InnerEye: A Tale on Images Filtered Using Instagram Filters - How Do We Interact with them and How Can We Automatically Identify the Extent of Filtering?

Gazi Abdur Rakib[1], Rudaiba Adnin[1(✉)], Shekh Ahammed Adnan Bashir[1], Chashi Mahiul Islam[1], Abir Mohammad Turza[1], Saad Manzur[1], Monowar Anjum Rashik[1], Abdus Salam Azad[1,2], Tusher Chakraborty[3], Sydur Rahaman[1], Muhammad Rayhan Shikder[1], Syed Ishtiaque Ahmed[4], and A. B. M. Alim Al Islam[1]

[1] Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
1505032.ra@ugrad.cse.buet.ac.bd
[2] University of California, Berkeley, USA
[3] Microsoft, Redmond, USA
[4] University of Toronto, Toronto, Canada

Abstract. Even though digitally filtered images are taking over the Internet for their aesthetic appeal, general people often feel betrayed if they are dealt with filtered images. Our study, comprising a series of structured surveys over images filtered using Instagram filters, reveals that different people perceive the filtered images differently. However, people have a common need for an automated tool to help them distinguish between original and filtered images. Accordingly, we develop an automated tool named *'InnerEye'*, which is capable of identifying how far an image is filtered or not. InnerEye utilizes a novel analytical design of a Neural Network Model that learns from a diverse set of images filtered using Instagram filters. Rigorous objective and subjective evaluations confirm the efficacy of InnerEye in identifying the extent of filtering in the images.

Keywords: Images · Filters · Survey · Deep Learning

1 Introduction

With the rise of social media, the world has witnessed an influx of images shared on the Internet. As of January 2022, an average of more than 100 million images are uploaded, and more than 4.2 billion likes are accumulated each day on Instagram alone [33], and these values are ever-rising. Part of the reason for having the images to be predominating on social media is associated with the availability

S. Longfei and P. Bodhi (Eds.): MobiQuitous 2022, LNICST 492, pp. 494–514, 2023.
https://doi.org/10.1007/978-3-031-34776-4_26

of mobile cameras and mobile applications specializing in image editing. Digital image processing is categorized into two classes such as pixel-level editing and parametric editing [36]. Using both of the classes of image editing, mobile 'apps' provide different preset ways to edit images reflecting the choice of a user. These preset ways are called 'Filters'. People use filters to achieve a stylized appearance of images without having any prior knowledge about digital image processing [27]. The filters are considered a type of parametric image editing system, which changes the outlook of an image while preserving the context and original objective content. As the notion of image filtering is extensively used nowadays and people are now substantially experiencing outcomes of image filtering, there arise several aspects related to real-life interactions with filtered images. The aspects include how people perceive the filtered images, how far they are confident in their ability to distinguish whether a filter has been applied to an image or not, and so on. Several related studies have focused on users' ability to find out image manipulation [19,32], however, very few studies focused specifically on the perspectives of users while interacting with filtered images. Therefore, this study focuses specifically on the perceptions of users with filtered images.

In this research, we performed a chronological series of online surveys to uncover the different aspects of people's perceptions of filtered images. To overcome the demographic variations that unavoidably occur in our chronological series of surveys and to verify whether the results of our surveys hold value over time, we performed a consolidated survey ensuring consistent demography of the participants. Analyzing the results of our surveys, we discovered people's experiences with filtered images. The results demonstrate significant variability in public opinions on categorization and quantification over original and filtered images. Irrespective of this variability, analysis of the survey results implies the need for an automated tool for detecting and measuring the extent of applying a filter to an image. Accordingly, we designed and developed a new web tool-based solution that employed a Neural Network Model specially crafted for predicting whether an input image is edited with filters or not. Here, our consideration of filtering mostly subsumes applying Instagram filters, as Instagram is the largest photo-sharing social media network that has a huge collection of built-in filters [37,38]. We present our solution to the public and sought out public opinion through yet another survey to evaluate the efficacy of our solution and its acceptability to the users. As a result, we encountered the following set of research questions in this study.

- *RQ1:* Is there any prevalence of filtered images shared on social media? Can users perceive the extent of applying filters over images with bare eyes?
- *RQ2:* Will general users appreciate the help of an automated tool in identifying the extent of applying filters over an image, and if so, why?
- *RQ3:* What are the components of an automated tool for identifying whether an image is filtered or not? Do people appreciate using such a tool?

Working with the above-mentioned research questions, we make the following set of contributions to this paper.

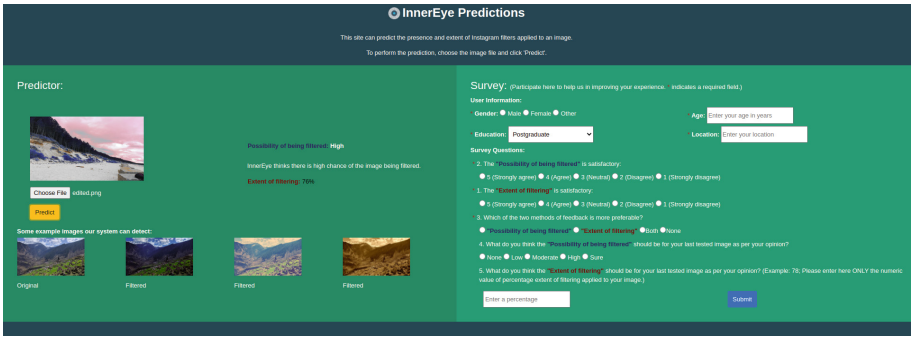


Fig. 1. The user interface of InnerEye

– We conducted a comprehensive series of chronological surveys that present an in-depth look into users' perspectives on interacting with filtered images, their preferences in interacting with specific types of filtered images, and their ability to distinguish between an original image and a filtered image. The findings of the surveys reveal salient aspects of users' experience with filtered images. This will help HCI researchers to better understand users' interaction with filtered images.
– Being inspired by the findings of our surveys, we developed a novel solution called *'InnerEye'* (Fig. 1) for detecting the extent of applying filters over an image. InnerEye is a web tool implemented with a Neural Network Model specializing in predicting the originality of an image in terms of being edited with image filters. Our implemented tool, InnerEye, informs the HCI community of a method regarding how a filtered image detection tool can be designed and implemented for users.
– We conducted a user evaluation with a survey presented within the InnerEye tool. This survey uncovers users' feedback on the performance of InnerEye as well as their appreciation of identifying the extent of image filtering with an automated tool. The user evaluation will assist HCI researchers to understand users' preferences for filtered image detection tools and facilitate the development of solutions in the realm of image filtering in novel ways.

2 Related Work

In this section, we provide an in-depth discussion of the prior research work pertaining to our study.

2.1 Social and News Media, and Edited Images

The upsurge in social media usage has led us to a world where images are shared across the globe and edited images play a significant role in people's interaction with social media. A study [26] takes an in-depth look at how social media

shaped people's perception of images and how it has affected their day-to-day life. Another study [23] investigates the factors that influence Facebook users' intentions to post digitally altered self-images. However, manipulated images lead to significant personal or societal impact [6]. Gupta et al., show how edited images are prevalent in prominent social media sites like Twitter, especially during disastrous events such as Hurricane Sandy, misleading the general users to believe in fake news, and how automated techniques can be used to detect fake images with adequate accuracy [14]. Boididou et al., explored the challenges of building an automatic detection framework for fake multimedia content in social media posts [5]. Another study [30] talked about a broad case study of Hurricane Sandy and the Instagram images shared over Twitter. They explored how news propagates among citizens in a post-newspaper era where uncontrolled access to images shapes people's viewpoints toward various groundbreaking incidents differently. Additionally, exposure to manipulated images directly led to lower confidence related to physical appearance [20], especially, girls with higher social comparison tendencies [20].

In addition, some research studies [7,32] investigate how well people can identify image tampering. Farid et al. [7] focus on people's ability to identify photo manipulations by creating a series of computer-generated scenes consisting of basic geometrical shapes. Furthermore, an exploratory focus group study [19] was done to discover how individuals react and evaluate the authenticity of images that accompany online stories in communications channels.

However, we know little about how people assess filtered images specifically and their judgments about the authenticity of such images. Therefore, we investigate the experiences of people with filtered images.

2.2 Image Forgery and Fake Image Detection

The field of image forgery detection has become of paramount importance in recent years. Jang et al. [17], theorized an apparatus capable of detecting fake images by learning the background information from input images and building a learned background and comparing it to the present background to detect discrepancies. The performance comparison between various novel passive techniques concludes that they need automation, i.e., removal of human intervention in output analysis [4]. In digital images, sometimes image portions are edited to cover up a major feature of the original image, which contributes to deceiving people. Fridrich et al. [9], studied detecting what is known as copy-move forgery, which involved copying a segment of an image to cover up another part. Several other studies also worked on dealing with the same problem. Some studies focus on detecting re-sampling [24] and interpolation [10] in digital images. Another recent work [25] addresses the problem of tampering localization. Another study [34] uses contextual cues to detect image forgery. With the recent development of generative neural network architectures, the world has begun witnessing the prevalence of DeepFake, which describes the phenomena of replacing one person's face with another's realistically in images or videos. Nguyen et al. [31] talk about the existing DeepFakes technology and how to detect DeepFakes in-depth.

Rössler et al. [35] showed us in their dataset collection work how excessive the growth of generated face images has been in recent years. Agarwal et al. [1] show us why detection cannot keep up with the generation of DeepFake and what is hindering the current development of DeepFake detection. Xuan et al. [42], show how preprocessing both real and fake training data can lead to a better generalization across DeepFakes generated using various technologies. In a more recent work [28] by Marra et al., it has been shown that Neural Network Models can achieve higher efficiency in detecting machine-generated images with variable properties. Tariq et al. [39], show how preprocessing and employing ensemble methods can achieve far greater success in GAN-generated fake images than edited images. Frank et al. [8], show us how analyzing images in the frequency domain could lead to a groundbreaking change in the sector of GAN-generated face images. Belkasoft's [3] forgery detection module is a comprehensive payware that can perform automatic detection of alteration, modification, and forgery, providing a confidence percentage score on the originality of an image. The tool employs error level analysis, clone detection, quantization table analysis, double compression artifacts analysis, double quantization effect analysis, and foreign artifact detection.

Due to the extreme prevalence of edited and fabricated images online, there has been a lot of development in the sector of building user-friendly software that can perform qualitative and forensic analysis on digital images. Jonas Wagner [40] built the tool Forensically that can perform a multitude of levels of forensic operation on digital images, such as digital magnification with three types of enhancement: histogram equalization, and auto contrast by channel; similar region detection through operations regarding minimal similarity, minimal detail, minimal cluster size, block-size, and maximal image size; error level analysis through comparison of the original and re-compressed image for manipulation detection; Level sweep for histogram analysis; principal component analysis; metadata analysis; and string extraction for hidden ASCII content in the image. Dr. Neal Krawetz [21] built a similar forensic analysis tool called FotoForensics that can perform metadata analysis, error level analysis, hidden content analysis, and service information extraction. Zampoglou et al. [43], built a tool for an image originality verification system, featuring a multitude of image tampering detection algorithms, along with metadata analysis, geolocation tagging, and reverse image searching through Google. In a study [13], we see how analyzing image histograms can reveal the statistical differences of the Hue and Saturation channel, and designing a histogram-based and feature-based detection system can lead to a decent performance in detecting fake colorization of images.

Analyzing the previously annotated research studies, we can observe that a limited number of studies focused on filtered images. Therefore, we uncover the perspectives of users with filtered images and a filtered image detection tool.

3 Methodology of Our Study

We recruited participants for a chronological series of online surveys to gather public opinion on the prevalence of image sharing and filtered images over social media as per their experiences and practices. Analyzing the survey result, we performed the task of filtered image detection. We built a web tool *'InnerEye'* by analyzing the alteration of the inherent color distribution of an image with a custom Neural Network model. Deploying our automated web tool, we performed a user evaluation to find out whether people appreciate using InnerEye which assists them by detecting filtered images online and quantifying the extent of filtering applied to images. Figure 2 shows our methodology (Fig. 3).

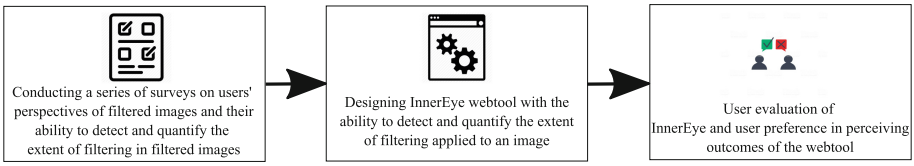


Fig. 2. An overview of the methodology of our work

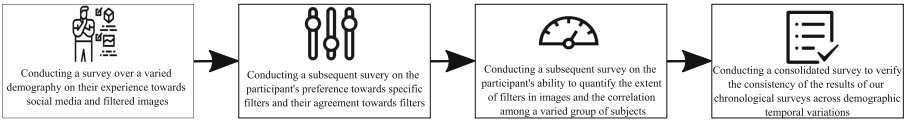


Fig. 3. A series of surveys conducted in our study in a chronological manner

4 Perception About Interacting with Alterations of Shared Filtered Images

During the rudimentary stages of our research, we designed three consecutive online surveys to understand how people perceive image sharing on social media, and what is the general experience while interacting with shared filtered images. All the questionnaires of the surveys were designed to attain the perspectives of users on how they interact with filtered images on social media. The participants of the surveys were recruited through open online communication via publicly available emails and public social media groups. This study was approved by the Ethics Committee of the institution of the corresponding author.

4.1 Surveying Users About Their Experiences with Interacting Through Images on Social Media

We designed this questionnaire to analyze the social media usage of the participants and their experiences while interacting with filtered images shared on

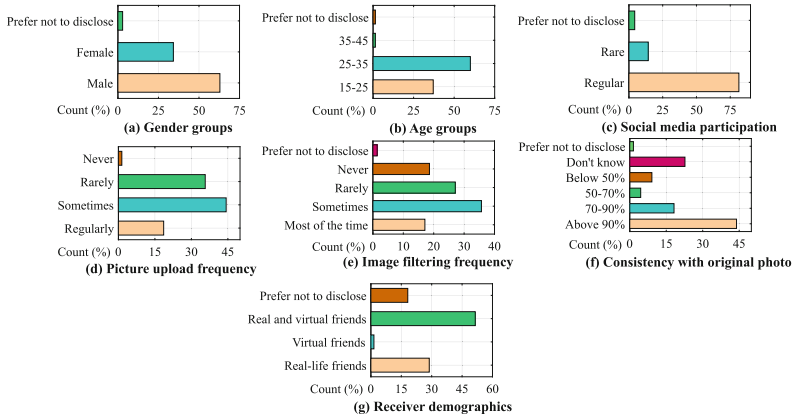


Fig. 4. Different graphs showing the percentage measures of demographic variables and different qualitative measures of social media activity and image-related information

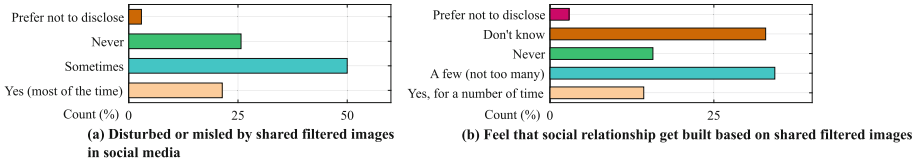
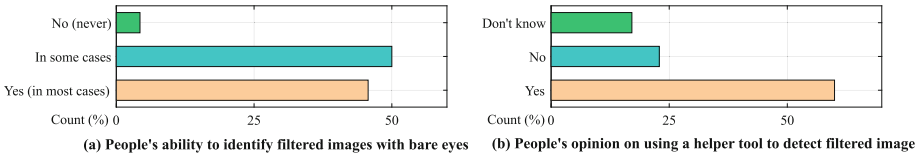


Fig. 5. Different graphs showing the general opinion about different aspects of relationships in social media

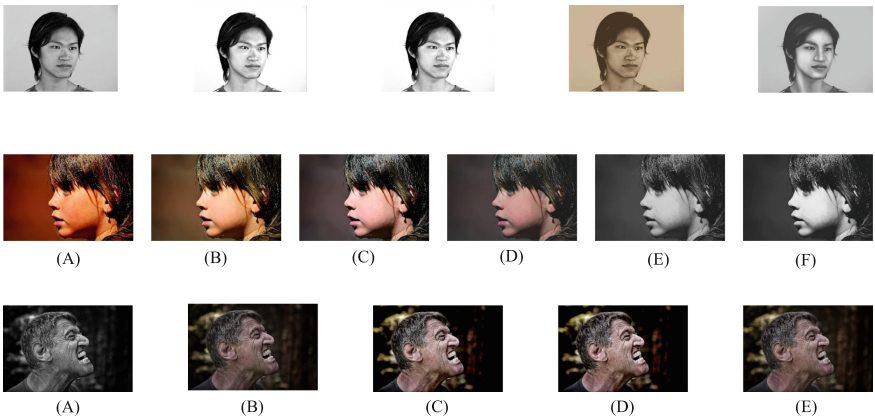
social media. From the results in Fig. 4, we observe that participants from different demographics use social media regularly and a lot of these social media users interact with filtered images, in the forms of uploading, and filtering. A considerable amount of participants filter images to be shared online and they have different tastes in filtering frequency and deciding the extent to which the filters will alter the images. The receivers of these filtered images are not convergent towards a specific category, which emphasizes more on the varied taste of general users when it comes to filtered images. Additionally, from the results in Fig. 5, we notice that many feel disturbed by filtered images shared on social media. Some of them further think that social relationships frequently get built based on filtered images.

In addition, from the results in Fig. 6, we observe that majority of the participants claim to be able to determine whether an image has been filtered or not with their own eyes, but at the same time, a vast majority will appreciate using an automated tool to detect filtering in images. These imply that users are not confident in their ability to detect image alterations by applying filters while interacting with images. Therefore, an objective opinion about the originality of images will assist them in their interaction with filtered images.



(a) People's ability to identify filtered images with bare eyes (b) People's opinion on using a helper tool to detect filtered image

Fig. 6. Different graphs showing the confidence of people's ability to distinguish filtered images and opinion on automated tools to help in differentiating between filtered and original images



(c) Five alternate versions of a picture of an angry person

Fig. 7. Different images used to quantify user preference for image filtering

4.2 Surveying Social Media Users About Their Preferences Towards Interacting with Types of Image Filtering

We designed this questionnaire with inspiration from the previous survey, to find an objective view of the preferences of the participants regarding filtered images. The first part of the survey collected demographic data. The second part of the survey showed the participants different filtered variants of the same images, shown in Fig. 7, filtered in terms of brightness, contrast, color saturation, and temperature, and asked which variant of effect in each image would they choose most likely when filtering an image. We decided to perform the same test on three images to present a subjective variability to the contents of the images so that we do not see any sort of subjective bias in the preference of the participants.

If we look at the results of this survey (Fig. 8), we observe that participants have varied tastes when it comes to image filtering preferences, across filters and subjective variables. Additionally, there is no universal agreement on what types

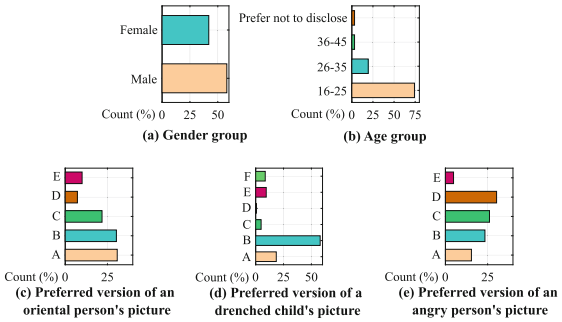


Fig. 8. Different graphs showing the demographic information, and the variability of preference between different filtered versions of three images

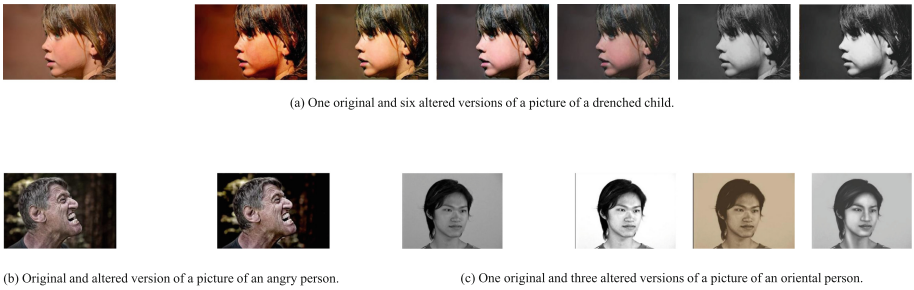


Fig. 9. All ten examples of the original image and their edited counterpart, used in the survey on quantification of image alteration

of filters participants will use. Thus, the user's preference varies when choosing a filter to apply to images. These imply that the task of detecting whether an image has been filtered or not is a vast problem space.

4.3 Surveying Capability of General Users in Quantifying Extent of Image Alteration

We designed this questionnaire to test out whether participants can come to an agreement when quantifying different extents of image filtering and whether they have any objective measurements of the properties of the images when they are considering the extent of filtering. The first section of the questionnaire served to collect demographic information. The second part aimed to judge the participants' ability to come to an agreement while quantifying the extent of filtering of different filtered versions of three images, as shown in Fig. 9. The third part extracted the qualitative information about the specifics of components based on which participants reported their quantification of filtering of the images in

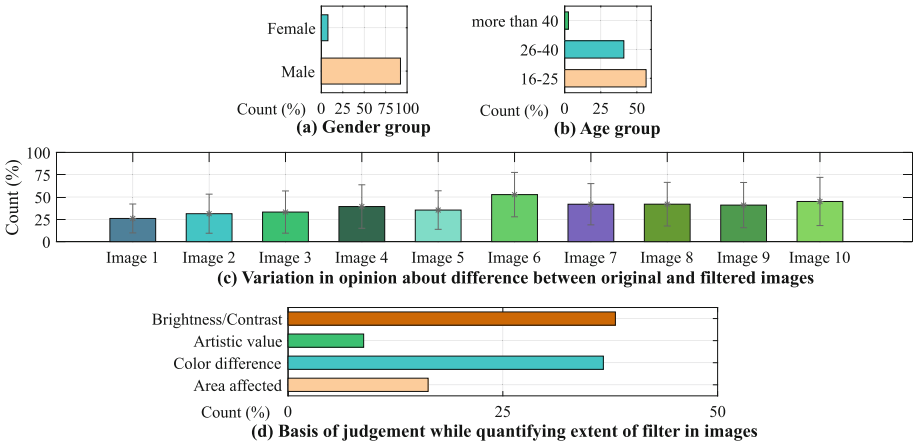


Fig. 10. Five graphs showing the demographic information, one graph showing median and standard deviation of the general perception towards the level of editing of 10 images, and one graph showing the basis of the judgment of quantifying image editing level

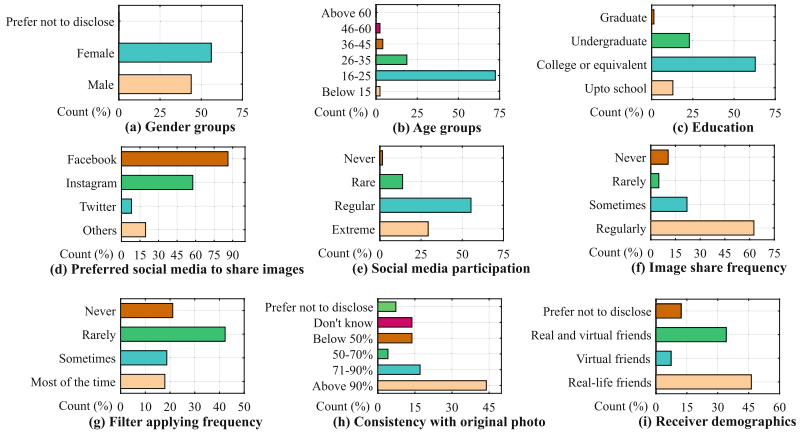


Fig. 11. Different graphs showing the percentage measures of demographic variables and different qualitative measures of social media activity and image-related information

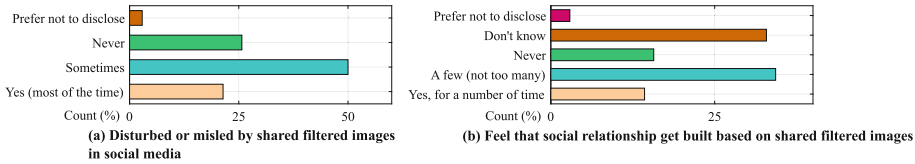


Fig. 12. Different graphs showing the general opinion about different aspects of relationships in social media

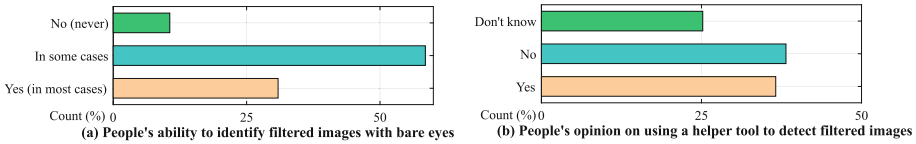


Fig. 13. Different graphs showing the confidence of people's ability to distinguish filtered images and opinion on automated tools to help in differentiating between filtered and original images

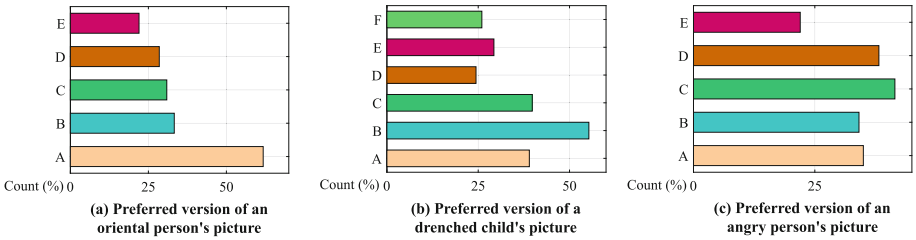


Fig. 14. Different graphs showing the variability of preference between different filtered versions of three images

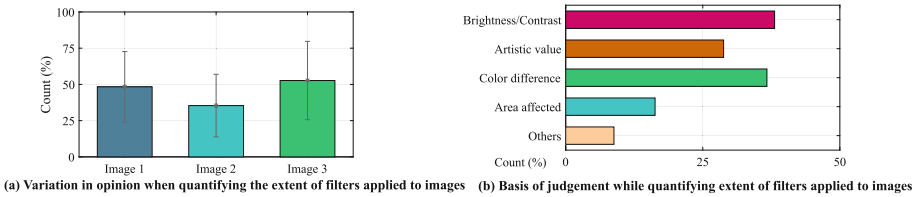


Fig. 15. One graph showing median and standard deviation of the general perception towards the level of editing of three images, and one graph showing the basis of the judgment of quantifying image editing level

the second part. We decided to present ten different variations of three different images to overcome the subjective bias due to regularity or pattern in questions that would arise if we presented the same number of variations of each image.

The participants had to answer some numeric value that quantified the level of image alteration in each instance. For all such items, the responses almost uniformly varied in the range of 10% to 90%. They were allowed to choose multiple answers if they used more than one criterion to compare two images. Therefore, when calculating, we used a weighted percentage manner. Each person's choice was multiplied by the inverse of the number of choices they made. Delving into the results of the questionnaire (Fig. 10), we observe that users perceive the same filtered versions of images to a varying degree and it is hard for them to achieve agreement while quantifying the extent of filters applied to images. Even when identifying image components that were changed during filtering the images in our questionnaire, there is a vast divergence among participants. Therefore, we

can hypothesize that users' opinions toward a tool that can detect a wide variety of filters are divergent similar to divergent opinions towards filtered images.

4.4 Consolidated Survey to Verify Our Findings Across Demographic and Temporal Variations

Due to conducting an open online series of chronological surveys designed respectively, we noticed a demographic variation in the three questionnaires. We understand that open online surveys have their limitations when it comes to capturing a wide range of demography. Additionally, we sequentially collected the results, which raises some questions about the validity of the surveys, as participants observed shifts in their online presence between the time they answered the three questionnaires. To verify the results of our series of surveys, we designed a consolidated one and we recruited participants in a manner that preserves the demographic proportion of the participants.

In this consolidated survey, we brought in a few changes compared to the previous ones while keeping all the questions similar to the previous three consecutive surveys. We collected more detailed demographic information, and we further presented the participants with three filtered images, compared to ten in the previous survey, as we noticed the consolidated survey would be rather time-consuming if we presented the participants with ten filtered images, which might discourage participants to complete the survey. After analyzing the results (Fig. 11, 12, 13, 14, 15) of our consolidated survey, we can see the results are comparable even in the presence of demographic and temporal variation.

5 Implementation of an Automated Solution to Detect Extent of Image Alteration: InnerEye

We developed a web tool, *'InnerEye'* to enable users to interact with filtered images by detecting filtered images automatically. InnerEye makes use of an image classifier to distinguish between unfiltered and filtered images and provides a confidence value (chance of being filtered) of the input image. We deployed the tool online and made it open for public use.

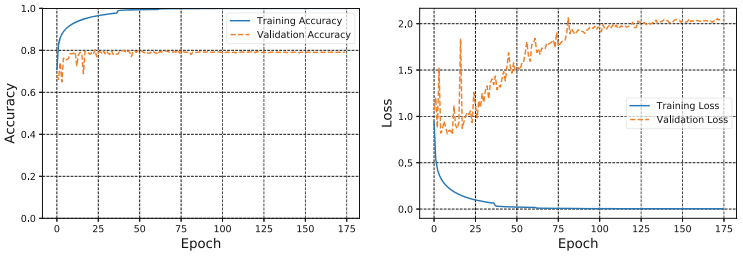
Table 1. Dataset division into train, validation, and test set for updated dataset

Dataset Description			
	Sampled Images	Filtered Images	Total Images
Train	7001	203029	210030
Validation	251	7279	7530
Test	501	14529	15030

Fig. 16. An original and three filtered versions of a single image that our tool can successfully distinguish between

5.1 Developing a Neural Network Model for Detecting Filtered Images

We implemented a classifier to classify input images into two types: Filtered and Unfiltered. The challenge was to recognize how the appearance of colors of the objects present in a filtered image is different from those in its unfiltered counterpart. Further, while constructing such a classifier, we had to consider the fact that, under different illumination conditions, e.g., early morning, mid-afternoon, and evening, the color appearance of the objects in an image changes. However, the real color of the objects in an image does not change over the day. Therefore, if we can have the classifier learn to recognize the invariant properties each color posses under all-natural illumination conditions, our classifier learns the features present in the unfiltered images. We refer to the invariant properties each color posses under all-natural illumination conditions to be the style of an image. The classifier can classify filtered and unfiltered images following the



(a) Classifier accuracy plot (b) Classifier loss plot

Fig. 17. Classification performance of the sequential Neural Network-based classifier on the larger, more varied dataset

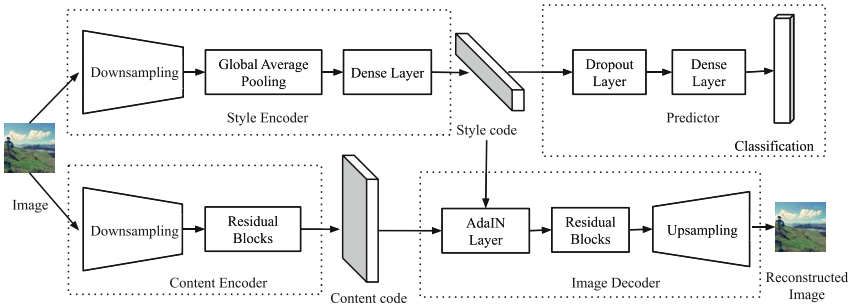


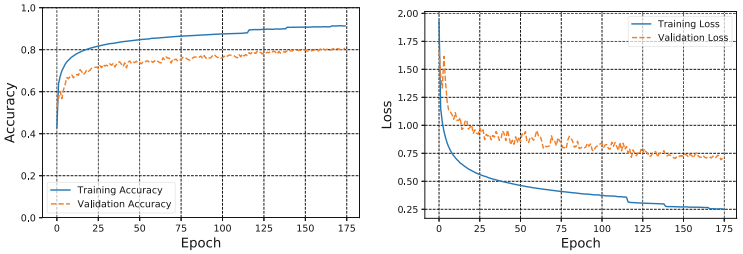
Fig. 18. Architecture of the analytical classifier

prescribed knowledge. Therefore, we are using supervised Learning to build the classifier. We compiled a dataset of Filtered and Unfiltered images to train the InnerEye classifier from the Google Landmarks dataset [12]. We included the image filters *Gotham*, *Lomo*, and *Sepia*, *Aden*, *Brannan*, *Brooklyn*, *Clarendon*, *Earlybird*, *Gingham*, *Hudson*, *Inkwell*, *Kelvin*, *Lark*, *Lofi*, *Maven*, *Mayfair*, *Moon*, *Nashville*, *Perpetua*, *Reyes*, *Rise*, *Slumber*, *Stinson*, *Toaster*, *Valencia*, *Walden*, *Willow*, and *Xpro2*. Description of these filters are given in [29] and we used the Python package [18] to generate the newly filtered images. Our compiled dataset description is given in Table 1.

After training our sequential classifier with the dataset, we can observe a larger case of overfitting due to the high variance in the dataset from our training log in Fig. 17. Thus We looked more closely into the problem of the classification of filtered and unfiltered images. We note that the application of an image filter only alters the color of the objects in an image, however, it does not alter the semantic content. Therefore, if we can have our classifier ignore the semantic content of images as much as possible, it can better learn the style and hence yield a better accuracy on the classes. Moreover, this approach is likely to reduce the variance that used to arise in the previous classifiers due to a large number of different filters. InnerEye classifier consists of a style encoder module, a content encoder module, an image decoder module, and a style predictor module. InnerEye classifier is co-trained for both image classification and image reconstruction tasks. This classifier does not overfit the training data, as evidenced in Fig. 19,

Table 2. Accuracy report on train, validation and test set

Accuracy Report	
	Accuracy
Train	91.39%
Validation	80.32%
Test	96.02%



(a) Classifier accuracy plot (b) Classifier loss plot

Fig. 19. Classification performance of the analytical classifier.

as there is a convergence tendency among the training loss and validation loss curve in the training logs.

In one recent work, Huang et al. described a mechanism, which the authors call MUNIT [15], for image-to-image translation. In image-to-image translation, some aspects of an image, which may be only style or only semantic content, or both, are preserved entirely or to some degree, and the rest are filtered according to those described in a different image. In MUNIT, the authors assumed that an image consists of domain-invariant semantic content code and domain-invariant style code. In the case of InnerEye, we can make the same assumption on the images and can work with the style code only to have the classifier learn better than before. Based on this assumption, we design a classifier, presented in Fig. 18, for InnerEye. We call such a classifier an *analytical classifier* because it analyzes and decomposes the image into content code and style code. The training logs are shown in Fig. 19. The accuracy reports are given in Table 2.

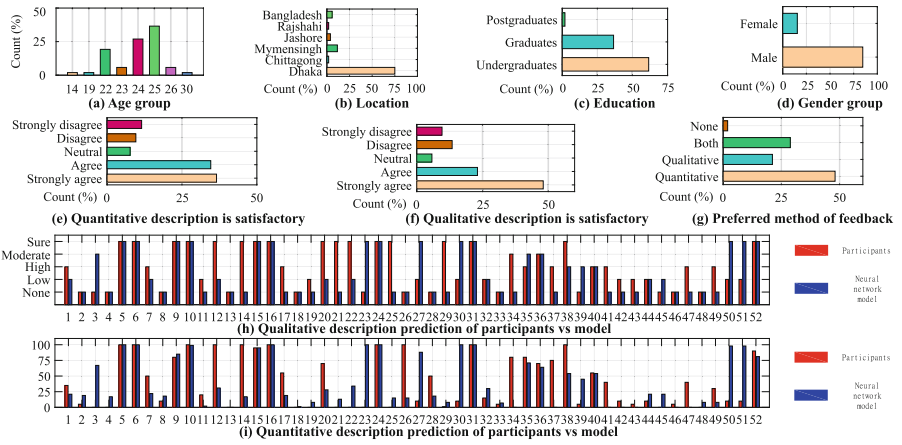


Fig. 20. Four graphs showing the demographic information, three graphs showing the general consensus about chances of being filtered, summary, and the preferred method of feedback, and two graphs showing contrast between the human prediction and the machine one

5.2 Developing a Web Tool to Deploy the Neural Network Model

We deployed our newly designed neural network model in the form of a web tool, InnerEye. The tool consists of two components, the Predictor and the Survey. The predictor serves to give an objective measurement in the form of a quantitative percentage value of the extent of a filter applied to an image and a qualitative message about the possibility of an image being filtered.

The left part of the web tool contains the predictor. It takes an image as an input and outputs two results: a quantitative description, a percentage score of the extent of any filter applied to the input image, and a qualitative description, a brief message about the possibility that the input image is filtered. The qualitative description has five values: None, Low, Moderate, High, and Sure. The message is delivered in the form of a sentence, for example, 'InnerEye thinks there is no chance of the image being filtered' if the output is None. We showed three examples of filtered images our system can detect, as shown in Fig. 16. The web tool contains a survey through which we collected feedback about our tool from our users. The survey starts with a section for collecting demographic information such as gender, age, education, and location. Consecutively, the survey collected user opinions regarding the quantitative and qualitative descriptions, the preference between the qualitative description and quantitative one, and the user-proposed values for quantitative and qualitative descriptions.

6 Evaluation of Performance of Our Solution

We made the web tool online after finishing development and made it public. We distributed the tool among participants ($n = 52$) recruited through open online participation via emailing to publicly available email addresses. We collected the user experiences through the survey integrated into the tool. The survey results can be observed in Fig. 20.

We observe that the general agreement and disagreement towards our solution are nearly equalized, which is undesirable but quite expected. We hypothesized that as the opinions of users towards filtered images are very divergent, their opinions toward a filtered image detection tool will be divergent. We analyzed the survey results on our baseline and verified this hypothesis. The results reflect this hypothesis. We see the general agreement towards the qualitative and quantitative description of our improved solution. We further observe the correlation between the model's prediction and the participant's predictions from the bar charts. Consequently, we can conclude that the participants prefer the results of our tool in a quantitative manner.

7 Discussion

In this section, we answer the key research questions of our study.

7.1 Users and Filtered Images

We identify that users frequently share images on social media and edit those images using filters. However, they have varied tastes when it comes to image filtering preferences. As a result, there is no fixed type of filters that users use. These extend prior work [23] where the intention of users to post edited images was studied.

As an unavoidable consequence, users frequently come across filtered images shared by others. They further perceive that some social relationships are built based on filtered images. Additionally, whilst some users prefer filtered images shared by others, many showed detest to such images probably because they often feel that those images can be misleading. This contributes to prior work [14,20,30] on how edited images can mislead people.

In addition, as found in our study, users feel somewhat confident about their capability of identifying filtered images with their bare eyes. However, that does not stop them from seeking assistance from a technological solution such as a tool to detect filtered images. This proves the dependency of users on technology to support their confidence [22].

Several prior studies [7,32] investigate the capability of users in identifying image manipulations. We extend these prior researches by uncovering that users perceive that they can detect filtered images with bare eyes even though different users quantify the extent of filters used on images in a substantially different manner. This indicates that users cannot uniformly (and perhaps accurately) identify filtered images or the extent of filters used on those images with bare eyes.

7.2 Variability in General Perception Towards Filtered Images

Users are often confident in their ability to distinguish filtered images shared on social media. This thought of users is reflected in our survey results as about 46% of the participants are highly confident and 50% of the participants are somewhat confident. However, while we let the users quantify the extent of image alteration for the same sets of images, we get a completely different story. We noticed that the quantification of the extent of filtering over the same image substantially deviates for different users. This suggests that the opinions of users about a filtered image can highly vary and they do not generally agree (or even somewhat agree) on a single value of quantification for the extent of image filtering.

This result, along with the confidence of users, demonstrates that users perceive filtered images differently, which complies with individuals' cognitive differences on visual images [2,11,30]. Thus, the same image that appeared to be highly filtered to one person may appear to be mostly untouched to another. The variability of users to quantify the extent of filters used on images and their desire for detecting filtered images point out a need for technological solutions enabling the detection of filtered images. In this context, most of the users in our study further expressed the need for an automated tool for detecting filtered

images. Thus, analysis of our survey results implied a common need for an automated tool for the purpose of detecting and measuring the extent of filtering to an image.

7.3 Appreciation of Automated Tools in Detecting Filtered Images

Although users feel that they can distinguish filtered images with their bare eyes, when asked whether they would appreciate an automated tool to detect a filtered image, 60% of the participants answered positively and about 17% of the participants were unsure (Fig. 6). We can further observe a different trend in our consolidated survey - about 37% of the participants were sure and about 25% of the participants were unsure. It sheds light on the fact that with the progression of time, users are perhaps turning indifferent toward filtered images. Accordingly, they are potentially becoming less conscious about the potential harms engendered from disturbing filtered images over social media.

When presented with an automated tool that specializes in detecting whether an image is filtered or not, the results obtained from the user evaluation shed light on some new interesting facts. Here, 37% of the participants strongly agreed with the chances of being filtered, and 35% agreed, compared to 11% disagreement and 12% strong disagreement (Fig. 20). Based on these results, we observe that users appreciate automated tools for detecting filtered images. Therefore, we provide a novel solution to detect filtered images by following a user-centric approach. This extends previous studies [3,21,40,43], where different online tools were proposed to detect edited and fabricated images.

In addition, we designed our tool with the option for both types of prediction, i.e. quantitative prediction based on chances of being filtered, and qualitative prediction based on the summary comment. We put an option in our survey (integrated with our tool) to check which one of the two ways users find preferable while receiving predictions from the tool. While analyzing the collected survey data, we noticed a phenomenon that users generally prefer quantitative assessments over qualitative ones (Fig. 20).

7.4 Limitations of Our Study

The values we obtained from the surveys were self-reported values collected from the participants. Determining the objective viability of these self-reported values is another challenging research topic. Additionally, considering the reality of having computational resource constraints, we used a subset of filters for training our implemented tool from a huge number of existing filters on Instagram. Despite these limitations, the findings of our study will be useful for technology design in the context of image filtering.

8 Avenues for Further Research

This study focuses on image filtering, which is a particular way of performing image editing. Other ways of image editing could be cropping, color adjustment,

removal of objects from images, selective color change, etc., which have not been covered in this study. We are aware of recent research studies [16,41,44], where the authors addressed cropping, cloning, and other types of forgeries. In future work, we plan to address these ways of image editing along with our considered image filtering in a single pipeline to detect clever frauds ever. Moreover, in this study, we built a classifier by decomposing an image into style and content. However, we want to explore how can we implement the decomposition in a better way to achieve more classifier accuracy. Therefore, in the future, we intend to study the construction and the properties of the image filters in-depth and make changes to the Neural Network architecture accordingly to enhance the classifier.

9 Conclusion

Image filtering presents a process of changing the style of the components of an image through various mathematical operations. Our survey results showed how the participants could not provide a unified view on the quantification of the extent of the filter of different images pointing out that users cannot agree on an objective value of the extent of image filtering and how they would appreciate the help of an automated tool in this regard. Inspired by the results of the surveys, we built a custom web tool, InnerEye, using a Neural Network Model that can detect images processed or filtered with popular social media filters. We conducted a user evaluation after deploying the tool online, which demonstrates the efficacy of InnerEye and its acceptance among users. Thus, our study contributes to the field of identifying the extent of image filtering and shows that an automated tool can assist in improving user experience substantially in this domain.

Acknowledgements. The work was conducted at and supported by the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh.

References

1. Agarwal, S., Varshney, L.R.: Limits of deepfake detection: a robust estimation viewpoint (2019)
2. Bandura, A.: Social cognitive theory of personality. In: Handbook of Personality, vol. 2, pp. 154–96 (1999)
3. Belkasoft: Belkasoft forgery detection module (2021). http://reveal-mklab.iti.gr/reveal/
4. Birajdar, G.K., Mankar, V.H.: Digital image forgery detection using passive techniques: a survey. Digit. Investig. **10**(3), 226–245 (2013)
5. Boididou, C., Papadopoulos, S., Kompatsiaris, Y., Schifferes, S., Newman, N.: Challenges of computational verification in social multimedia. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 743–748 (2014)
6. Dearden, L.: The fake refugee images that are being used to distort public opinion on asylum seekers. Independent **16**(9), 15 (2015)

7. Farid, H., Bravo, M.J.: Image forensic analyses that elude the human visual system. In: Media Forensics and Security II, vol. 7541, pp. 52–61. SPIE (2010)
8. Frank, J., Eisenhofer, T., Lea, S., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition (2020)
9. Fridrich, J.A., Soukal, D.B., Lukáš, J.A.: Detection of copy-move forgery in digital images. In: Proceedings of Digital Forensic Research Workshop. Citeseer (2003)
10. Gallagher, A.C.: Detection of linear and cubic interpolation in jpeg compressed images. In: The 2nd Canadian Conference on Computer and Robot Vision (CRV 2005), pp. 65–72. IEEE (2005)
11. Gartus, A., Klemer, N., Leder, H.: The effects of visual context and individual differences on perception and evaluation of modern art and graffiti art. Acta Physiol. (Oxf) **156**, 64–76 (2015)
12. Google: Google landmarks dataset (2020). https://www.kaggle.com/google/google-landmarks-dataset
13. Guo, Y., Cao, X., Zhang, W., Wang, R.: Fake colorized image detection. IEEE Trans. Inf. Forensics Secur. **13**(8), 1932–1944 (2018)
14. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 729–736 (2013)
15. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189 (2018)
16. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting fake news: image splice detection via learned self-consistency. In: Proceedings of the European Conference on Computer Vision, pp. 101–117 (2018)
17. Jang-Hee, Y., Kim, Y., Kyoungho, C., Soonyoung, P., Moon, K.Y.: Method and apparatus for determining fake image (2013). US Patent 8,515,124
18. Kamakura, A.: pilgram 1.1.0 (2019). https://pypi.org/project/pilgram/
19. Kasra, M., Shen, C., O'Brien, J.F.: Seeing is believing: how people fail to identify fake images on the web. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2018)
20. Kleemans, M., Daalmans, S., Carbaat, I., Anschütz, D.: Picture perfect: the direct effect of manipulated instagram photos on body image in adolescent girls. Media Psychol. **21**(1), 93–110 (2018)
21. Krawetz, D.N.: Fotoforensics (2021). http://fotoforensics.com/
22. Lee, Y., Lee, J., Lee, Z.: Social influence on technology acceptance behavior: self-identity theory perspective. ACM SIGMIS Database DATABASE Adv. Inf. Syst. **37**(2–3), 60–75 (2006)
23. Lowe-Calverley, E., Grieve, R.: Self-ie love: predictors of image editing intentions on Facebook. Telematics Inform. **35**(1), 186–194 (2018)
24. Luo, W., Huang, J., Qiu, G.: Robust detection of region-duplication forgery in digital image. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 4, pp. 746–749. IEEE (2006)
25. Maigrot, C., Kijak, E., Claveau, V.: Context-aware forgery localization in social-media images: a feature-based approach evaluation. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 545–549. IEEE (2018)
26. Manovich, L.: Instagram and contemporary image. CUNY, Nova Iorque (2017)
27. Marques, O.: Innovative Technologies in Everyday Life. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45699-7

28. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of GAN-generated fake images over social networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 384–389 (2018)
29. Messieh, N.: How instagram filters work, and can you tell the difference? (2018). https://www.makeuseof.com/tag/instagram-filters-work-can-tell-difference/
30. Murthy, D., Gross, A., McGarry, M.: Visual social media and big data. Interpreting instagram images posted on twitter. Digit. Cult. Soc. **2**(2), 113–134 (2016)
31. Nguyen, T., Nguyen, C.M., Nguyen, T., Nguyen, D., Nahavandi, S.: Deep learning for deepfakes creation and detection: a survey (2019)
32. Nightingale, S.J., Wade, K.A., Watson, D.G.: Can people identify original and manipulated photos of real-world scenes? Cogn. Res. Principles Implications **2**(1), 1–21 (2017). https://doi.org/10.1186/s41235-017-0067-2
33. Omnicore: Instagram statistics (2022). https://www.omnicoreagency.com/instagram-statistics/
34. Papadopoulou, O., Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Web video verification using contextual cues. In: Proceedings of the 2nd International Workshop on Multimedia Forensics and Security, pp. 6–10 (2017)
35. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: a large-scale video dataset for forgery detection in human faces. arXiv (2018)
36. Russotti, P., Anderson, R.: Digital Photography Best Practices and Workflow Handbook: A Guide to Staying Ahead of the Workflow Curve. Taylor & Francis, Milton Park (2010)
37. Statista: Instagram number of daily active instagram stories statistics (2021). https://www.statista.com/statistics/730315/instagram-stories-dau/
38. Statista: Instagram number of monthly active user statistics (2021). https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/
39. Tariq, S., Lee, S., Kim, H., Shin, Y., Woo, S.S.: Detecting both machine and human created fake face images in the wild. In: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, pp. 81–87 (2018)
40. Wagner, J.: Forensically (2021). https://29a.ch/photo-forensics/
41. Wu, Y., AbdAlmageed, W., Natarajan, P.: Mantra-net: manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9543–9552 (2019)
42. Xuan, X., Peng, B., Wang, W., Dong, J.: On the generalization of GAN image forensics. In: Sun, Z., He, R., Feng, J., Shan, S., Guo, Z. (eds.) CCBR 2019. LNCS, vol. 11818, pp. 134–141. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31456-9_15
43. Zampoglou, M.: Reveal - image verification assistant (2021). http://reveal-mklab.iti.gr/reveal/
44. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, vol. 9, pp. 1053–1061 (2018)