marked photos. For these three platforms, custom web crawlers have been developed to download the required data. For example, the photos in some platforms could be downloaded with *gallery-dl*,[7] however, this tool does not download, e.g., social data. Because of this, it was needed to develop own crawlers. They are mostly developed for the purpose of downloading the data at the

[7] https://github.com/mikf/gallery-dl

specific time and will be invalid in the future, due to changes in the corresponding photo-sharing platforms. For this reason, the code of the crawler is not published, because maintenance would be required. The general crawling procedure is as follows:

- open the webpage with chrome (automated with python selenium[8])
- collect links of shown photos in the selected list, e.g. upcoming or new photos; this selection is specific to the platform
- after storing all links and pre-selecting using the social-data/meta-data and license (CC0):
  - open each link individually with Chrome and get the image in the highest resolution and additional data
  - in case of Pixabay/Flickr screenshots of the shown webpage and HTML dump are included, which can be used for later verification

B. AVAILABLE IMAGE APPEAL DATASETS

As described in Section II-C, there are various publicly available datasets for image appeal, aesthetic or quality assessment.

In Table 2 several published image appeal or quality-related datasets are summarized. Although a number of databases are contained, this list is not complete. Here, the focus is on datasets that can be extended by additional data (e.g. social or meta-data that is not included in most of the datasets). For example, the YFCC100m does not include subjective ratings and rather is a "generic" dataset to assemble specific collections, such as, for example, the KonIQ-10k [12], [13] dataset which is based on YFCC100m [42]. The most promising datasets are AVA [10], AADB [11], and KonIQ-10k, however, they do not include high-resolution images. In the case of KonIQ-10k, this can be bypassed. Here the high-resolution images can be downloaded using the YFCC100m dataset and a Flickr-specific crawler, based on a mapping of image IDs in the KonIQ-10k to URLs to the YFCC100m dataset.

C. DATASET SAMPLING

For the datasets mentioned in the previous section or several other related works in the field of image appeal, aesthetics, or liking prediction, images are typically sampled out of a larger dataset. This larger dataset could be considered to be the set of photos available on the internet or for specifically selected photo-sharing platforms. For this paper, photo-sharing platforms are preferred with their included social data, otherwise, datasets such as ImageNet [49] could be considered as the main source for the database as well. In [49] the focus – image object classification – is different, thus it is not required that the ImageNet photos cover a wide range of appeal and furthermore, there are no social data included. For example, Datta et al. [58] uses 3581 photos from photo.net, and the overall sampling is based on the platform scores considering liking. Gelli et al. [59] use a dataset based on Flickr, with a randomly selected subset of images. In other works, e.g., [11], [14], [60], the specific sampling method is not mentioned. The KonIQ-10k dataset [12], [13] uses a random selection, then filters by license, then samples by tag distribution. The final selection uses specific extracted image features.

The dataset generation and sampling for the **AVT-ImageAppeal-Dataset** is based on the following steps. First, several image sources are considered. Each of the different image sources is sampled based on either subjective appeal ratings or like values available from the photo-sharing platform. The given ratings are normalized to a [0,5]-scale and rounded to integers. Afterward, for each of the integer bins, a uniform random selection is applied to derive the sample, e.g., 40 images of each bucket. With this approach, it was possible to sample around 150-220 images per given data source, compare Table 3. It is noted that some of the integer bins do not include 40 images, which is due to the rareness of e.g. highly appealing images in some of the published datasets.

The final **AVT-ImageAppeal-Dataset** is assembled from six different image sources, namely AVA [10], AADB [11], KonIQ-10k [12], [13], Pixabay, 500px_cc0 and OWN images. In Table 3, an overview of the sampled images with the selection criteria is given. The data sources AVA and AADB are included in the dataset to enable a cross-comparison to previously conducted appeal research. After the sub-sampling of the original dataset, another filtering or cleaning step was performed. Here, photos that were not natural (e.g. synthetic photos, or photos of drawings), photos with topics covering "army", "military", "weapons", "babies" or "explicit sexual content" have been removed. Further, some images have been excluded because they had white or black borders. In the case of AVA, these borders have manually been removed. The KonIQ-10k images are high-resolution replacements of the images initially published with KonIQ-10k, and the ratings are based on expert annotation (with additional filtering of non-fitting content). This annotation was performed manually using AvrateNG[9] considering quality with one expert. The 500px_cc0 images are downloaded from the platform using the #cc0 image tag, reflecting that a previously created dataset (see [4]) did not include a sufficient number of CC0 licensed photos. The photos from Pixabay (pixabay_first50k) are downloaded with a custom crawler, overall the meta-data has been downloaded first, and then according to the sampling criterion, which was $log(like)/log(view)$, the final images have been selected. The first 50k images sorted by date of the Pixabay webpage have been considered for the sampling. The image source indicated by "OWN" are images from the first author, here no social data is included.

[8] https://www.selenium.dev/

[9] https://github.com/Telecommunication-Telemedia-Assessment/avrateNG

TABLE 1. Overview of different photo sharing platforms.

Name	URL	License	Meta-data	Social-data
1x	https://1x.com	closed	yes	favourites, views, no likes
unsplash	https://unsplash.com	free	yes	views, down., no likes
pixabay	https://pixabay.com	free	yes	views, down., likes, favourites
freeimages	https://freeimages.com	no sharing	yes	likes, dislikes
deviantart	https://deviantart.com	per image	no	favourites, comments
imageshack	https://imageshack.com	unclear	no	no
flickr	https://flickr.com	per image	yes	favourites, views, comments
instagram	https://instagram.com	hashtag-based	no	likes, comments, no views
500px	https://500px.com	CC0 tags	yes	likes, views,
fotki	https://fotki.com	unclear	yes	views
vero	https://vero.co	unclear	no	yes
behance	https://behance.ne	unclear	no	yes
tumblr	https://tumblr.com	unclear	yes	yes
signatureedits	https://signatureedits.com	unclear	no	no
youpic	https://youpic.com	by user	yes	likes, views
pxhere	https://pxhere.com	per image	yes	likes, shares
DPChallenge	https://www.dpchallenge.com/	unclear	no	likes/ratings, views
photo.net	https://www.photo.net/	unclear	no	views, favourites, impression

TABLE 2. Overview of different image appeal or image datasets. low resolution = below 1080p height.

Name	URL	Source	# Photos	Comment
AVA	https://github.com/mtobeiyf/ava_downloader	[10]	250k	low resolution
AADB	https://github.com/aimerykong/deepImageAestheticsAnalysis	[11]	10k	low resolution
KonIQ-10k	http://database.mmsp-kn.de/koniq-10k-database.html	[12, 13]	10k	quality oriented, no social data, based on YFCC100m
LIVE Wild	https://live.ece.utexas.edu/research/ChallengeDB/	[14]	1.2k	quality oriented, low resolution
YFCC100m	http://projects.dfki.uni-kl.de/yfcc100m/	[42]	100m	based on flickr; only meta-data: images need to be downloaded separately
A&A	http://ii.tudelft.nl/iqlab/A&A.html	[41]	200	not public, low number of images

TABLE 3. Overview of image sources for the AVT-ImageAppeal-Dataset with the sampling criteria and the number of selected/filtered images.

Source	Sampling based on	# Sampled images	# Images after cleaning
AVA	ratings	201	186
AADB	ratings	212	183
KonIQ-10k	expert ratings	200	197
Pixabay	$log(like)/log(view)$	199	191
500px_cc0	$log(like)/log(view)$	190	168
OWN	none	69	136
			$\sum = 1061$

In total, the resulting dataset comprises 1061 images after cleaning, from different realistic photo sources and is referred to in the following as **AVT-ImageAppeal-Dataset**. For the different datasources we considered to sample approximately 200 images before the manual cleaning step, the exact numbers for each data source are listed in Table 3, to have approximately 1000 images in total in the dataset. E.g., for the AVA dataset we sampled 201 images and included in the final dataset 186 images. Furthermore, the **AVT-ImageAppeal-Dataset** consists of images from various sources, to also represent a wide range of contents. In case appeal ratings are used for sampling, this ratings can be later used for appeal prediction analysis. In case likes and views are used for the sampling, these values can be analyzed considering the appeal ratings of the overall **AVT-ImageAppeal-Dataset**.

Even though Instagram could have been considered as one of the sources, here the platform license disallows downloading and further distribution. In Figure 2, a brief overview of the included images is shown. In comparison to other datasets, our dataset includes higher resolution images, like and view values for a subset of the images, other appeal ratings from the state-of-the-art datasets for cross comparison. All images follow a `CC0` license and the likes and views are not associated with specific users, furthermore the annotations which are done in the lab and crowd tests are also anonymously shared.

D. PROPERTIES OF THE AVT-ImageAppeal-DATASET

In the following, we briefly characterize the dataset using image features and properties.

First, image resolution is important and some published datasets only include lower-resolution images. In Figure 3, an overview of the image resolutions included in the **AVT-ImageAppeal-Dataset** is given. Here, the heights and widths of all images are extracted, and each subset is individually visualized. It can be seen that the images from AVA and AADB have the lowest resolution, which is followed by the images from Pixabay. The subsets 500px_cc0, koniq10k, and OWN have the highest resolutions, also, e.g., the images from koniq10k have been updated by the high-resolution versions. Overall, the heights/widths are in a range from 267 px to 5616 px.

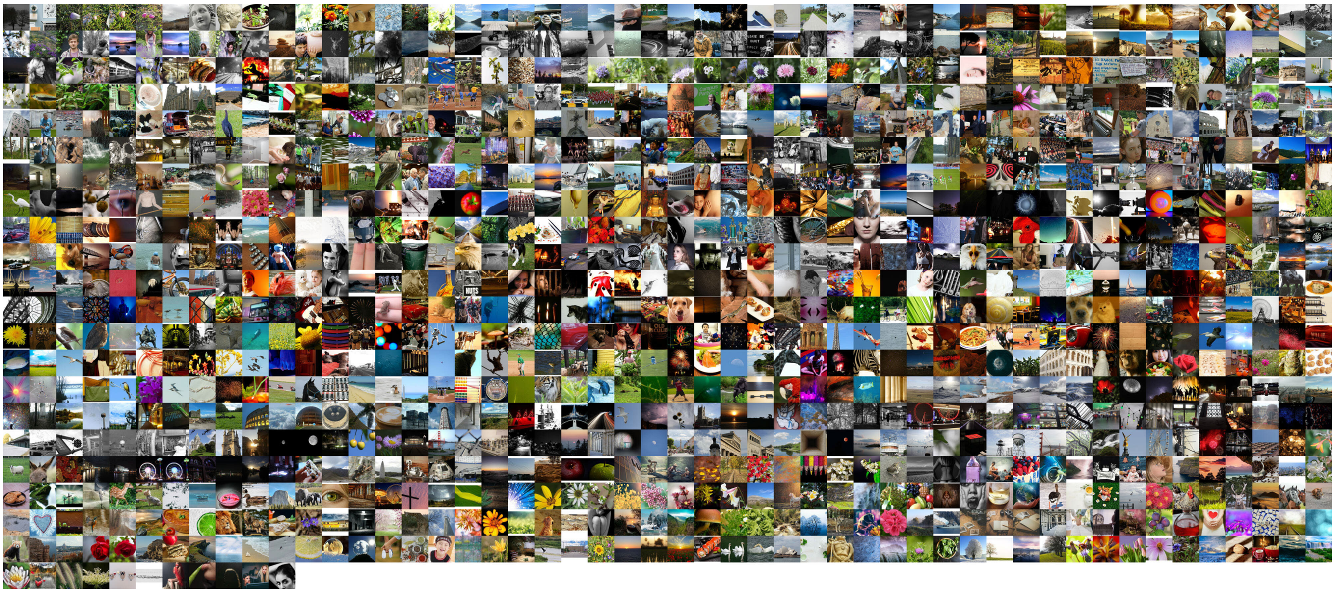


FIGURE 2. All images of the AVT-ImageAppeal-Dataset; shown are centre-cropped variants of each image individually.

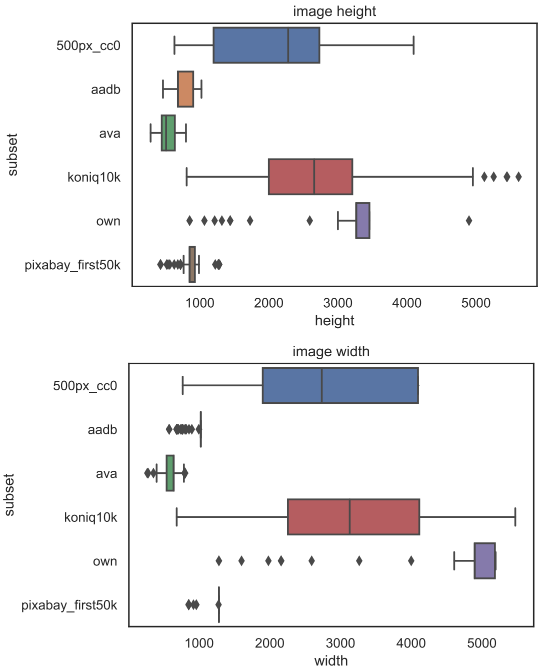


FIGURE 3. Height and width boxplots for the AVT-ImageAppeal-Dataset.

We calculated several state-of-the-art features,[10] which are used in the context of image appeal or image quality. In total, we calculated 9 features, namely blur strength,[11] colorfulness, contrast, CPBD,[12] NIQE, noise, saturation, spatial information (SI), and tone. All features have been calculated for all images which have been resized to a common height of 1000px for unification. All features with their corresponding references are listed in Table 4. Note that the CPBD feature has been showing good performance for image appeal prediction in the case of mobile game images, according to Ling et al. [51].

TABLE 4. Image appeal and quality-related features with their corresponding sources.

Feature	Source
blur strength	[61]
colorfulness	[62]
contrast	[43]
cumulative probability of blur detection (CPBD)	[63]
natural image quality evaluator (NIQE)	[64]
noise	[65]
saturation	[66]
spatial information (SI)	[67]
tone	[66]

In Figure 4, example boxplots for the features CPBD and SI are shown. It can be seen that each subset of the **AVT-ImageAppeal-Dataset** has a similar value range for the two features. Especially in the case of SI, the ranges are mostly overlapping. For the CPBD feature, the AVA subset spans a wider range, however, the other ranges are similar and overlapping. In Table 5, a full overview including all calculated features is provided. Here, only statistical values, min, max mean, median, and standard deviation are listed. For most of the included features, a wide range of values is used, which indicates that the content in the dataset is diverse.

In addition to the aforementioned low-level image features, we calculated saliency maps, image depth maps, semantic segmentation maps, and visual sentiment as feature maps or values targeting a more high-level view of the images.

Image appeal or aesthetics usually also depends on the photo topic and selection, for this reason, saliency models can

[10] implementation: https://github.com/Telecommunication-Telemedia-Assessment/sophoappeal_image_features_tool

[11] code: scikit-image [68]

[12] code: https://github.com/0 × 64746b/python-cpbd

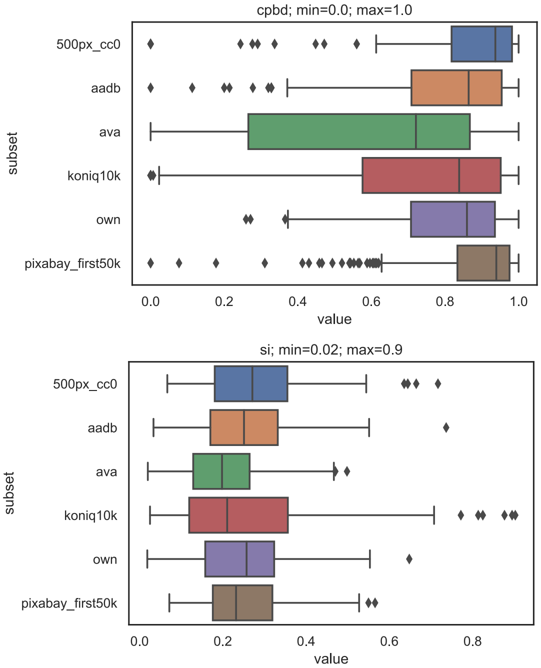


FIGURE 4. CPBD and SI feature boxplots for the AVT-ImageAppeal-Dataset.

TABLE 5. Statistical description of image feature values for all images of the AVT-ImageAppeal-Dataset; values are rounded to two decimals.

Feature	min	max	mean	median	std
blur strength	0.13	0.79	0.37	0.36	0.11
colorfulness	0.00	187.56	115.97	132.44	48.48
contrast	-111.50	99.78	86.49	88.41	13.24
CPBD	0.00	1.00	0.77	0.87	0.27
niqe	14.97	27.08	19.43	19.70	1.43
noise	0.00	22.75	1.03	0.58	1.48
saturation	0.00	98.72	36.93	34.10	22.37
SI	0.02	0.90	0.25	0.23	0.13
tone	0.00	0.99	0.59	0.65	0.26

be seen as a set of features to estimate where the important part or region of a photo is located. An overview of several state-of-the-art saliency models and their corresponding performance is given in [69]. Here, UNISAL [70] and DeepGaze II [71] are listed in the top-5 performing models and have therefore been selected. The **AVT-ImageAppeal-Dataset** includes saliency map predictions of the previously mentioned models and values for a statistical evaluation as additional metadata. For example, in Figure 5 an example image with the corresponding saliency map predictions is shown. The estimated saliency maps will be used as further input for feature calculation.

In Figures 6 and 7, the evaluation for the estimated saliency models is shown. For both saliency prediction models the saliency maps have been extracted. Using these maps, the mean and standard deviation are estimated. Moreover, for each of the saliency maps, a binary thresholding was applied followed by labeling the connected components using scikit image's `measure.label`[13] [72], [73]. The estimated connected components are an indicator for whether there is only one region of interest or several. For this reason, the plots 6 and 7 include the number of connected components. It should be mentioned that in the case of the UNISAL model, the plot is reduced because it seems that the UNISAL model is creating many separated regions of interest. Comparing both saliency models, it can clearly be stated that the DeepGaze II model seems to create less separated components in the saliency maps. However, both models are well performing in the saliency competition. We include them to also have a wider range of feature values for the appeal prediction.

Similarly to saliency maps, depth information may be important for a scene of an image. For this reason, we calculated image depth maps, and furthermore segmentation maps. In Figure 8, estimated depths and image segmentation using the deep learning approach described in [74] and [75] are shown. For all images of the **AVT-ImageAppeal-Dataset**, such depth and segmentations maps have been extracted and are included in the dataset.

For the depth maps, mean and standard deviation have been calculated. Furthermore, for the segmentation maps, the number of segments and mean and standard deviation of the map (using an RGB-color to number mapping) are calculated.

In Figure 9, an overview of the estimated depth values is shown, for both mean and standard deviation boxplots are given. Furthermore, in Figure 10 the semantic segmentation boxplots for mean, standard deviation, and the number of segments are presented. The depth maps indicate that the dataset has various images included. Furthermore, the semantic segmentation maps verify that most of the images contain only a few segments, which is an indicator of image complexity.

Vadicamo et al. [76] proposed an approach using DNNs to estimate the visual sentiment of an image; an open source implementation of the prediction model has been made available, too.[14] In general, the visual sentiment of an image can be -1 (negative), 0 (neutral), or 1 (positive) and refers to the emotion that a person would associate with the image [77]. In Figure 12, boxplots for the three sentiments negative, neutral, and positive (neg, neu, pos) for the **AVT-ImageAppeal-Dataset** are shown, as calculated using the model described in [76]. Furthermore, in Figure 11, example images for the three sentiment cases are shown (selection was based on top-10 values for each of the categories). It can be seen that most of the images are neutral or positive, considering Figure 12. Because image appeal is also related to emotions and thus connected to visual sentiment, this feature may be relevant for the appeal prediction.

To which extent photographic rules have been considered may also have a strong influence on the appeal of an image,

[13] https://scikit-image.org/docs/dev/api/skimage.measure.html#skimage.measure.label

[14] https://github.com/fabiocarrara/visual-sentiment-analysis

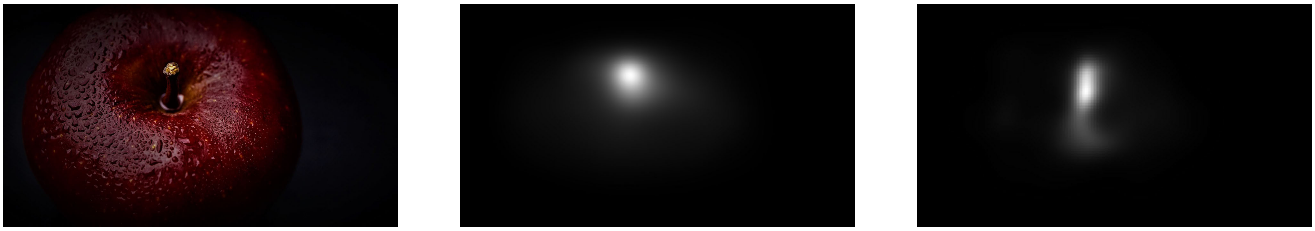


FIGURE 5. Saliency prediction: example image (left), and UNISAL (middle), Deepgaze II (right).

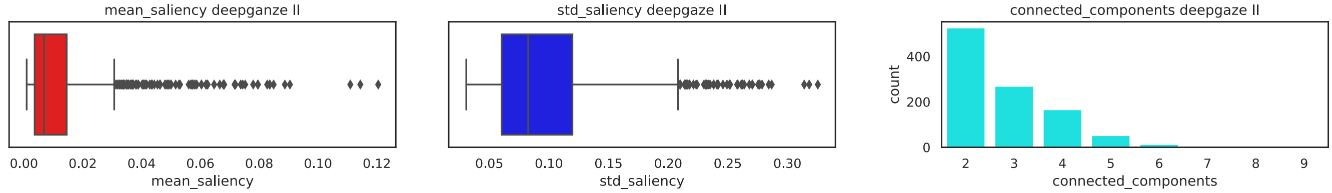


FIGURE 6. Plots for saliency evaluation considering the DeepGaze II model: mean (left), standard deviation (middle) and number of connected components (right).

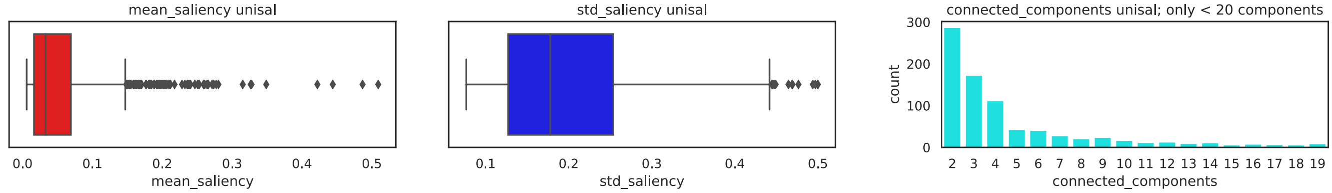


FIGURE 7. Plots for saliency evaluation considering the UNISAL model: mean (left), standard deviation (middle) and number of connected components (right).

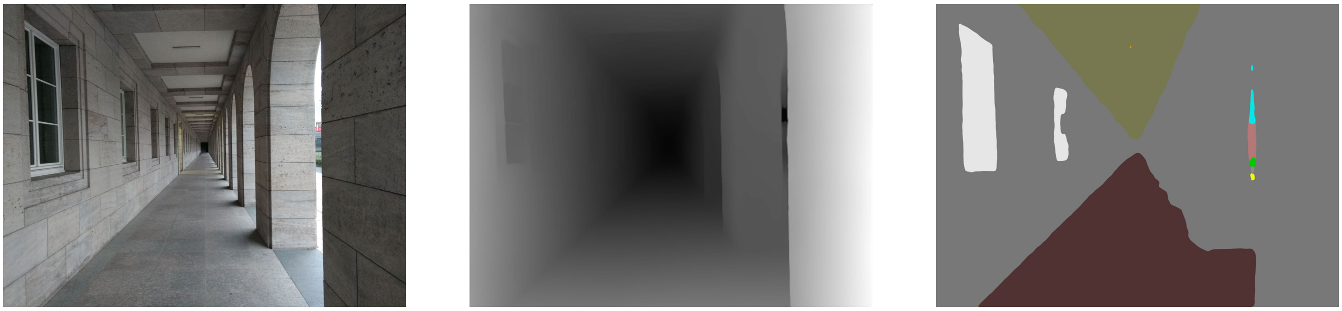


FIGURE 8. Image (left), depth map (middle), and semantic segmentation of the image (right).

for example, whether an image is following the rule of thirds or not. Another aspect may be to consider that simpler images are easier to parse and highlight more of the important part of the image. For this reason, we developed a deep neural network-based approach [9] to classify images in terms of whether they follow certain photo rules or not. The neural network and software implementation are made publicly available.[15]

In Figure 13, prediction results for both rules, namely the rule of thirds and image simplicity, are summarized. In general, it can be stated, that, for example, the AVA, Pixabay, and Koniq10k subsets of the dataset include more images that are classified as simple. Furthermore, the AVA and 500px_cc0 subsets include more images that follow the rule of thirds. Overall, the automatic annotation indicates that the **AVT-ImageAppeal-Dataset** spans a wide range of images, and the prediction results can be used as further features for image appeal prediction.

IV. SUBJECTIVE TESTS AND EVALUATION

To evaluate the appeal or aesthetics of the images of the **AVT-ImageAppeal-Dataset**, several instances of subjective evaluation tests have been carried out. The tests, which are described in the following, were handled as crowd-sourcing and lab-based tests to enable a comparison between both paradigms, similar to the work in [34] and [35] for quality assessment. As instances for the subjective tests, the following two variants are considered. First, a pure appeal

[15] https://github.com/Telecommunication-Telemedia-Assessment/sophoappeal_rule_prediction

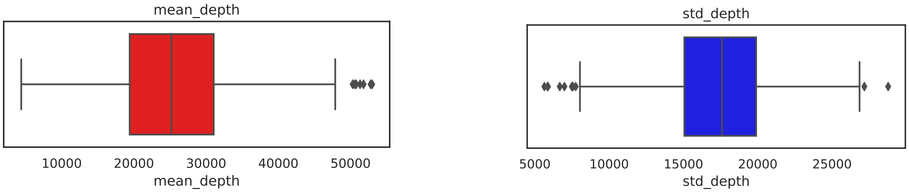


FIGURE 9. Boxplots for depths evaluation: mean (left) and standard deviation (right).

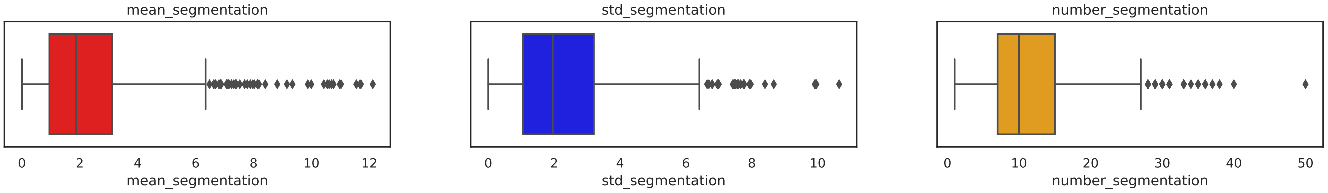


FIGURE 10. Boxplots for segmentation evaluation: mean (left), standard deviation (middle) and number of segments (right).

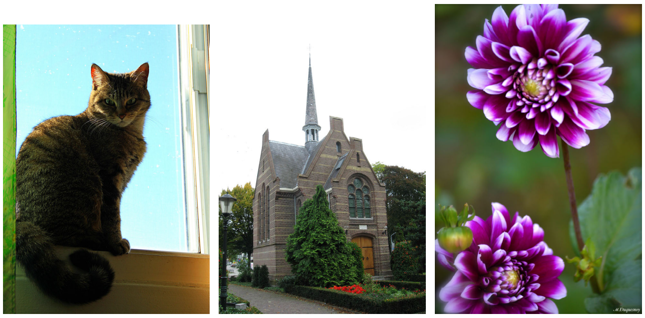


FIGURE 11. Visual sentiment examples: max negative (left), max neutral (middle), and max positive (right) for the AVT-ImageAppeal-Dataset.

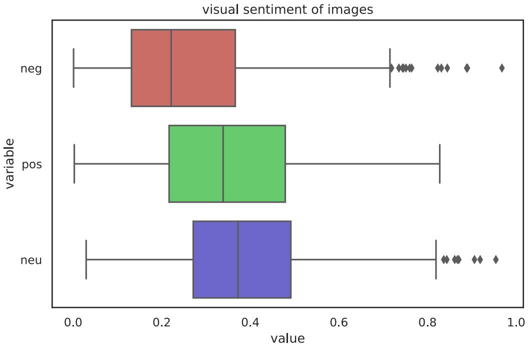


FIGURE 12. Boxplot of predicted visual sentiment of the AVT-ImageAppeal-Dataset.

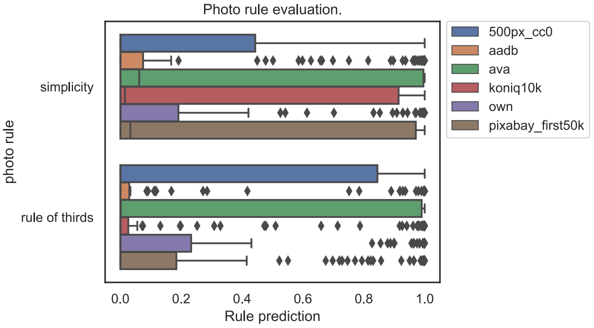


FIGURE 13. Boxplots for Image simplicity and rule of thirds prediction for the AVT-ImageAppeal-Dataset on a subset level, x-axis: in case a rule is followed the prediction is 1, otherwise 0.

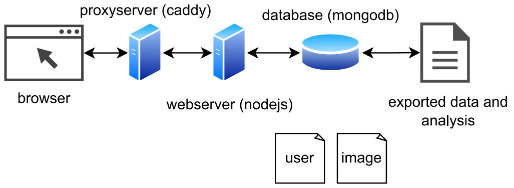


FIGURE 14. Web-based client-server test framework.

rating, where only the image is shown. And second, a rating where the image and additionally like and view statistics are shown.

A. GENERAL RATING FRAMEWORK

The general rating system is based on a client-server web-based architecture. The lab and the crowd tests both use the same system, where the focus in the development of the rating framework was the crowd-sourcing tests. The core itself is similar to AvrateNG.[16] However, it was newly developed to improve stability, similar to AVrate Voyager [36] that is an additional result of the work described in this paper, providing a generic online test framework as open source.[17]

In Figure 14, the general communication architecture is shown. The participant connects via a web browser to the test software and the proxy server (Caddy[18] is used) redirects the requests to the web application. The web application uses

[16] https://github.com/Telecommunication-Telemedia-Assessment/avrateNG

[17] https://github.com/Telecommunication-Telemedia-Assessment/AVrateVoyager

[18] https://caddyserver.com/

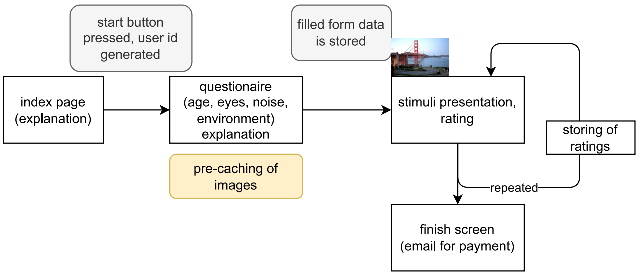
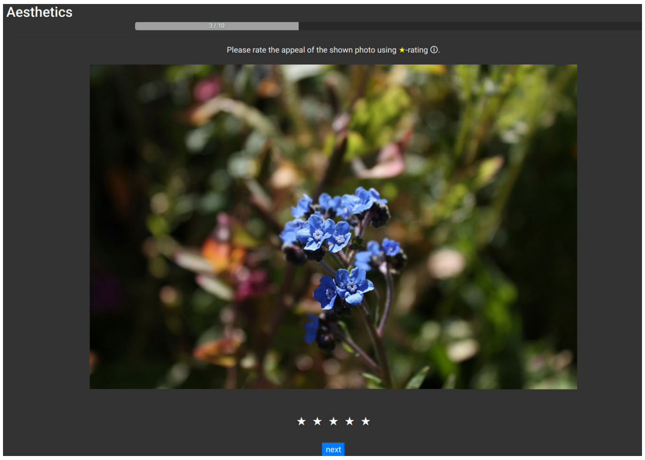

start button pressed, user id generated
filled form data is stored
index page (explanation)
questionaire (age, eyes, noise, environment) explanation
stimuli presentation, rating
storing of ratings
pre-caching of images
repeated
finish screen (email for payment)

FIGURE 15. **Test procedure with back-end specific steps.**

nodejs[19] for the application logic and mongodb[20] to store the data. For nodejs the express[21] framework is used. The database stores two types of data, first the image filename with the rating decision that is associated with a specific user and the data of the user. The user is represented using a unique user id, and the questionnaire data is assigned to the user. The overall data is anonymized. Cookies are used to prevent users to cheat or modify different rating aspects, e.g., a user cannot perform the test twice (only if the cookies were deleted). The views follow a responsive web design and use HTML5 and CSS along with bootstrap.[22] AVrate Voyager [36] can be seen as an extension of this procedure, using Python 3 and different web frameworks in the core.

In Figure 15, the general test procedure including web application-specific steps is shown. First, the participant opens the test URL with a given browser, and due to the usage of modern web frameworks, there is no limitation for the used browser. On the first page, a check whether the screen resolution is sufficient is performed, which is estimated based on the browser window size. Afterward, the unique user id is generated, the shown images are defined in the server back-end, and the questionnaire is shown. While filling out the questions and reading the explanation text, the stimuli are pre-cached, this ensures that the test can be performed smoothly also with lower bandwidth connections. Here, it should be mentioned that the test still requires a minimum of 3 Mbit/s internet connection, and it is assumed that most crowd users have access to such internet speed. After the questionnaire is filled, the data is stored in the back end and the stimulus rating procedure starts. Depending on the test instance the procedure varies, in general, for each stimulus shown the ratings are stored. Then, after the stimuli are rated, a final screen is shown, where the participant has the possibility to enter their email address used to get the payment or to get the confirmation code for the crowd-sourcing platform and subsequent payment. For data protection reasons, the email address itself is not stored in the back-end, only an email is generated and sent to start the payment procedure.

The tests are based on Absolute Category Rating (ACR). Here, the image (with or without additional data) is shown



Please rate the appeal of the shown photo using ★-rating ⓘ.

FIGURE 16. **ACR style rating; the image is shown in the centre of the screen, and below a star rating from 1-5 is visible, where the user needs to rate the appeal.**

and the user is asked to rate the appeal using a discrete scale from 1 to 5. An example is visualized in Figure 16, which is purely focusing on the image and an appeal rating, which forms test #1. In contrast to traditional quality assessment ratings, a 1-5 ⋆ star rating is used. Before the test is started, an explanation of the used rating scheme, a definition of image aesthetics, and some basic questions (age, self-reported vision, self-reported environment (noise), and device type) are provided as test instructions. In the case of a crowd test, only a sub-sample of all the images of the **AVT-ImageAppeal-Dataset** are shown and rated by a given participant. Here, 200 images are randomly selected from all images (uniform distribution). Because such a reduction leads to lesser time required for the test, the approach is changed for the lab test, where each user rated 600 or 800 randomly selected images (originally 600 was selected and changed to 800 because participants were faster than expected in the lab test). In both cases, we verified that the overall duration of the tests is fitting, e.g. for crowd tests approximately 15 minutes, and for the lab test maximum of 60 minutes, breaks were allowed in-between. A similar sampling approach for individual participants has been successfully used in [34] and [35].

In Figure 17, the extension of the rating view, including views and likes of the given images, is shown. In the following, this setup is referred to as test #2. Only 359 images in the dataset have like and view ratings because they originate from real photo-sharing platforms. In test #2, each of the images is shown twice, thus having 718 stimuli to be rated by the participants in a randomized order. The image is shown with the real like and view values and another time with simulated like values. The simulated like values are referred to as *fake_likes* in the following, and are calculated as shown in Equation 1.

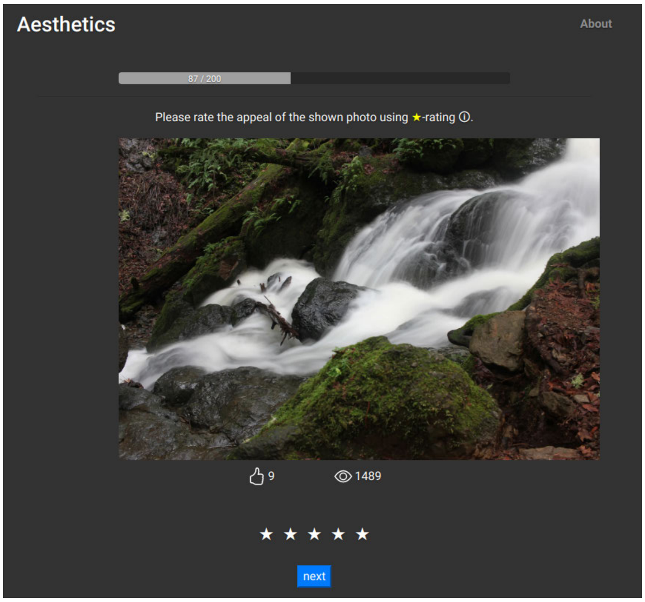
$$fake\_likes = int\left(views \cdot clip\left(max\_ratio - \frac{likes}{views}\right)\right) \quad (1)$$

[19] https://nodejs.org/en/

[20] https://www.mongodb.com/de

[21] https://expressjs.com/

[22] https://getbootstrap.com/



Please rate the appeal of the shown photo using ★-rating ⓘ.

FIGURE 17. Likes and Views Extension of the rating window with ACR style rating; the image is shown in the centre of the screen, with the 1-5 rating, and additionally likes and views of the image are shown.

TABLE 6. Number of ratings and participants for the differently tests and their types; `tuil` and `click` are online tests, while `lab` are lab tests.

Test	Type	# Participants	# Ratings
#1	`tuil`	148	24924
#1	`click`	139	26240
#1	`lab`	34	26400
#2	`click`	108	20938
#2	`lab`	24	16514

For *max_ratio*, a value of 0.5 has been estimated using the overall distribution of all *ratio* values in the dataset. *clip* is a function, which limits the values between 0 and 1, while *int* converts the float number to integer with truncation. Furthermore, *likes* and *views* are the corresponding likes and view values of the image. The idea for the *fake_likes* generation is to mirror the likes from e.g. many likes to few or vice versa. Hence, the focus of test #2 is twofold. Firstly, to check to which extent the real like and view values have an influence on the appeal rating. And secondly, to compare the real and fake like values and to analyze how they may influence the overall rating.

B. CROWD AND LAB TEST FOR TEST #1

For test #1 we conducted three different iterations. The first iteration was performed within the university, where participants have been recruited via email reflectors (students and staff members) and took part in the study with their own devices online. As recompensation, the participants took part in a lottery and could win 10 € as a voucher. This iteration is marked as `tuil` in the following. Furthermore, we recruited participants from clickworker[23] for the second instantiation of test #1, which is referred to as `click` in the following. Each of the clickworker participants got a payment of 2.38 €, which was recommended for a test with 15 minutes duration. Furthermore, we conducted the same test in a lab setting, which is referred to as `lab`. In contrast to the online tests, which were planned with a duration of 15 minutes, the `lab` test was designed with a duration of 60 minutes for each participant. This was realized with the increase of the number of images to be rated by the participants from 200 (for the online tests) to 600 or 800. Participants in the `lab` test were asked to undergo a simple vision test using Snellen charts as recommended in ITU-T Rec. P.910 [67]. The `lab` test was conducted in a controlled lab setting following ITU-T Rec. BT.500-13 [78] and ITU-T Rec. P.910 [67] with appropriate lighting conditions and a viewing distance of 1.5×H, with H being the height of the screen, which is similar to a typical desktop screen setup.

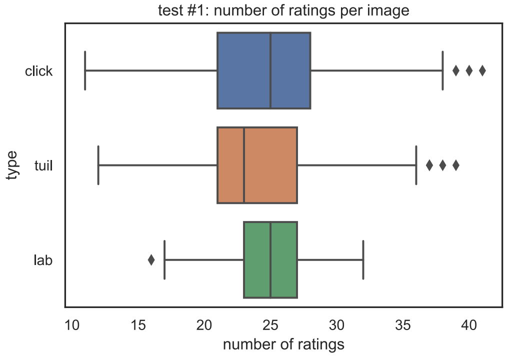


FIGURE 18. Number of ratings per image for `click`, `tuil`, and `lab` in case of test #1.

For example, overall 148 participants took part and provided 24924 ratings for test #1 in case of the `tuil` instance, see Table 6. Similarly, in the case of the `click` recruitment, 139 participants took part and gave 26240 ratings. Overall, the `tuil` instance has more users, however lesser overall ratings, due to the fact that not all participants completed rating all the images. In general, filtering of participants has been performed. Here, participants who rated everything with 1 or rated less than 50 images have been excluded. Thus there are participants who have not rated all images in the case of the `tuil` setting, however, 50 images may be enough for participants to get an understanding of the task. Overall, the `tuil` is also considered only for comparison and has not been done in the test #2, and the participants from the clickworker platform never had incomplete runs. The `lab` test had a total of 34 participants, which provided an overall of 26400 ratings.

Because only a subset of the overall images is shown and rated by individual participants in all three test variants, we analyzed how often an image has been rated in the respective test instance. The results of this analysis are summarized in Figure 18. Overall for all three variants, the number of ratings per image is in a similar range, as shown in the boxplots. For example, the median for `click` and `lab`

[23] https://www.clickworker.com/

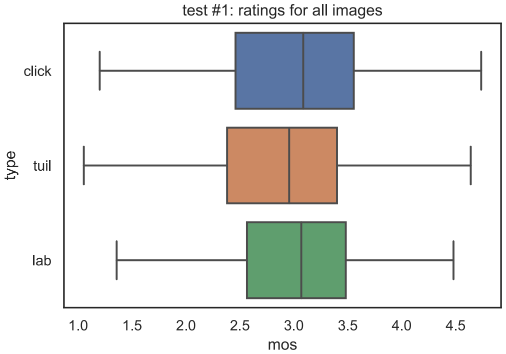


FIGURE 19. Rating distribution for mean appeal ratings ("MOS" – mean opinion scores) for `click`, `tuil`, and `lab` in case of test #1.

TABLE 7. Correlation values and RMSE for the comparison of the different test variants of test #1.

Comparison	Pearson	Kendall	Spearman	RMSE
`click` vs. `lab`	0.845	0.634	0.823	0.385
`tuil` vs. `lab`	0.886	0.689	0.868	0.357
`click` vs. `tuil`	0.872	0.677	0.862	0.379

is 25 and for `tuil` is 23. The number of ratings for `click` is in the range from 11 to 41, for `lab` from 16 to 32, and `tuil` 12 to 39. These ranges are similar to the crowd tests described by Göring et al. in [34] in the context of visual quality evaluation. Furthermore, the number of ratings for each individual image is similar to state-of-the-art datasets, which have been also conducted in crowd setups.

In addition to the number of ratings per image, we also analyzed the overall distribution of mean scores ("MOS" – mean opinion scores), as visualized in Figure 19. All three tests show a similar general range of ratings, and also the median and other quantiles are similar to each other. Therefore, considering the rating distribution only, the three test variants seem to be similar.

To verify this similarity, and also to check the reliability of each test, other approaches can be used. In Figure 20 we performed a Standard deviation of Opinion Scores (SOS) analysis [79] with the equation $SOS(x)^2 = a(-x^2+6\cdot x-5)$. The SOS analysis is a method to check for the reliability of a test. The a values found by curve-fitting are the following: $a_{click} = 0.297$, $a_{tuil} = 0.297$, and $a_{lab} = 0.336$. In [80], an a value of 0.27 is reported by Siahaan et al. for image appeal, and our estimated values are in a similar order of magnitude.

In Figure 21 and Table 7, a pairwise comparison of all three different types for test #1 is summarized. Overall, it can be stated that the results of all three tests are highly correlated with each other, and are therefore similar. We further added the number of ratings per image in the scatterplots for `click` and `tuil`, to check whether there is an influence on the overall correlation, which cannot be observed. However, correlation values for the repetition of lab tests in several labs for quality assessment are usually in a higher range, as it is, e.g., shown by Pinson and Wolf [81], where the Pearson correlation values are ranging from 0.902 to 0.935 for such inter-lab comparisons. Here, the more subjective view on image appeal for each participant, as also indicated by the SOS analysis, may play an important role.

TABLE 8. Correlation values of the different test variants of test #1 with the AVA and AADB subsets, sorted by Pearson Correlation Coefficient per subset.

Subset	Type	Pearson	Kendall	Spearman
AVA	`tuil`	0.811	0.610	0.812
AVA	`click`	0.777	0.577	0.772
AVA	`lab`	0.756	0.540	0.746
AADB	`tuil`	0.635	0.490	0.670
AADB	`lab`	0.615	0.477	0.650
AADB	`click`	0.614	0.477	0.651
KonIQ-10k	`tuil`	0.549	0.373	0.530
KonIQ-10k	`lab`	0.511	0.328	0.467
KonIQ-10k	`click`	0.463	0.295	0.425

The **AVT-ImageAppeal-Dataset** includes appeal or quality ratings for some subsets. In Table 8, a comparison of the subsets for AVA, AADB, and KonIQ-10k considering different correlation coefficients for the already included appeal and quality values is shown. We did not calculate RMSE values, because the respective rating schemes of the datasets are different in comparison to our used 1-5 ACR rating. It can be seen that the best matching test #1 instance is `tuil` for the three datasets. However, overall the correlation values have a lower range than, e.g., the comparison of the individual instances had (>0.84 for Pearson Correlation Coefficient). Only between AVA and the `tuil` variant, the Pearson Correlation is similarly high to our inter-test evaluation. The values for KonIQ-10k are the lowest, however, it must be considered that KonIQ-10k targets image quality, which is only slightly related to appeal. The AADB shows medium correlation values, here the reason could be that the database targets not only image appeal, but also other attributes [11], such as lighting, color, and symmetry, and each image has only been rated by 5 different participants, and using Amazon Mechanical Turk. These differences may explain the lower correlation with our dataset, which targets only image appeal in the subjective evaluation.

C. CROWD AND LAB TEST FOR TEST #2

Test #2 consists of 359 images which are shown twice to the participants so that 718 stimuli are to be rated by participants. Two settings have been considered for test #2, namely `click` and `lab`. We did not conduct test #2 with the `tuil` setting, because it has been shown that the clickworker results in test #1 are highly similar to the lab setting and it is simpler to recruit clickworker. Further, also the `tuil` setting is highly similar to `click` in test #1.

For example, as can be seen from Table 6, for test #2 in the case of the `click` instance, overall 108 participants took part and provided 20938 ratings. For the `lab` variant of test #2, overall 24 subjects participated and 16514 ratings have been collected. The `lab` test had a duration of 45 min, and each participant rated all stimuli, in contrast to the `click`

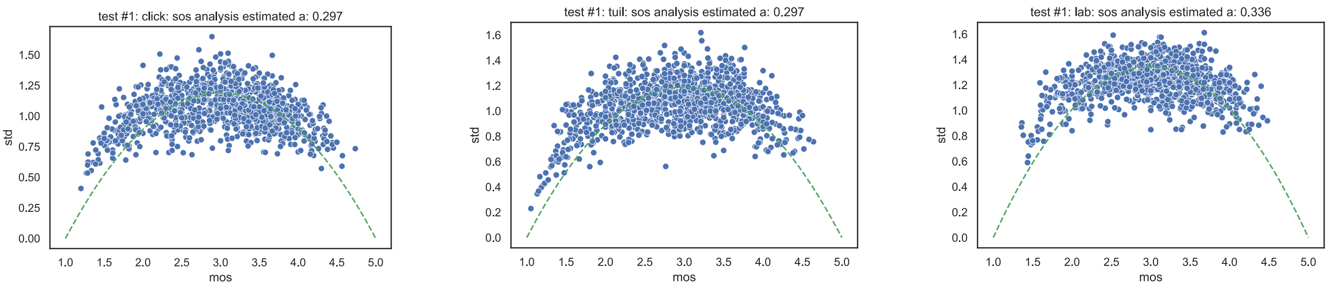


FIGURE 20. Standard deviation of Opinion Scores (SOS) analysis for `click`, `tuil`, and `lab` in case of test #1.

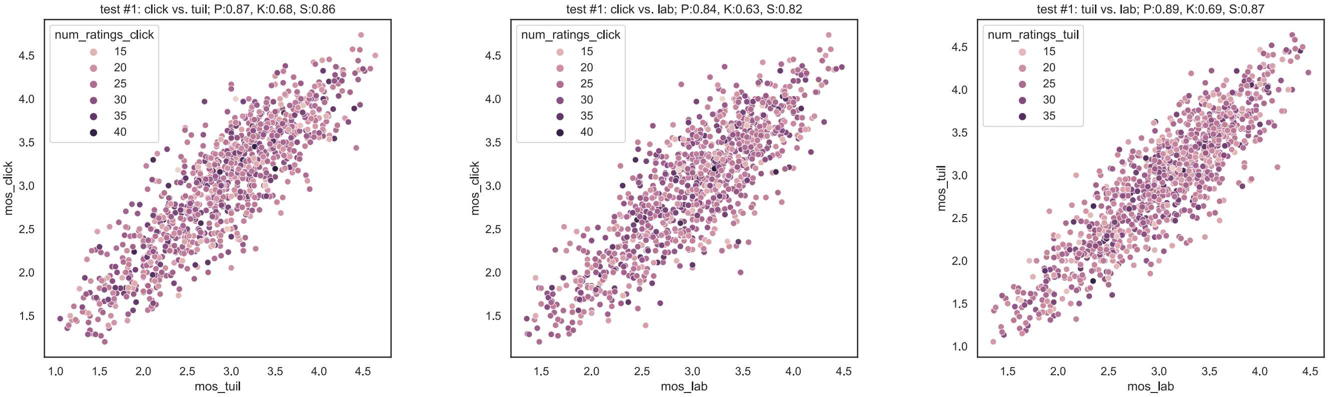


FIGURE 21. Pairwise comparison of mean appeal ratings (mos) for `click`, `tuil`, and `lab` in case of test #1.

test, where each participant rated 200 randomly selected stimuli. Therefore for `lab` all stimuli have 24 ratings each. No outliers have been detected. For the `click` instance of test #2, the stimuli have at least 14, at most 47 ratings, with 29 ratings as the median. As we did for test #1, we also performed an SOS analysis for test #2. We calculated the values $a_{click} = 0.298$ and $a_{lab} = 0.331$ for test #2. Both values are similar to test #1 ($a_{click} = 0.297$, $a_{tuil} = 0.297$, and $a_{lab} = 0.336$).

For the `lab` test setting of test #2, the mean appeal ratings are in the range from $[1.43, 4.39]$, with a median value of 3.13. The `click` variant has a similar range for the ratings of $[1.5, 4.5]$, with a similar median value of 3.14.

The test #2 targets the influence of presenting also likes and views on the image appeal rating. To analyze this, in Figure 22 scatterplots for the comparison of real likes and fake likes are shown for both test scenarios. Overall, it can be seen that in both cases the ratings for real and fake likes are highly correlated, with, e.g., Pearson Correlation values of 0.95, which indicates that there is no difference. Furthermore, we performed statistical tests (paired t-test) which also showed that there is no statistical difference between showing fake and real likes in both variants.

In addition, in Table 9 comparisons of test #2 to test #1 considering the fake and real likes are summarized. In general, in both settings, real and fake likes have a similar correlation to the results of test #1. Furthermore, the Pearson Correlation Coefficient is in the same range, as for the `lab`, `click`, and `tuil` pairwise evaluation in test #1. Thus it can be stated that there is only a minor influence of the like and view numbers for the overall appeal rating. In Table 9, also a comparison of the `lab` and `click` setting of test #2 is included. Here, the values are lower than in the case of test #1, where we use the `click` variant for comparison. However, the lower values may originate from the observation of participants that images are shown twice. This and the duration of the test would need further evaluation, which is not the focus of our analysis, however, it will be addressed in future work.

TABLE 9. Comparison of fake likes and real likes, and `lab` vs. `click` for test #2, in case a comparison with test #1 the values for the `click` type are taken.

Type	Comparison	Pearson	Kendall	Spearman
`click`	real likes vs. test #1	0.87	0.67	0.86
`click`	fake likes vs. test #1	0.87	0.67	0.85
`lab`	real likes vs. test #1	0.83	0.63	0.81
`lab`	fake likes vs. test #1	0.82	0.61	0.80
	`lab` vs. `click`	0.65	0.45	0.63

We further analyzed the connection of likes and views with the appeal ratings of test #2. For this, we selected the ratings for the real likes and compared them with the likes and views. We extended the like and view values also by the ratio of likes and views and logarithmic variants of each of the individual values. To ensure that there are no numerical issues, we added $+1$ in some calculations, which has only

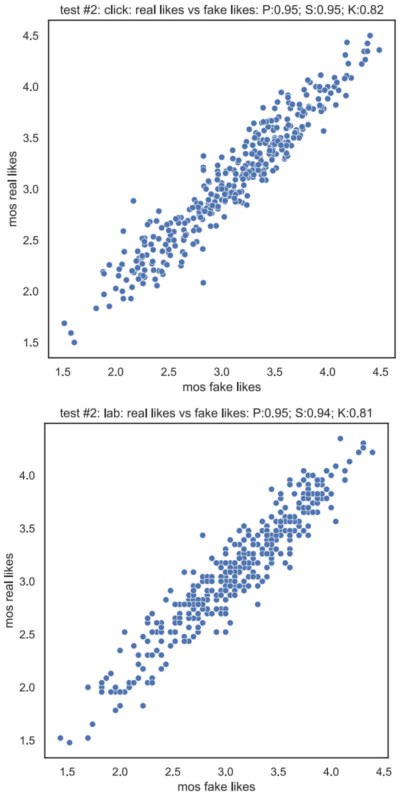


FIGURE 22. Evaluation for test #2 of mean appeal ratings (MOS) for `click`, `lab`, and considering real and fake likes.

TABLE 10. Likes, views, and derived values compared to mean appeal rating (MOS) for test #2; sorted by Pearson and rounded to 3 decimals.

Comparison to MOS	Pearson	Kendall	Spearman
$\log(likes + 1)$	0.432	0.286	0.414
$\log(views + 1)$	0.384	0.233	0.347
$likes$	0.329	0.286	0.414
$\log(likes + 1)/\log(views + 1)$	0.300	0.216	0.317
$views$	0.262	0.233	0.347
$likes/views$	-0.001	-0.045	-0.071
$\log(views + 1)/\log(likes + 1)$	-0.230	-0.150	-0.222

a minor impact. In Table 10, the results for this evaluation are summarized. It is visible, that there is a small to medium correlation for $\log(views + 1)$, $\log(likes + 1)$, and *likes*, thus likes and views have only a small correlation with the appeal rating, which could be also explained by social-temporal network effects [82], where likes and views change over time and may not yet be stabilized as we collected them for the dataset. These conclusions may be limited, due to the nature of an online or lab test, which has a certain task for each participant, in contrast to a real-world photo-sharing platform where no task is given.

V. PREDICTION MODELS

In the following, we will evaluate state-of-the-art models for image appeal, introduce new or derived features extracted from image information, and present our own newly developed prediction models for image appeal. Our developed models are based on traditional machine learning models, such as random forest regression and classification, or use transfer learning with pre-trained deep neural networks. In all experiments, we use the `click` results from test #1 as mean appeal scores.

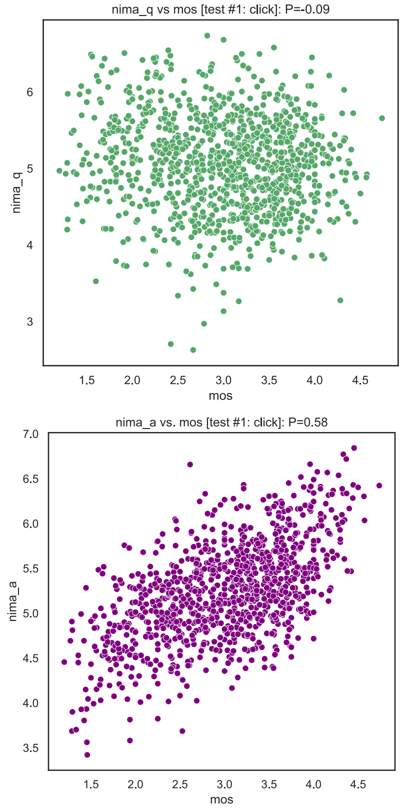


FIGURE 23. NIMA predictions for quality (*nima_q*) and appeal (*nima_a*) for the AVT-ImageAppeal-Dataset.

A. EVALUATION OF STATE OF THE ART MODELS

The NIMA model [5], [15] is open source and can be used for image appeal and quality prediction. The NIMA model is based on transfer learning and uses a pre-trained deep neural network that has been fine-tuned for the corresponding tasks. We calculated the appeal prediction, which is referred to as *nima_a*, and quality prediction (*nima_q*) using the provided models with no modifications. Furthermore, we estimated the NIQE scores [64] for quality prediction. The results for both NIMA predictions (*nima_a* and $nima_q$) are visualized in Figure 23, and all results are listed in Table 11. In addition, we calculated image popularity (*popularity*) using the model provided by Ding et al. [83]. This open-source popularity model is a deep learning-based model, which is trained using transfer learning and photos from Instagram.

The NIMA appeal predictions (*nima_a*) are the best considering all three correlation coefficients, however, there is still only a medium correlation. NIMA quality (*nima_q*) has nearly no correlation to the mean appeal scores, which is because there are no quality degradations within the

TABLE 11. NIMA appeal, quality predictions, image popularity, and NIQE for the AVT-ImageAppeal-Dataset; sorted by Pearson and rounded to 3 decimals.

Model	Pearson	Kendall	Spearman
nima_a	0.584	0.388	0.550
popularity	0.383	0.239	0.349
nima_q	-0.092	-0.057	-0.088
niqe	-0.171	-0.125	-0.186

TABLE 12. Additional features for image appeal prediction.

Feature name	Source	Idea/Meaning/based on
blur	[43]	blurriness
fft	[43]	resolution feature
mean_dominant_color		average of dominant color
deepgaze_connected_components	[71]	saliency maps of DeepGaze II
deepgaze_mean_saliency	[71]	saliency maps of DeepGaze II
deepgaze_std_saliency	[71]	saliency maps of DeepGaze II
unisal_connected_components	[70]	saliency maps of UNISAL
unisal_mean_saliency	[70]	saliency maps of UNISAL
unisal_std_saliency	[70]	saliency maps of UNISAL
mean_segmentation	[74, 75]	semantic segmentation maps
number_segmentation	[74, 75]	semantic segmentation maps
std_segmentation	[74, 75]	semantic segmentation maps
mean_depth	[74, 75]	semantic segmentation maps
std_depth	[74, 75]	semantic segmentation maps
sentiment_neg	[76]	image sentiment
sentiment_neu	[76]	image sentiment
sentiment_pos	[76]	image sentiment
rule_of_thirds	[9]	photographic rules
simplicity	[9]	photographic rules
nima_a	[15, 5]	NIMA prediction for appeal
nima_q	[15, 5]	NIMA prediction for quality

dataset. NIQE has a small correlation with the appeal ratings. Furthermore, image popularity has a low to medium correlation with appeal ratings. However, we checked to which extent the popularity prediction and *nima_a* are matching and there is a medium correlation of both values (Pearson Correlation Coefficients of ≈ 0.51).

B. FEATURES

In Table 5, we already introduced some features to characterize the dataset, namely blur strength, colorfulness, contrast, CPBD, niqe, noise, saturation, SI, and tone. These features are extended by the features listed in Table 12.

For all features, we calculated Pearson, Spearman, and Kendall correlations considering the mean appeal rating. Because these three correlation coefficients are showing similar results, we focus on Pearson Correlation Coefficient. In Figure 24, the individual Pearson Correlation values for all features are summarized. Similarly, as for the state-of-the-art model prediction, the *nima_a* values correlate as best with the MOS. This is followed by CPBD, fft, and sentiment_pos, and also considering absolute numbers by mean_depth, niqe, and sentiment_neg.

Overall, it can be seen that there is no feature that has a strong correlation to the appeal ratings. Most of the features do only have a small or weak correlation.

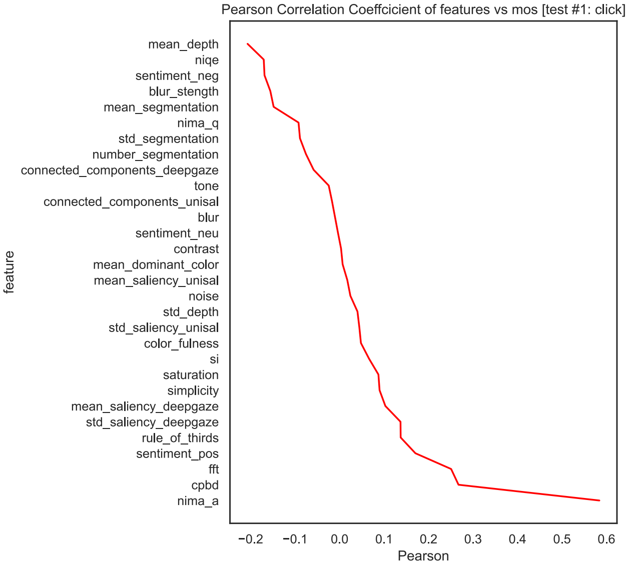


FIGURE 24. Pearson Correlations for features in comparison with mos; sorted by Pearson.

In addition to the mentioned features, we further checked other features, such as the ones shared and used in [84] by Zakrewsky et al. for item popularity prediction using images. Zakrewsky et al. use image appeal and image features, however only minimal correlations can be observed with the appeal ratings. And some of the features are already partially covered semantically in our feature set, e.g., there are blur, rule of thirds, and simplicity features.

C. MACHINE LEARNING BASED MODELS

At first, we focused on traditional machine learning models such as e.g, random forest models, as they have been successfully used for quality prediction [43], [85]. We use 75% of the images from the **AVT-ImageAppeal-Dataset** for training and 25% for validation. For the training, we use auto-sklearn [86], [87], which is an automated machine-learning framework. In general, auto-sklearn trains several machine-learning models and creates an ensemble of models for the final prediction. We trained a regression model for appeal prediction. Auto-sklearn selected the following machine learning models to be in the final ensemble, namely, gaussian_process extra_trees, gradient_boosting (2 variants), and ard_regression.

The prediction results for the 25% validation data are shown in Figure 25.

We further evaluated the impact of each feature individually, and have similar results, as compared to Figure 24, where, e.g., the *nima_a* feature has the highest correlation and also the highest impact for the model prediction. In comparison to the pure NIMA model, the auto-sklearn ensemble has a slightly better performance (e.g. 0.697 Pearson vs. 0.58).

In addition to the mentioned features, we further used a pre-trained DNN (VGG19 [88]) from Keras [89] to extract

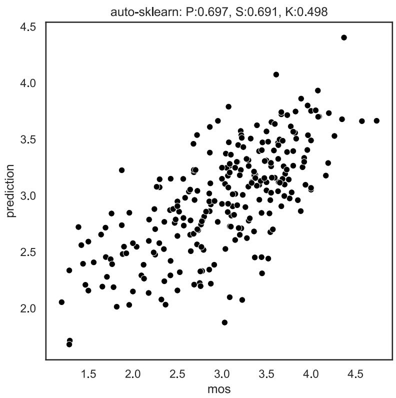


FIGURE 25. Auto-sklearn model for appeal prediction.

deep neural network (DNN) features. We removed the last application-oriented layer from the VGG19 network, and use the DNN as a feature extractor similar to [90] and [91]. We performed three experiments, one using only the deep neural network features (DNNf), another one using only the signal features (features), and one which combines the signal features with the extracted deep neural network features (DNNf + features). For the evaluation, we used random forest regression (scikit-learn [92], 100 trees, and default parameters) and the same 75%-25% split for training and validation. Auto-sklearn [86], [87] was not feasible due to memory restrictions with the DNN features. In Table 13 the results are summarized. The best-performing model is the DNNf+features variant with a Pearson Correlation Coefficient of approximately 0.76. The features or DNN features alone (features, or DNNf) also have a good performance considering the Pearson value, however, the combination indicates a gain. To analyze the gain of each individual signal feature, we performed a leave-one-out experiment, a summary of the Pearson values is shown in Figure 26. In this setting, we use all DNNf+features and remove individual signal features, and for each of the analyzed features, a new Random Forest Regression model is trained and validated with the same validation split. Overall, this experiment showed that for the prediction, the top-3 features (which have the highest drop in performance when left out for the model) are *nima_a*, fft, and sentiment_neu. The least impact on the performance had si, std_depth, and std_segmentation the reason for this could be that some aspects are already covered by the deep neural network features or other variants of the signal features. However, overall, the leave-one-out experiments have a similar Pearson value to using the full set of features, namely the DNNf+features setting. Also, the performance drop of individual features is low in the leave-one-out experiment, considering that the Pearson correlation values are in the range of 0.75 to 0.758.

D. DEEP LEARNING BASED MODELS

To evaluate specific trained deep neural network models, we use transfer learning [93], [94]. We trained several models, namely VGG19 [88], VGG16 [88], Xception [95], DenseNet121 [96], DenseNet201 [96], EfficientNetV2L [97], MobileNetV2 [98], MobileNetV3Large [98], InceptionV3 [99], and ResNet50 [100]. More models could have been considered, however, the overall idea is to check how well transfer learning of deep neural networks can be used for the appeal prediction. We used 50 epochs for training, because results indicated that more epochs improve the performance only minimally. Furthermore, we used a 75%-25% train-validation split. As image augmentation, we used vertical flip, brightness range (0.2, 0.8), and a zoom range of 0.1, assuming that such augmentations have only a minimal effect on the appeal rating. For further processing, the appeal ratings have been [0, 1]-normalized. Each of the DNNs is used without the last layer and extended by a global average pooling, a dense layer (1024 values, and ReLu activation), and another dense layer with a sigmoid activation and one output value for the final prediction. The input layer for all DNNs is (224, 224, 3), thus the input images are rescaled and have three channels with a resolution of 224 × 224 pixels. For the training, we used the Adam Optimizer (with default parameters of Keras) with the mean squared error as loss.

We trained in total 10 DNNs and performed a validation with the 25% of unknown data from the **AVT-ImageAppeal-Dataset**, which is similar to the previously performed validations. The results of the validation are summarized in

TABLE 13. Evaluation for DNN as feature extractor and signal features.

Feature set	Pearson	Kendall	Spearman
DNNf+features	0.756	0.555	0.748
features	0.673	0.459	0.645
DNNf	0.658	0.474	0.666

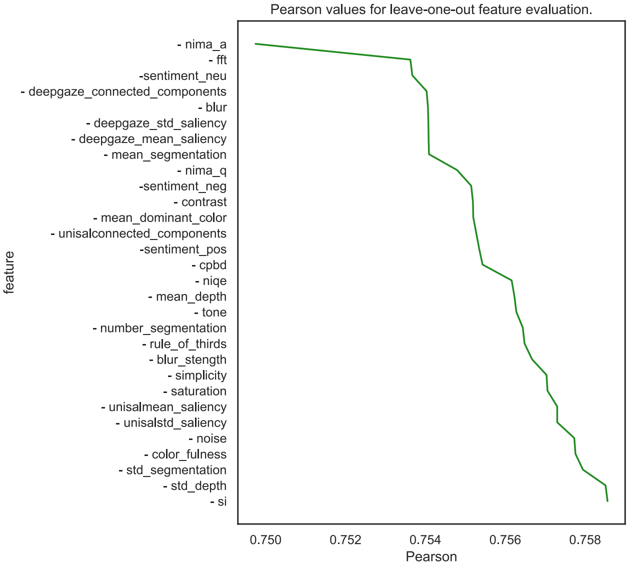


FIGURE 26. Pearson Correlations for leave-one-out feature evaluation, DNNf+features is used as full feature set and individual signal features are removed; sorted by Pearson.

TABLE 14. Validation results for transfer learning considering several DNNs; values are rounded to three decimals.

Model	Pearson	Kendall	Spearman
MobileNetV2	0.636	0.440	0.630
DenseNet121	0.616	0.425	0.607
Xception	0.562	0.377	0.538
VGG16	0.433	0.283	0.422
VGG19	0.384	0.270	0.404
ResNet50	0.127	0.072	0.111
EfficientNetV2L	0.105	0.064	0.095
MobileNetV3Large	0.049	0.020	0.027
DenseNet201	-0.020	0.125	0.188
InceptionV3	-0.084	0.101	0.127

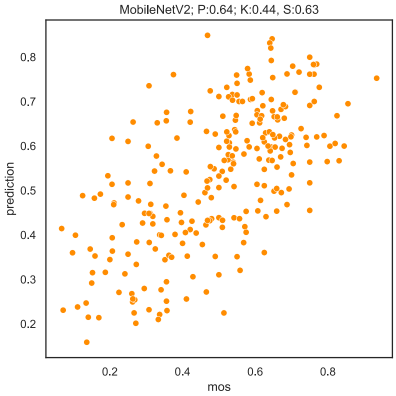


FIGURE 27. MobileNetV2 prediction results for the transfer learning evaluation.

Table 14. MobileNetV2 is the best-performing model, followed by DenseNet121 and Xception. The results are similar to the NIMA (*nima_a*) model [5], [15] (which, unmodified, has a Pearson value of 0.58), wherein also MobileNet is used. The worst-performing models are DenseNet201 and InceptionV3, here some additional layers or re-training of some of the model-specific layers could improve the performance.

The results for MobileNetV2 are shown in detail in Figure 27. The overall prediction range matches the [0, 1]-normalized appeal ratings.

It is important to mention that the considered deep neural network models are trained for image classification in the context of the ImageNet competition [49]. Therefore, for appeal prediction, some more re-training of more layers may be required. For example, the worst performing model, namely InceptionV3, with the current setup has $2,099,201$ trainable parameters. In case we extend the trainable layers by the last 15 layers of the network, which results in $2,494,081$ trainable parameters, we can achieve a better overall prediction performance. For example, we got a Pearson value of 0.655 for the validation data. However, the more layers we add to the training, the more images would be required so that the model does not overfit, which cannot be fully ensured. For this reason, the shown experiments are just a proof of concept, and it can be seen that deep neural networks are well-suitable for the prediction of image appeal.

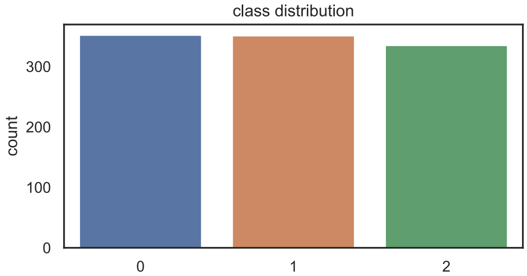


FIGURE 28. Distribution of appeal classes (three classes) using the described conversion.

TABLE 15. Results for the classification evaluation (three classes, $n = 3$); values are rounded to three decimals.

Model	Accuracy	F1	Precision	Recall
RFC_3	0.604	0.597	0.597	0.597
GBC_3	0.554	0.550	0.550	0.550
SVC_3	0.442	0.427	0.427	0.427

However, it is also visible, that the prediction performance of the deep learning models is similar to the signal based models which are shown in the previous experiments.

E. IMAGE APPEAL AS A CLASSIFICATION PROBLEM

In addition to the mean appeal ratings, it is also possible to handle the image appeal prediction as a classification problem, e.g., with 3 different discrete appeal classes (low, medium, high appeal). For the training of the classification models, we use the DNNf+features and define appeal classes based on the majority of votes (m) for a specific discrete value of the used 1-5 rating scheme. Afterwards, we threshold this value m, if $m \leqq 2$ then $class = 0$ (low), if $m = 3$ then $class = 1$ (medium), otherwise $class = 2$ (high). The threshold has been selected to ensure a uniform distribution across the different classes.

In Figure 28, a distribution plot for the three appeal classes is shown.

Because in the regression evaluation, the random forest regression showed good results, for the classification we also selected a random forest classifier (RFC), along with a support vector classifier (SVC) and a gradient boosting classifier (GBC). For the classification with n classes, we use as notation RFC_n, GBC_n, and SVC_n. The used parameters are default values for scikit-learn [92], thus in the case of RFC and GBC 100 trees have been used. Similarly, as compared to the regression evaluation, we use a 75%-25% train-validation split of the **AVT-ImageAppeal-Dataset**.

In Table 15, the results for the classification are summarized. The best-performing model considering accuracy is the RFC_3 followed by the GBC_3 model. The SVC_3 model is the worst. The same ranking holds for the other calculated classification metrics. The values are comparable to state-of-the-art models, however, a bit lower, which is also due to the 25% split, whereas in state-of-the-art usually 10-fold cross-validation has been performed. To visualize the results of the

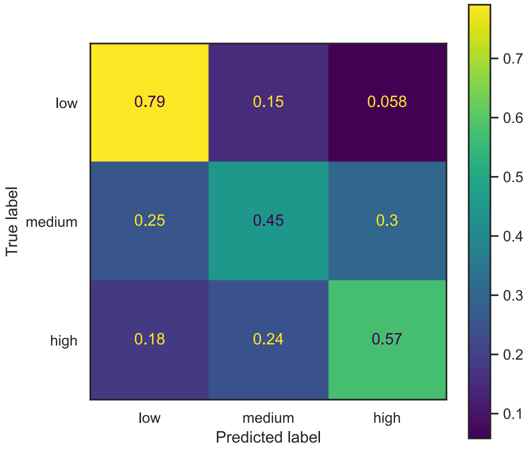


FIGURE 29. Confusion matrix for the RFC$_3$ model (three classes).

TABLE 16. Results for the classification evaluation (two classes, $n = 2$); values are rounded to three decimals.

Model	Accuracy	F1	Precision	Recall
GBC$_2$	0.804	0.804	0.804	0.804
RFC$_2$	0.758	0.758	0.758	0.758
SVC$_2$	0.569	0.552	0.552	0.552

best-performing model, we also created a confusion matrix, which is shown in Figure 29.

In Figure 29, it is visible that low appealing images are well recognized. Furthermore, high and medium-appealing images are not well recognized, however, this may also depend on the approach used to create the class labels.

For this reason, we also evaluated a split into two appeal classes (low and high appealing images). The class labels have been defined by the mean appeal rating, in case the rating is below or equal to 3.0, it is assumed to be $class = 0$ (low), otherwise $class = 1$ (high). We checked the distribution of the classes and both classes are equally represented in the dataset. The evaluation uses the same 75%-25% split, and the performance metrics are listed in Table 16

Similar to the three-class predictions, the RFC$_2$ and GBC$_2$ models are the best performing. Here, the GBC$_2$ model has a higher performance, and therefore we also show the confusion matrix in Figure 30.

In the confusion matrix, it can be observed that low- and high-appealing images are well-classified, and only a small amount of images are not correctly classified. Overall the results are comparable with the three-class evaluations.

VI. DISCUSSION

We described the **AVT-ImageAppeal-Dataset**, and characterized it. Using the dataset, we conducted several subjective evaluation tests considering image appeal. Here, it can be stated that the crowd-sourcing paradigm can be used for image appeal evaluation, considering the required adjustments outlined in the paper. Furthermore, we found that there is only a small influence of like- and view statistics shown alongside the image, therefore the users are focused on the pure appeal with their rating. This may be influenced by the given tasks of the test approaches – in both lab and crowd. In a more natural setup, e.g., a real-world photo-sharing platform, such aspects as likes and views or more social aspects may influence the decisions of users in a stronger manner. Considering the recruitment of the participants both lab and crowd tests showed highly similar results, however it should be stated that, e.g., a pure expert panel would may be rate image appeal differently, however this was not the purpose of the conducted tests and could be targeted in future work. Besides appreciating a picture for its content and/or the aesthetic appeal of the image, other reasons may be that users follow each other and/or wish to appreciate their own work via establishing social media connections and interactions. With the described approach, such effects cannot be evaluated and would need further investigation. The **AVT-ImageAppeal-Dataset** consists of images from other state-of-the-art image appeal datasets, and we compared the appeal ratings of our tests with the included ones, where it was shown that there is a good correlation. In addition, it is also shown that the a values of the performed SOS analysis are higher compared to image quality tests, which is also shown in other conducted work. This effect indicates that image appeal rating has a stronger subjective component and hence stronger inter-subject variability than image quality assessment, where the criteria based on which judgments are made are more universal. In addition to the image appeal assessment, we evaluated several machine learning models for image appeal prediction. It is shown that NIMA appeal (*nima_a*) is a good prediction model for our database. However, we describe approaches employing other models, which use signal and deep learning features and show a better performance. Furthermore, we also trained deep neural networks using transfer learning, and these models perform similarly well as compared to the signal models. Here, for a mode of robust training, a larger dataset would be required. The prediction models handle the appeal prediction

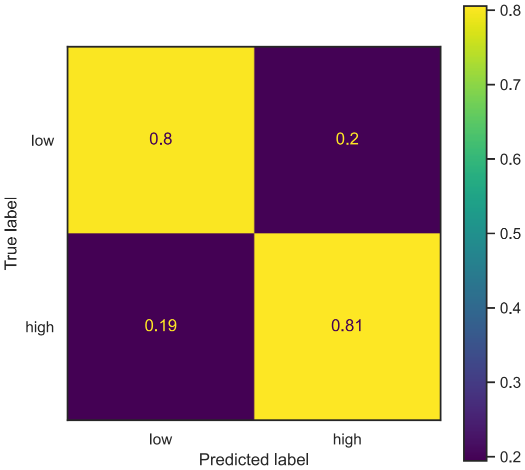


FIGURE 30. Confusion matrix for the GBC$_2$ model (two classes).

problem as a regression. As a complementary approach, we describe models that handle the problem as a classification with 3 and 2 classes. The class labels are derived from the raw ratings from the conducted subjective tests. For both classification cases, the used machine learning models are shown to perform well, e.g., random forest and gradient boosting models are the best. The used features are based on signal and deep learning models because such a hybrid setup showed the best results in the regression case.

VII. CONCLUSION

Image appeal is a crucial part of photography and is also important for photo-sharing platforms, or for users to decide whether an image is of high appeal or not. To evaluate the image appeal of photos, we introduced the **AVT-ImageAppeal-Dataset**, which is a newly constructed image appeal dataset including high-resolution images from several real-world sources. This dataset also includes images from other state-of-the-art datasets partly comprising complementary metadata such as the number of likes and views. We compare crowd and lab tests, where we also included like and view statistics along with the photo. We found that the results of lab and crowd tests are highly similar and that the likes and views have only a minor impact on the appeal rating. Furthermore, we describe various features, which we extracted from the images, and use them to develop machine learning models for image appeal prediction as classification and regression. These models are also using deep neural network features, and we also train deep neural networks. Our developed models show similar and better performance than state-of-the-art models. The models, code, images, and ratings are publicly accessible for reproducibility in the context of open science. Overall, it can be seen that image appeal prediction is still a challenging task, and that additional factors may have an influence on the ratings. For example, models could also include a more per-user oriented view, which would be analyzed in future work.

ACKNOWLEDGMENT

The authors would like to thank the participants of the conducted subjective tests. Furthermore, they want to thank the "AG Wissenschaftliches Rechnen" of the Technische Universität Ilmenau for providing computing resources.

REFERENCES

[1] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *Brit. J. Psychol.*, vol. 95, no. 4, pp. 489–508, Nov. 2004.

[2] P. Lebreton, A. Raake, and M. Barkowsky, "Evaluation of aesthetic appeal with regard of user's knowledge," *Electron. Imag.*, vol. 28, no. 16, pp. 1–6, Feb. 2016.

[3] P. Lebreton, A. Raake, and M. Barkowsky, "Studying user agreement on aesthetic appeal ratings and its relation with technical knowledge," in *Proc. 8th Int. Conf. Quality Multimedia Exp. (QoMEX)*, Jun. 2016, pp. 1–6.

[4] S. Göring, K. Brand, and A. Raake, "Extended features using machine learning techniques for photo liking prediction," in *Proc. 10th Int. Conf. Quality Multimedia Exp. (QoMEX)*, Sardinia, Italy, May 2018, pp. 1–6. [Online]. Available: https://www.researchgate.net/publication/325479182_Extended_Features_using_Machine_Learning_Techniques_for_Photo_Liking_Prediction

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

[6] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, Nov. 2015.

[7] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "Predicting image aesthetics with deep learning," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.* Cham, Switzerland: Springer, 2016, pp. 117–125.

[8] H. Takimoto, F. Omori, and A. Kanagawa, "Image aesthetics assessment based on multi-stream CNN architecture and saliency features," *Appl. Artif. Intell.*, vol. 35, no. 1, pp. 25–40, Jan. 2021.

[9] 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. [Online]. Available: https://ieeexplore.ieee.org/document/9733554

[10] 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.

[11] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 662–679.

[12] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10k: Towards an ecologically valid and large-scale IQA database," 2018, *arXiv:1803.08489*.

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

[14] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[15] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[16] N. Zangwill, "Aesthetic judgment," in *The Stanford Encyclopedia Philosophy*, E. N. Zalta, Ed. Stanford, CA, USA: Metaphysics Research Lab, Stanford Univ., Fall 2014.

[17] I. Kant, *Critique of Judgment*, Trans. Meredith, Ed. New York, NY, USA: Oxford Univ. Press, 1790.

[18] C. Sartwell, "Beauty," in *The Stanford Encyclopedia Philosophy*, E. N. Zalta, Ed. Stanford, CA, USA: Metaphysics Research Lab, Stanford Univ., Winter 2017.

[19] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation* (Signals and Communication Technology). Berlin, Germany: Springer, 2005. [Online]. Available: https://books.google.de/books?id=8OM7CmWQYAcC

[20] K. Brunnström et al., "Qualinet white paper on definitions of quality of experience," White Paper, 2013.

[21] A. Raake and S. Egger, "Quality and quality of experience," in *Quality of Experience*, S. Möller and A. Raake, Eds. Cham, Switzerland: Springer, 2014, doi: 10.1007/978-3-319-02681-7_2.

[22] R. Schifanella, M. Redi, and L. Aiello, "An image is worth more than a thousand favorites: Surfacing the hidden beauty of Flickr pictures," 2015, *arXiv:1505.03358*.

[23] L. Hsieh, W. H. Hsu, and H. Wang, "Investigating and predicting social and visual image interestingness on social media by crowdsourcing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4309–4313.

[24] B. Krages. (2012). *Photography: The Art Composition*. Allworth. [Online]. Available: https://books.google.de/books?id=bmqCDwAAQBAJ

[25] J. T. Smith, *Remarks Rural Scenery; With Twenty Etchings Cottages, From Nature; Some Observat. Precepts Relative to Pictoresque. By John Thomas Smith*. London, U.K.: Engraver Antiquities London. Jun. 1797.

[26] I. Biederman and E. Vessel, "Perceptual pleasure and the brain a novel theory explains why the brain craves information and seeks it through the senses," *Amer. Scientist*, vol. 94, no. 3, pp. 247–253, 2006.

[27] P. Jonas, *Photographic Composition Simplified* (A Modern Photoguide). New York, NY, USA: Amphoto Books, 1976. [Online]. Available: https://books.google.de/books?id=OSBEAQAAIAAJ

[28] O. Le Meur and Z. Liu, "Saccadic model of eye movements for free-viewing condition," *Vis. Res.*, vol. 116, pp. 152–164, Nov. 2015.

[29] S. Ma, Y. Fan, and C. W. Chen, "Finding your spot: A photography suggestion system for placing human in the scene," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 556–560.

[30] L. Mai, H. Le, Y. Niu, Y.-C. Lai, and F. Liu, "Detecting rule of simplicity from photos," in *Proc. 20th ACM Int. Conf. Multimedia*, Oct. 2012, pp. 1149–1152.

[31] L. Mai, H. Le, Y. Niu, and F. Liu, "Rule of thirds detection from photograph," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2011, pp. 91–96.

[32] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[33] M. Maleš, A. Hedi, and M. Grgic, "Compositional rule of thirds detection," in *Proc. ELMAR*, Sep. 2012, pp. 41–44.

[34] S. Göring, R. R. R. Rao, and A. Raake, "Quality assessment of higher resolution images and videos with remote testing," *Quality User Exper.*, vol. 8, no. 1, p. 2, 2023.

[35] R. R. R. Rao, S. Göring, and A. Raake, "Towards high resolution video quality assessment in the crowd," in *Proc. 13th Int. Conf. Quality Multimedia Exp. (QoMEX)*, Jun. 2021, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9465425

[36] S. Göring, 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. [Online]. Available: https://ieeexplore.ieee.org/document/9733561

[37] J. Redi, E. Siahaan, P. Korshunov, J. Habigt, and T. Hossfeld, "When the crowd challenges the lab: Lessons learnt from subjective studies on image aesthetic appeal," in *Proc. 4th Int. Workshop Crowdsourcing Multimedia*, Oct. 2015, pp. 33–38.

[38] J. Redi and I. Povoa, "Crowdsourcing for rating image aesthetic appeal: Better a paid or a volunteer crowd?" in *Proc. Int. ACM Workshop Crowdsourcing Multimedia*, 2014, pp. 25–30.

[39] R. Schifanella, M. Redi, and L. M. Aiello, "An image is worth more than a thousand favorites: Surfacing the hidden beauty of Flickr pictures," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 9, no. 1, 2015, pp. 397–406.

[40] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang, "Image aesthetics assessment using deep Chatterjee's machine," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 941–948.

[41] J. A. Redi and I. Povoa, "The role of visual attention in the aesthetic appeal of consumer images: A preliminary study," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2013, pp. 1–6.

[42] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, Jan. 2016.

[43] S. Göring, 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

[44] Y. Tan, P. Tang, Y. Zhou, W. Luo, Y. Kang, and G. Li, "Photograph aesthetical evaluation and classification with deep convolutional neural networks," *Neurocomputing*, vol. 228, pp. 165–175, Mar. 2017.

[45] L. Abdenebaoui, B. Meyer, A. Bruns, and S. Boll, "UNNA: A unified neural network for aesthetic assessment," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, Sep. 2018, pp. 1–6.

[46] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. CVPR*, Jun. 2011, pp. 1657–1664.

[47] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1784–1791.

[48] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 83–92.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[50] H. Lee, K. Hong, H. Kang, and S. Lee, "Photo aesthetics analysis via DCNN feature encoding," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1921–1932, Aug. 2017.

[51] 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 Exp. (QoE) Vis. Multimedia Appl.*, 2020, pp. 47–53.

[52] 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.

[53] S. A. Amirshahi, G. U. Hayn-Leichsenring, J. Denzler, and C. Redies, "Color: A crucial factor for aesthetic quality assessment in a subjective dataset of paintings," 2016, *arXiv:1609.05583*.

[54] Y. Bo, J. Yu, and K. Zhang, "Computational aesthetics and applications," *Vis. Comput. Ind., Biomed., Art*, vol. 1, no. 1, pp. 1–19, Dec. 2018.

[55] M. Xu, J.-X. Zhong, Y. Ren, S. Liu, and G. Li, "Context-aware attention network for predicting image aesthetic subjectivity," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 798–806.

[56] J. Hou, S. Yang, and W. Lin, "Object-level attention for aesthetic rating distribution prediction," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 816–824.

[57] K. Ghosal and A. Smolic, "Image aesthetics assessment using graph attention network," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 3160–3167.

[58] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2006, pp. 288–301.

[59] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S.-F. Chang, "Image popularity prediction in social media using sentiment and context features," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 907–910.

[60] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 271–280.

[61] 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*, vol. 6492, 2007, pp. 196–206.

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

[63] 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.

[64] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.

[65] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, Sep. 1994.

[66] T. O. Aydin, 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.

[67] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document P. 910, Int. Telecommun. Union, Geneva, Switzerland, 2008.

[68] 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.

[69] M. Kümmerer, Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. *Mit/tübingen Saliency Benchmark*. Accessed: Jul. 7, 2023. [Online]. Available: https://saliency.tuebingen.ai/

[70] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 419–435.

[71] M. Kümmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge, "Understanding low- and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4799–4808.

[72] C. Fiorio and J. Gustedt, "Two linear time union-find strategies for image processing," *Theor. Comput. Sci.*, vol. 154, no. 2, pp. 165–181, Feb. 1996.

[73] K. Wu, E. Otoo, and A. Shoshani, "Optimizing connected component labeling algorithms," in *Proc. SPIE*, Apr. 2005, pp. 1965–1976.

[74] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.

[75] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," 2021, *arXiv:2103.13413*.

[76] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 308–317.

[77] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on visual sentiment analysis," *IET Image Process.*, vol. 14, no. 8, pp. 1440–1456, Jun. 2020.

[78] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT.500-13, Int. Telecommun. Union, Geneva, Switzerland, 2014.

[79] 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.

[80] 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.

[81] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *Proc. SPIE*, vol. 5150, pp. 573–582, Jun. 2003.

[82] B. Wu, T. Mei, W.-H. Cheng, and Y. Zhang, "Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.

[83] K. Ding, K. Ma, and S. Wang, "Intrinsic image popularity assessment," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1979–1987.

[84] S. Zakrewsky, K. Aryafar, and A. Shokoufandeh, "Item popularity prediction in e-commerce using image quality feature vectors," 2016, *arXiv:1605.03663*.

[85] R. R. R. Rao, S. Göring, and A. Raake, "AVQBits—Adaptive video quality model based on bitstream information for various video applications," *IEEE Access*, vol. 10, pp. 80321–80351, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9846967

[86] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 2962–2970.

[87] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: Hands-free AutoML via meta-learning," 2020, *arXiv:2007.04074*.

[88] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[89] F. Chollet et al. (2015). *Keras*. Accessed: Jul. 7, 2023. [Online]. Available: https://keras.io and https://keras.io/getting_started/faq/#how-should-i-cite-keras

[90] S. Göring, J. Skowronek, and A. Raake, "DeViQ—A deep no reference video quality model," in *Human Vision and Electronic Imaging 2018, Burlingame, CA, USA, 28 January 2018–2 February 2018*, B. E. Rogowitz, T. N. Pappas, and H. de Ridder, Eds. Ingenta, 2018, pp. 1–6, doi: 10.2352/ISSN.2470-1173.2018.14.HVEI-518.

[91] S. Göring 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 pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8611703

[92] 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, pp. 2825–2830, Oct. 2011.

[93] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on Machine Learning Applications and Trends: Algorithms*. Hershey, PA, USA: IGI global, 2010, pp. 242–264.

[94] *Transfer Learning & Fine-Tuning*. Accessed: Jul. 7, 2023. [Online]. Available: https://keras.io/guides/transfer_learning/

[95] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 1251–1258.

[96] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[97] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[98] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[99] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

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

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

ALEXANDER RAAKE (Member, IEEE) received the Dr.-Ing. degree 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 held a Senior Researcher, Assistant, and later Associate Professor positions at TU Berlin's An-Institut T-Labs, 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 at LIMSI-CNRS, Orsay, France. His research interests include audiovisual and multimedia technology, speech, audio, and video signals, human audiovisual perception, and quality of experience. Since 1999, he has been involved with ITU-T Study Group 12's Standardization work on QoS and QoE assessment methods. He is a member of the Acoustical Society of America, the AES, VDE/ITG, and DEGA.