Identifying Computer-Generated Portraits: an Eye Tracking Study

Pallavi Raiturkar*, Hany Farid†, Eakta Jain*
*University of Florida
†Dartmouth College

***Abstract*—We conducted two eye tracking studies to understand attention allocation while participants identified portrait images as computer-generated or photographic. In the first study, participants viewed the images for six seconds and tagged them as either CG or photographic, while we collected eye tracking data. In the second study, we measured reaction times as participants performed the same task as quickly and accurately as possible. Eye tracking data revealed that participants view CG faces with the same "eyes-nose-mouth" pattern in which they view photographs. Overall, we found that participants were highly accurate at the task but that there were no systematic differences in eye movements between CG and photographic portraits. Our results suggest that observers employ a mostly holistic approach to identify photographs, whereas they use a combination of holistic and feature-based approaches to identify a CG portrait. We believe that CG artists should continue to improve the realism of facial features, particularly eyes, but simultaneously attempt to enhance the "holistic" realism of CG faces.**

***Index Terms*—eye tracking, visual perception, faces, realism, computer-generated**

I. Introduction

Computer-generated characters are central to animated movies, games, training and simulation, and social virtual reality. Movies such as "The Matrix", "Avatar" and "Wonderwoman", and video games such as "The Last of Us" and "Call of Duty" have realistic virtual characters in lead roles. Previous work has shown that avatar realism increases co-presence and improves the quality of communication in immersive virtual environments [1]. Human faces have such critical evolutionary importance that people are "hard-wired" to recognize them [2]–[5]. Researchers across disciplines have studied face processing from perceptual, cognitive, neural, developmental, and computational perspectives. This extensive body of work has collected evidence that some aspects of face processing are local, i.e., involving the detection of specific features [6], some aspects are configural, i.e., involving spatial relationships between local features [7], [8], and some aspects involve global processing without high frequency information [9].

In our work, we ask the question: What cues do people use to identify if a face is real or computer-generated? In other words, do people focus on rendering artifacts in the eyes, examine imperfections in the rendered hair, or something else? We conducted two studies where we presented participants with portrait images and asked them whether the picture was computer-generated (CG) or photographic (photo). In the first study, participants were given six seconds to view each image. In the second study, participants were asked to respond as quickly and accurately as possible. We recorded gaze data and participants' responses.

We hypothesized that participants use local features to make the judgment on "CG" or "photo". Based on the popular "Polar Express eyes" trope in the computer graphics community, and the eyes-nose-mouth pattern of eye movements observed on photographic human face images, we hypothesized that participants would look at the eye region in all the portrait images. We expected that either participants might look closer (fewer fixations with longer dwell times) at CG eyes because they appear glassy and lifeless, or, that participants might have an aversive reaction to the eyes due to the uncanny valley, and would thus avert their gaze from the eye region. Therefore, we hypothesized that the dwell time on the eye region, as a percentage of the total time spent on the face, would be different between CG and photographic images.

Our next hypothesis was that if it was not just the eyes, but that other cues were used, such as imperfectly rendered hair, then these cues would likely be different for different CG images, but participants would still exhibit more local gaze behavior (fewer and longer fixations) for CG images because they would look at artifacts closely. In contrast, there would be more exploratory behavior for photo images (shorter fixations, longer saccades) because participants were searching for artifacts.

This work is not intended to be an evaluation of the start-of-the-art in computer graphics, rather it is an investigation into how people discriminate images of CG faces from photographic faces. The experiments are also not intended to make the task hard, or to test if it is possible to fool a participant into "suspending disbelief". The main goal of our experiments is to ascertain whether eye movements are different for "real" versus "fake" faces, similar in spirit to results for real and fake smiles by Calvo et al. [10]. Our findings have implications for creators of computer-generated characters. If eye movements are different, gaze information could inform CG artists of which features to improve. If eye movements are not different across the two groups, that suggests that people use holistic information, and that computer graphics has entered an era where higher-level properties need to be actively considered.

II. Background

Humans can be considered true "face experts", in that we can perform tasks such as face detection and even some degree

Fig. 1: Left: Sample images from dataset. Answer key on Page 9. Right: Experimental setup.

of recognition as infants [4], [5], [11]. Studies employing electrophysiology have shown that the processing of faces involves a special neural sub-system, which is reflected in one of the components of event-related potentials (ERPs) called $N170$ [2], [3]. Depending on the task, there are several modes of face processing. We briefly describe these approaches below:

Feature-specific. Primary facial features such as eyes, nose, mouth, and chin are used to recognize faces and perform other tasks such as gender identification [12].

Configural. This approach involves the use of secondary features, such as distance between eyes, or distance between eyes and mouth, to process faces and has been found to be employed in recognition and identification tasks [7], [8].

Holistic. This theory suggests that face processing involves creating a "gestalt" (perceptive whole independent from its constituent parts) [13]. This causes a decrease in performance on inversion, in attractiveness judgments [14] and identification [15].

Humans are sensitive to the realism of a presented face. In particular, humans have been shown to perform the task of discerning whether a presented face is photographic or computer-generated quite well [16]–[18]. Mader et al. [18] observed that masking out eyes made this task more difficult, suggesting a feature-specific approach.

Previous work in face recognition has shown that inverting a face causes a significant decrease in face identification accuracy, providing evidence that face inversion causes an inability to perceive the individual face as a whole rather than as a collection of features [19]. Farid and Bravo [20] found a significant degradation in performing the CG/photo task when the images were inverted. This finding also supported the idea that participants employ some high-level perceptual judgment, as opposed to some low-level feature-based judgment in order to complete the task. Fan et al. [21] asked participants to make visual realism judgments on real and CG faces. Their results indicated that both holistic and feature-based processing are involved in visual realism perception of faces, with holistic processing becoming more dominant when resolution is lower. Holmes et al. [17] had also tested for the effect of image resolution on performance. They found a very small effect (d' dropped from 1.68 to 1.45 for a resolution change from $600px$ to a mere $100px$). This is suggestive of a holistic approach, where accuracy is fairly high even with low spatial frequencies.

Looser and Wheatley [22] investigated the perception of "life" by presenting participants with morphed images created from animate (human) and inanimate (mannequin) faces. They found that manipulating the eye region had the biggest impact on perceived animacy of faces, suggesting that the eyes are a very important cue in conveying whether a face is "alive". MacDorman et al. [23] conducted four studies by manipulating a computer-generated human character's facial proportions, skin texture, and level of detail to investigate their effect on perceived eeriness, human likeness, and attractiveness. They found that texture photo-realism and polygon count increased human likeness, texture photo-realism heightened the accuracy of human judgments of ideal facial proportions. Moreover, a mismatch in the size and texture of the eyes was especially prone to make a character eerie.

Previous eye tracking studies on perception of real faces have consistently revealed a "Y"-shaped pattern of fixations over the eye, nose and mouth regions [24], [25]. Janik et al. [26] showed that subjects spent 40% of the time on eyes while free viewing facial photographs. In another study, Hsiao et al. [27] found that only two fixations contributed to the optimal face recognition rate. The distribution of the first fixation was just to the left of the center of the nose, and that of the second fixation was around the center of the nose.

Carter et al. [28] found that viewers attend to the face of an avatar for approximately 40% of the viewing time. However, we still lack a clear understanding of the features that people use to identify if an image is "CG" or "photo". To our knowledge, there have been no eye tracking studies

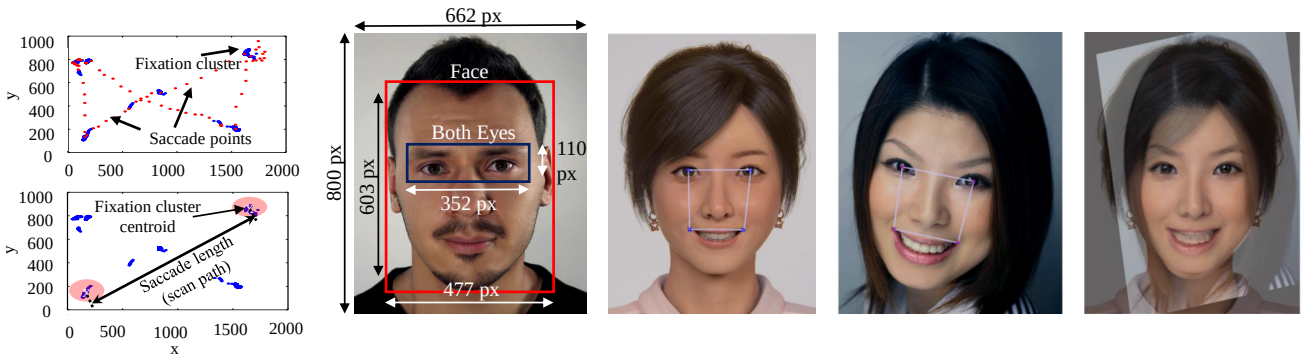

Fig. 2: (a) Top: Gaze samples classified as fixation (blue) or saccade (red) points. Bottom: Fixation cluster centroids are labeled as fixations, and a saccade forms between two consecutive fixations. (b) Example CG image with area of interest marked out around both eyes and face of actor. (c) A CG image annotated with four features marked in blue. (d) Photographic image to be aligned, with strategic four feature points marked in red. (e) Images superimposed after the affine transformation was applied.

conducted on computer-generated faces. This motivated us to use eye tracking to gain insights into where people look while performing the CG/photo task.

III. Experiment

We present two novel eye tracking studies on computer generated (CG) and photographic portraits. In the first study, participants viewed these images for six seconds and tagged them as "CG" or "photo". In the second study, participants were asked to perform the same task as quickly and accurately as possible.

A. Study 1: Fixed Time

Thirty six students from within the university community participated under an IRB approved protocol. Three participants volunteered without any compensation, and the other thirty three students were compensated with course credits. Students who participated for course credits had the option of submitting an extra assignment to receive the same number of credits. All participants had normal or corrected-to-normal vision.

Eye tracking data consisting of gaze position (x, y) and pupil diameter was collected using a remote infra-red eye tracker (SensoMotoric Instruments iView RED-m) at a sample rate of $120Hz$. Stimuli were presented using SMI's presentation software Experiment Center, on an external monitor of resolution 1680×1050 ($18in \times 11in$). Participants were required to support their head with a chin rest and forehead band. Participants viewed the monitor from a distance of 60-62cm. One degree of visual angle is approximately 37 pixels in this setting. The study was conducted in a quiet, well lit room within an indoor lab. A standard USB mouse and keyboard was used to record responses to self-report questions. Figure 1 (Right) shows our experimental setup.

We used the same image dataset as Holmes et al. [17]. This dataset consists of 10 CG images and 10 photographic images as training images, and 30 CG images and 30 photographic images as testing images. All images had a vertical resolution of 600px. These images were downloaded from computer graphics websites such as www.cgsociety.org, www.3dtotal.com, and www.cgarena.com and the work of a CG artist (http://www.romans3d.ru) [29]. All images had a human face posed outward, and a render date between 2013 and 2014. The photographic images were chosen to match the race and gender distribution of the CG set, mostly from http://www.flickr.com. The photographic images were subsequently chosen to match these in age, gender, race, pose and accessories. The backgrounds in both picture categories were comparable (CG had 11 textured backgrounds and 19 smooth backgrounds, while photographic had 13 textured backgrounds and 17 smooth backgrounds).

Holmes et al. adjusted the images with respect to brightness and contrast. The size of faces in CG images had a width = $11.3° \pm 2.4°$ visual angle and a height = $13.7° \pm 3.6°$ visual angle. For photographs, faces had a width = $11.9° \pm 2.5°$ visual angle and a height = $14.2° \pm 3.3°$ visual angle. Images were displayed in their original size and aspect ratio, centered on the screen with a neutral gray background. The order of presentation was randomized.

Upon arrival, participants were asked to sign a consent form. They were given instructions regarding the task both in person as well as on screen. Each session began with a five-point calibration and a five-point validation procedure, provided by SMI's Experiment Center. The experimenter recorded the validation error.

Each trial started with the presentation of a neutral gray slide with a small white fixation cross for two seconds. The fixation screen brightness was matched to the average brightness of the stimuli level (gray level = 127). Participants were instructed to fixate on the cross whenever it appeared on the screen. Next, the training stimulus image was displayed for six seconds. Participants were asked to tag the image they saw as either male or female, and either photographic or computer generated. For these training images, the correct

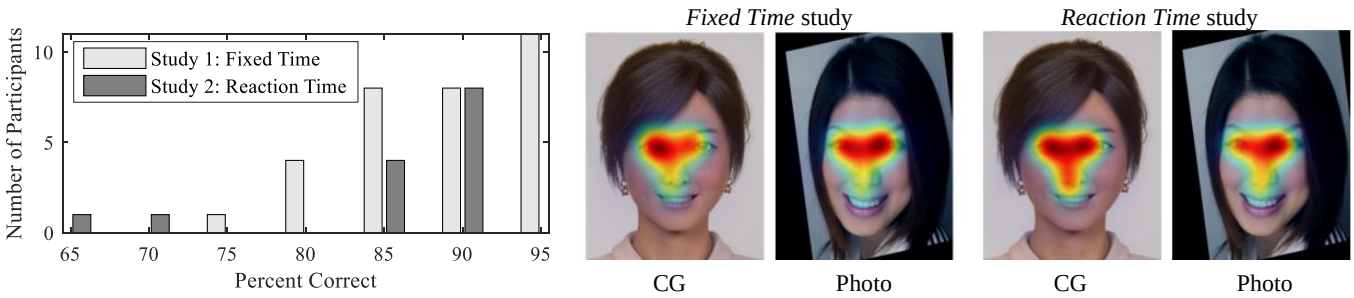

Fig. 3: Left: Histograms show participant accuracy in identifying the picture type as photographic or computer generated (percent correct responses) for Study 1 (light gray) and Study 2 (dark gray). Chance performance is 50%. Right: Spatial saliency maps across all images and participants for both studies. All saliency maps exhibit the "Y" pattern.

answer (male/female and CG/photo) was displayed below the image. The task of identifying gender was to serve as a check that participants were not distracted.

After completing the training session, participants proceeded to perform the same task on the 60 testing images. Finally, participants were asked demographic questions including age range, ethnicity, gender, academic level and major. We also asked them how often they played video games, and about their experience with photo editing software like Adobe Photoshop.

B. Study 2: Reaction Time

Fourteen students from within the university community participated under an IRB approved protocol. While five students were compensated with course credits, nine participants volunteered without any compensation. Students who participated for course credits had the option of submitting an extra assignment instead of participating in the study. The experimental setup conditions were the same as in Study 1.

Following the CG/photo training, participants went through a second reaction time training. Participants were asked to identify a shape (either a circle or a square) as quickly and accurately as possible. This simple task served as a baseline. The outline of the shape was displayed in black, centered on a neutral gray background. Participants were asked to press the space bar on a keyboard as soon as they had identified the shape. After completing ten trials of the reaction time training, participants performed the CG/photo task with the 60 testing images. In this session, we only asked them to tag the images as "CG"/"photo", and not identify the gender.

At the end of the experiment, we asked the same self-report questions as in Study 1. In addition, we asked the participants to rate the the task difficulty on a scale of 1 (easy) to 5 (hard), and which cues they used to perform the discrimination between photographic and computer-generated faces. We provided the following options for cues used: hair, facial hair, eyes, teeth, and skin texture and "other". Participants could select more than one option, and also input any other cues they used as freeform text.

IV. Preliminary Analysis

We discarded data from participants who had a validation error greater than one degree of visual angle (s016, s035 in Study 1, none in Study 2). The remaining 33 participants in Study 1 had an average validation error of $0.53°$ visual angle ($\sigma = 0.13°$). Participants in Study 2 had an average validation error of $0.57°$ visual angle ($\sigma = 0.14°$). We also checked if the remaining participants had reported the gender incorrectly more than two times in Study 1, but there were no such participants. Eye tracking data processing was done in MATLAB R2017a on an AMD FX(tm)-8350 eight-core processor. We used SMI's BeGaze software and R for creating AOIs, computing metrics and generating graphs.

A. Metrics

We define the metrics used to perform preliminary analyses on the collected data, including response accuracy, observer sensitivity/bias, and inter-observer consistency.

Response Accuracy. Each stimulus image has a label assigned to it: 'CG' or 'photo'. For a given participant, response accuracy is computed as the ratio of number of images tagged correctly, to the total number of images viewed by that participant. This metric is a measure of how well a participant performed on the task. Chance performance is 50%.

Observer Sensitivity and Bias. A hit is defined as correctly identifying a computer-generated image whereas a false alarm is incorrectly labeling a photographic image as computer generated. We then computed d' (observer sensitivity) and β (observer bias) [30]. A value of $\beta = 1$ indicates no bias, a value of $\beta > 1$ means that observers are biased towards classifying an image as photographic, and $\beta < 1$ indicates that observers are biased towards classifying an image as CG. If the participant has no information to make a decision, $d' = 0$. When $\beta = 1$ (no bias), a value of $d' = 1$ represents an overall accuracy of 76%. High values of d' imply greater observer sensitivity.

Inter-observer Consistency. Fixations are clusters of gaze points which are close in time and space [31]. Saccades are eye movements used to move the fovea rapidly from one fixation to the next. We used a velocity threshold algorithm [32] to classify gaze points as low or high velocity points, with

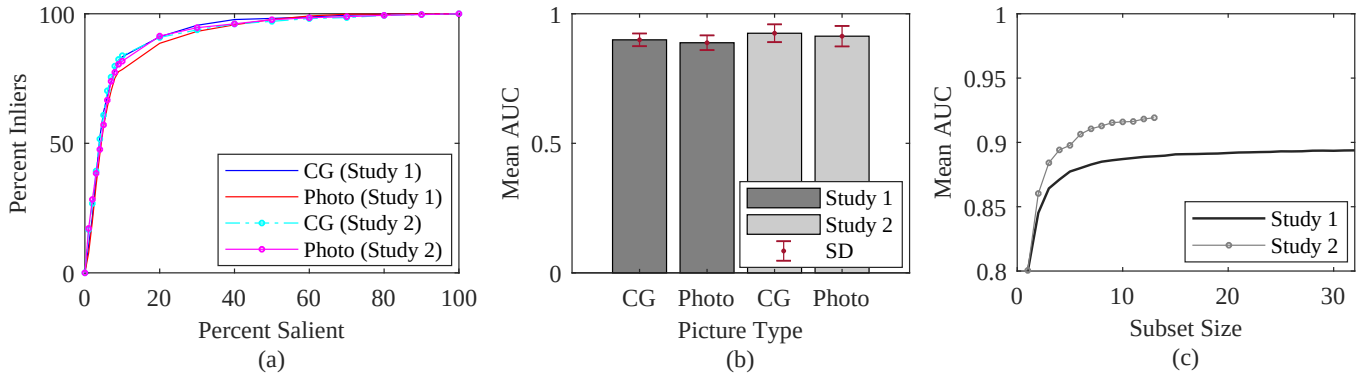

Fig. 4: Inter observer consistency (a): Mean ROC curves for CG and Photographic portraits in both studies. (b): Bar chart shows mean and standard deviation of AUC. (c): Mean AUC increases and achieves convergence with increase in subset size.

a threshold of $130^\circ/s$ (shown in the top panel of Figure 2 (a)). Consecutive low velocity points are clustered into a single fixation, with the cluster centroid marking its location. There is a saccade between every two fixations (bottom panel of Figure 2 (a)).

Spatial Saliency Maps. Because the faces in our stimuli images varied in size, orientation and pose, we aligned them such that the eye and mouth regions would overlap. In order to perform the alignment, we used a 6-parameter affine transform. We manually selected four corresponding features on each face, estimated the affine transform, and then registered all images. The features were: Center of left eye, center of right eye, left corner of mouth and right corner of mouth. Figure 2 (c) and Figure 2 (d) show the selected feature points for an example CG image and an example photographic image. An overlay after the transformation is shown in Figure 2 (e). All stimuli images were aligned to match the image in Figure 2 (c). We generated a fixation map for all participants on each aligned image, by summing up the number of fixations at each pixel location. We averaged fixation maps across the 30 images in each category. We convolved each fixation map with a Gaussian kernel with $\sigma = 1^\circ$ visual angle to generated the heat maps in Figure 3(Right).

Area under ROC Curve (AUC). We created a fixation map for one participant by marking a pixel location as 1 if a fixation fell on it and 0 otherwise. For all other participants, we generated an average fixation map by summing up the number of fixations at each pixel location for each participant, and dividing by the number of participants. We convolved this map with a Gaussian kernel of $\sigma = 1^\circ$ visual angle to get a saliency map. For a given threshold, the percentage of total area inside the binary map is the percent salient for a particular threshold value, and the percentage of fixations that fall inside the saliency map are called inliers. We plotted an ROC curve by varying the value of the threshold. An ideal score is 1 while random classification provides 0.5. We followed the algorithm in [33]. Other variants are presented in [34], [35].

Reaction Time. Reaction time is defined as the time duration between stimulus onset and the keyboard space bar press. For the reaction time analysis, we discarded trials where the reaction time was more than two standard deviations away from the mean, which accounted for any trials with reaction times greater than 2s in the baseline (circle/square) task and 7s in the CG/photo task.

B. Results

We present the results for each metric, first for the *Fixed Time* study and then for the *Reaction Time* study.

1) Study 1: Fixed Time: **Response Accuracy.** Figure 3 (Left) shows the histogram of the accuracy (percentage of correct responses) across all testing images in Study 1 (light gray bars). Participants had an average accuracy of 90.8% ($\sigma = 5.82\%$).

Observer Sensitivity/Bias. The average d' was 2.78 ($\sigma = 0.56$) and the average β was 0.99 ($\sigma = 0.59$). The d' value is higher compared to the results reported by Holmes et al. [17] with Mechanical Turk participants possibly because we conducted the study in a lab setting, where participants were more focused on the task. A β value close to 1 implies that there was no overall bias towards tagging an image as "CG" or "photo".

Spatial Saliency Maps. The overall saliency maps for computer-generated and photographic portraits are presented in Figure 3 (Right). The heat maps are overlayed onto a single example image, but contain data averaged across all viewers and all images in that category. The heat maps show that both CG and photographic faces were explored in a "Y" pattern. This pattern is consistent with earlier findings on human faces [26], [36].

AUC. We observed a mean AUC value of 0.8994 ($\sigma = 0.031$) across all CG images, and a mean AUC value of 0.8882 ($\sigma = 0.029$) across all photographic images. The mean ROC curves for each picture type, by leaving out each observer in turn, are presented in Figure 4 (a). The AUC value for all images was above 0.8, showing high agreement across participants. Additionally, Figure 4 (b) shows a barchart of the mean ROC area for each picture type.

We computed the mean AUC value across all images by performing the ROC analysis with each observer's fixation

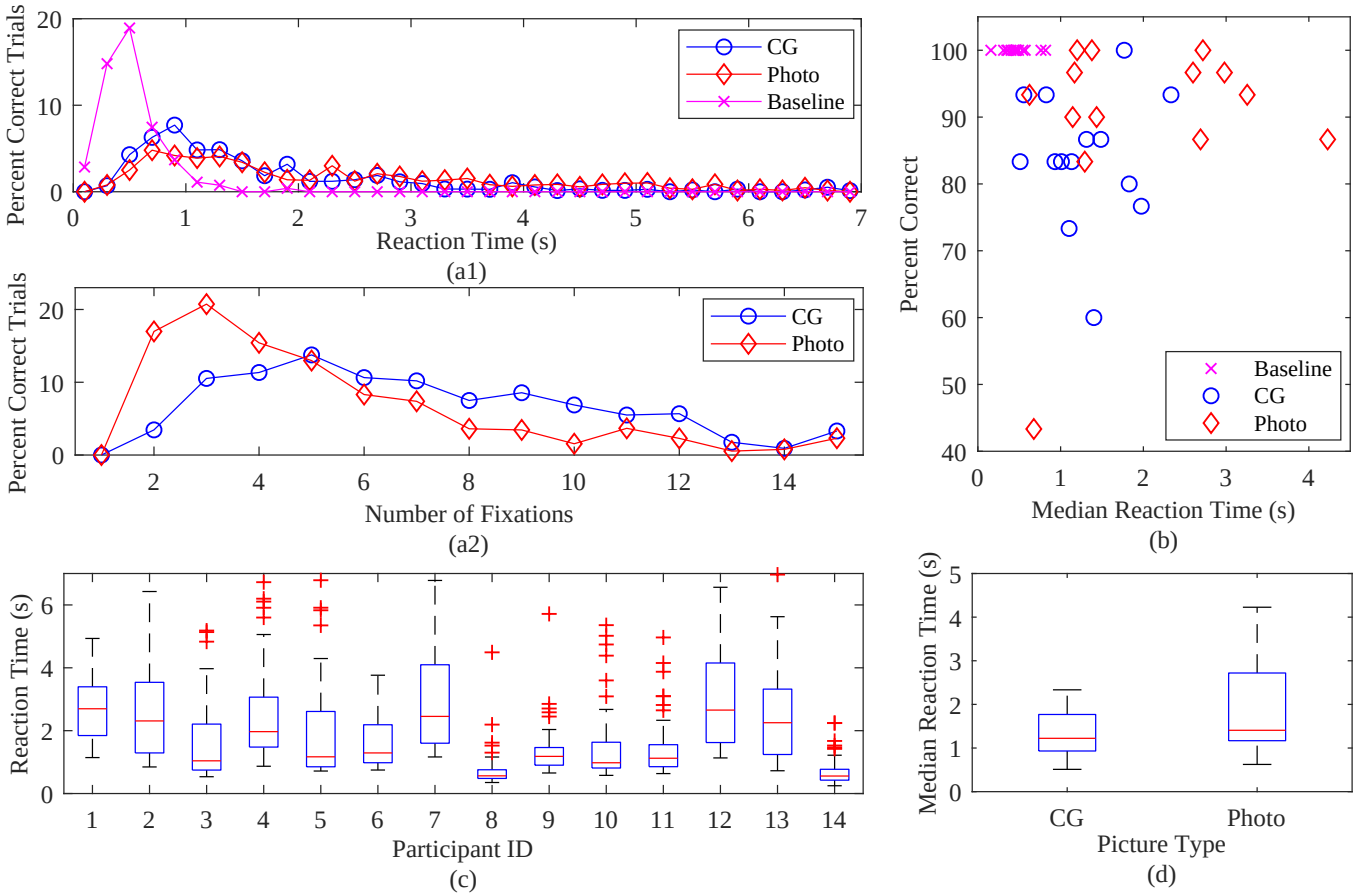

Fig. 5: (a1) Percent correct trials as a function of reaction time. Reaction times are divided into bins of 0.2s. (a2) Percent correct trials as a function of number of fixations. Number of fixations are divided into bins of 1. (b) Plot shows participants' accuracy as a function of their median reaction time. (c) Box plots show participants' reaction times. (d) Median reaction times averaged across all CG and Photographic images.

map and a saliency map generated from a random subset of the other observers. The subset size was varied between 1 to 32. Figure 4 (c) plots the mean AUC at each subset size, along with the AUC standard deviation across participants. The AUC value rises sharply and then flattens out.

We computed the AUC scores without taking into account any fixations inside the AOI on both eyes. The mean AUC dropped slightly to 0.8767 ($\sigma = 0.041$) across all CG images, and to 0.8542 ($\sigma = 0.044$) across all photographic images. This shows fairly high agreement between observers, even without the eye fixations (where participants looked at for half their viewing time).

2) Study 2: Reaction Time: **Response Accuracy.** When participants were asked to perform the CG/photo task as quickly as possible, they were still able to perform the task with an average accuracy of 86.9% ($\sigma = 7.48$) (dark gray bars in Figure 3(Left)).

Observer Sensitivity/Bias. Observer sensitivity dropped to $d' = 2.46$ ($\sigma = 0.54$), while bias went up to $\beta = 2.18$ ($\sigma = 1.45$), i.e. participants were biased towards tagging an image as photographic.

Spatial Saliency Maps. The saliency maps for CG and photographic portraits for Study 2 are presented in Figure 3 (Right). The heat maps show that both CG and photographic faces were explored in the same "Y" pattern, even when participants were asked to perform the task as quickly as possible.

AUC. We observed a mean AUC value of 0.9249 ($\sigma = 0.027$) across all computer-generated images, and a mean AUC value of 0.9135 ($\sigma = 0.026$) across all photographic images.

We computed the mean AUC value across all 60 images by performing the ROC analysis with each participant's fixation map and a saliency map generated from a random subset of the other participants. The subset size was varied between 1 to 13. The AUC value increases sharply and then flattens out, similar to results in the *Fixed Time* study ((Figure 4 (c))).

The mean AUC dropped to 0.8926 ($\sigma = 0.046$) across all CG images, and to 0.8794 ($\sigma = 0.048$) across all photographic images, when we removed fixations on the region marked *both eyes*.

Reaction Time. The recorded reaction times showed that participants were able to identify CG and photographic images in 1.42s (median). This is higher than the median reaction

time for the baseline task (0.51s), which involved assessing a single feature, i.e. circle or square. These reaction times include the participant's motor response. Figure 5 (a) shows percent correct trials for the 14 participants. Figure 5 (b) shows each participant's accuracy on the baseline task and on the CG/photo task. There was no apparent speed-accuracy tradeoff: participants had a high accuracy both at low and high reaction times. Participants had varying reaction times across various images (see Figure 5 (c)). CG images had a lower median reaction time (1.29s) compared to photographs (1.64s) (Figure 5 (d)). A paired samples t-test revealed that this difference was statistically significant ($t(13) = -3.71$, $p < 0.05$).

V. Main Analysis

We present the metrics used to analyze the collected eye tracking data, followed by the results for Study 1 and Study 2.

A. Metrics

AOI Analysis (Dwell time). We defined an area of interest (AOI) around both eyes and the face, for each stimulus image. An example is presented in Figure 2 (b). The areas of interest were marked manually using the AOI editor provided by SMI's BeGaze software. For the eye region, we included eyebrows. For the face, we used ears, forehead hairline and chin as boundary markers. For a given picture and participant, we calculated the percentage dwell time on eyes as the ratio between the dwell time on both eyes and the dwell time on the entire face.

Number of fixations. For each participant, we summed up the total number of fixations on each image. We computed the average number of fixations across all CG and photographic images.

Average fixation duration. Fixation duration is defined as the difference between the onset of a fixation and the onset of the next saccade. First, we averaged the fixation duration across all CG images for a particular participant. This procedure was repeated for photographic images. We compared the average fixation duration on CG and photographic images for all participants.

Average saccade length. Saccade length is the distance between two consecutive fixations in pixels (see bottom panel of Figure 2 (a)). We computed average saccade length for each participant across all CG and all photographic images.

B. Results

We present the results of the described metrics, first for Study 1 and then for Study 2. The means and standard deviations of the metrics we used are presented in Table I.

TABLE I: Means and standard deviations of metrics used, for Study 1 and Study 2.

Metric (Study 1)		CG	Photo	All
% Dwell Time Both Eyes	μ	53.3	53.8	53.5
	σ	11.5	11	11.2
Number of Fixations	μ	14.7	15.5	15.1
	σ	4	4.7	4.3
Average Fixation Duration (s)	μ	0.5	0.48	0.5
	σ	0.27	0.34	0.3
Average Saccade Length (px)	μ	162.8	166.8	164.8
	σ	22.6	22.8	22.6
Metric (Study 2)		CG	Photo	All
% Dwell Time Both Eyes	μ	56.4	55.5	55.9
	σ	22.9	23.1	22.9
Number of Fixations	μ	6.3	6.9	6.6
	σ	3.4	3.5	3.2
Average Fixation Duration (s)	μ	0.4	0.4	0.4
	σ	0.06	0.06	0.06
Average Saccade Length (px)	μ	152.8	157.3	161.9
	σ	20.2	20.1	21.7

1) Study 1: Fixed Time: **Percent Dwell Time on Eyes.** Figure 6 (a) shows box plots of the percent dwell time on both eyes for thirty three participants, on CG and photographic images respectively. On average, participants spent approximately 54% of their time focusing on the eye region of the face ($\sigma = 11\%$). We performed a Shapiro-Wilks test and found that the data was not normal ($p < 0.05$). A Wilcoxon signed-rank test revealed no significant difference between the mean percent dwell time on both eyes for computer-generated and photographic images ($Z = -0.65$, $p = 0.51$).

Number of Fixations. Figure 6 (b) shows a box plot of the number of fixations across all participants for CG and photographic images. CG images had approximately one fewer fixation than photographs. A paired samples t-test showed that the difference in means was significant ($t(32) = -2.89$, $p < 0.05$).

Average Fixation Duration. A box plot of the average fixation duration (in ms) for CG and photographs is shown in Figure 6 (c). A Shapiro-Wilks test showed that the data was not normal ($p < 0.05$). The average fixation duration on CG was 20ms more than on photographic images. This difference was statistically significant (Wilcoxon signed-rank test, $Z = -2.74$, $p < 0.05$).

Average Saccade Length. Figure 6 (d) shows a box plot of the average saccade length for all participants across CG and photographic images. Photographs have a slightly longer saccade length than CG images by about 4 pixels. The Shapiro-Wilks test showed that data was not normal ($p < 0.05$). A Wilcoxon signed-rank test showed that this difference is significant ($Z = -2.46$, $p < 0.05$).

2) Study 2: Reaction Time: **Percent Dwell Time on Eyes.** Figure 6 (e) shows box plots of the percent dwell time on both eyes for CG and photographic images respectively. Participants spent nearly 56% of their time on the eye region of the face ($\sigma = 22\%$). There was no significant difference in group means (paired samples t-test, $t(13) = 0.84$, $p = 0.42$).

Number of Fixations. A box plot of the total number of fixations across all participants for CG and photographic images is presented in Figure 6 (f). A paired samples t-test showed no significant differences in means ($t(13) = -1.03$, $p = 0.31$).

Average Fixation Duration. Figure 6 (g) shows a box plot of the average fixation duration (in ms) for CG and photographs.

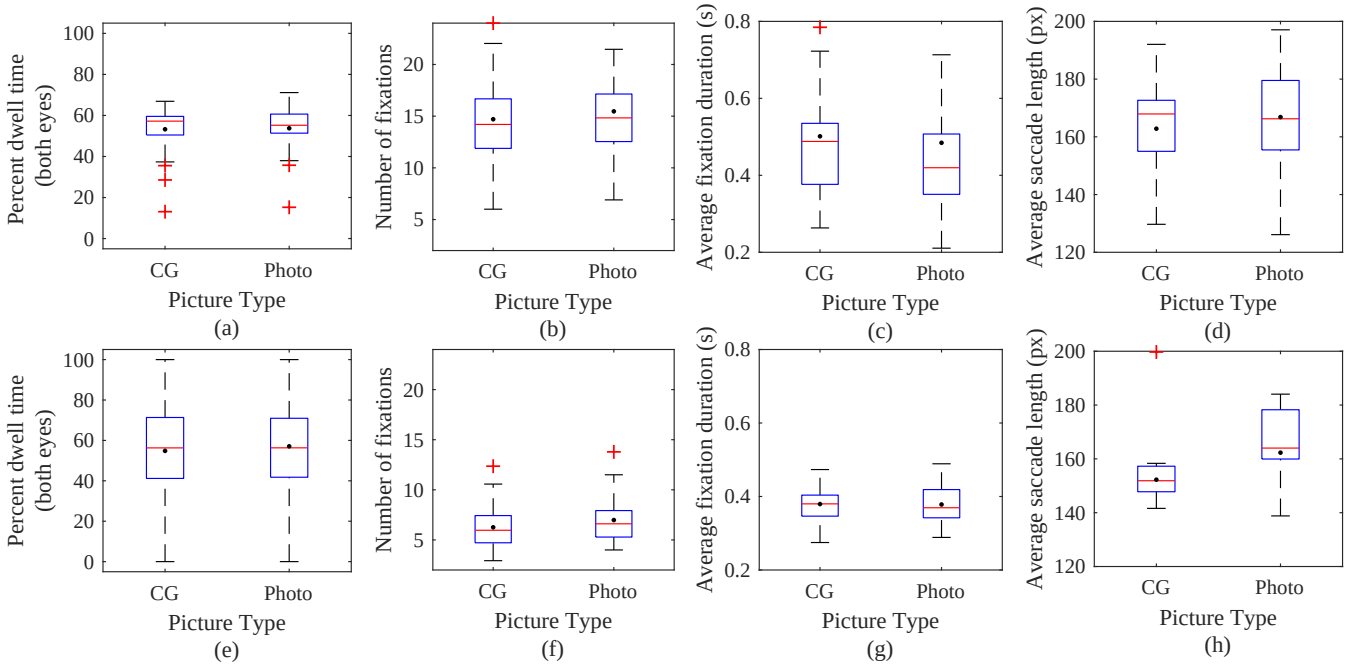

Fig. 6: Box plots showing (a,e) percentage dwell time on both eyes, (b,f) number of fixations, (c,g) average fixation duration, and (d,h) average saccade length across all CG and photographic images in Study 1 and Study 2 respectively. On each box, the line inside represents the median, and the black dot marks the mean. Bottom and top edges of the box indicate the 25^{th} and 75^{th} percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers ($\pm 2.7\sigma$). Individual outliers are marked using the $+$ symbol. On average, participants spent approximately 55% of the time on the eye region in both studies.

There was no significant difference between the average fixation duration on computer-generated and photographic images (paired samples t-test, $t(13) = 0.14$, $p = 0.88$).

Average Saccade Length. Figure 6 (h) shows a box plot of the average saccade length for all participants across CG and photographic images. The Shapiro-Wilks test showed that data was not normal ($p < 0.05$). A Wilcoxon signed-rank test showed a significant difference in means ($Z = -2.29$, $p < 0.05$).

VI. Discussion

Our first hypothesis was that the dwell time on the eye region, as a percentage of the total time spent on the face, would be different between CG and photographic images. However, our results showed no significant differences. As expected, participants spent almost half their viewing time time on eyes (54% and 56% in the first and second study respectively). Although participants spent most of their time on eyes, the proportion of fixation time that each participant spent fluctuated widely in the second study. Even though on average the eyes were fixated on for the most time, the values of individual participants ranged between 0% and 100% (Figure 6). This suggests that under time constraints, people adopt different strategies to perform the same task: While eyes are definitely an important cue in the first study, participants may have relied on other features in the second study.

We also found no systematic differences in gaze behavior (number of fixations, average fixation duration, or average saccade length) between computer-generated and photographic images. Participants viewed faces in the "eyes-nose-mouth" pattern.

In the first study, participants had an average accuracy of 90.8% in the *Fixed Time* study and 86.9% in the *Reaction Time* study, which is higher than previously observed accuracy by Holmes et al. [17]. These differences are most likely because our experiment was conducted in a laboratory setting causing an increase in participant focus, and hence the overall increase in accuracy.

Prior work has shown that people can tell the difference between paper, plastic and fabric within $500ms$ of viewing an image [37]. Naive observers can also recognize the gist of a novel scene in a brief glance (40ms) [38]. In our study, the median reaction time to identify images as "CG" or "photo" was 1.4s. Accounting for motor response time, it is likely that the participants made the judgment sooner than that. Participants were able to identify almost 40% of photographs correctly within the first 3 fixations, while they took almost 6 fixations to identify that many CG faces correctly (Figure 5(a2)). This suggests that participants require more information in addition to the first holistic glimpse to make the judgment on whether a face is "CG", i.e. they look at least at local features (eyes-nose-mouth) to make this call (Figure 7(b)). This is

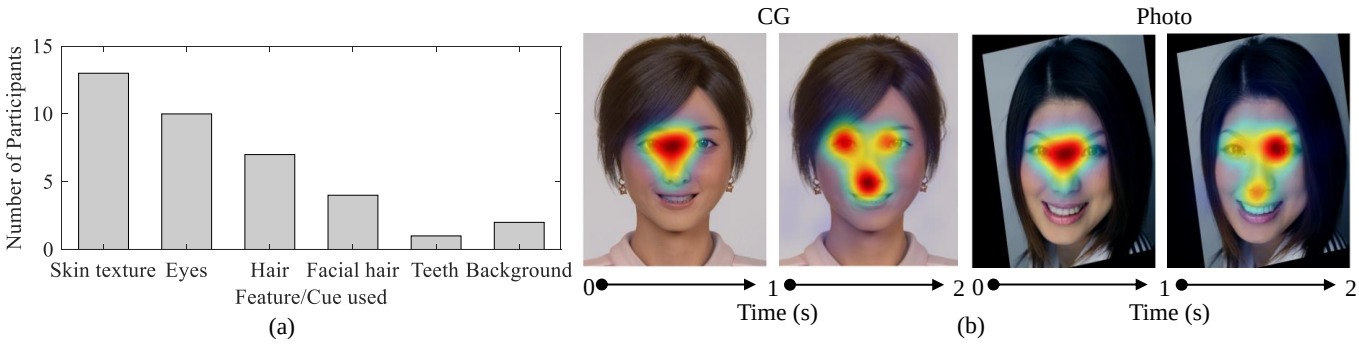

Fig. 7: (a) Bar graph shows the self-reported cues that participants used to perform the CG/photo task. (b) Spatial saliency maps for CG and photographic portraits in the first 2 seconds of viewing.

consistent with the findings of Fan et al. [21]: inversion made realism judgments harder for photographic images but not for CG ones, suggesting a combination of holistic and feature-based approaches to perform the CG/photo task. Unlike face recognition, identifying if an image is computer-generated or photographic takes more than just two fixations [27].

We found evidence for both holistic and configural modes of processing while performing the CG/photo task. Eyes were definitely an important cue, but while some participants carefully analysed the primary features before responding, some responded correctly in under 1.4 seconds. We believe that CG artists should continue to improve the realism of eyes, but also direct efforts at providing CG faces with a more "holistic" realism.

Familiarity is an important factor while studying the visual perception of faces [39]. The study by Mader et al. [18] showed that observers' performance with familiar faces is slightly better. However, because we are studying differences between CG and photographic images, and both groups had matching faces, familiarity is balanced in our experiment.

We had asked participants which cues they used to perform the discrimination (see Figure 7(a)). Eye color (light or dark) is matched between CG and photographs in our dataset, however, eyes in the photographs had significantly more asymmetry. Our dataset had a closely matching number of textured and smooth faces. Interestingly, 19 out of 30 photographs had diffused lighting, whereas only 11 CG images had diffused lighting. However, lighting was not explicitly stated as a cue used by any participant in their freeform response. One limitation of our self-reported responses is that skin texture was the only option that was not a distinct facial feature, i.e. it is possible participants assumed any non-facial feature to be skin texture (lighting/shadows, wrinkles, graininess).

Limitations and Future Work. This work focused on static CG portraits. When motion is available as a cue, it is likely that a different approach would be used to process faces [40]. While we have a better understanding of which visual regions of the face humans tend to focus on while identifying images as "CG" or "photo", we still do not know how they cognitively process this visual information. In future, pupil diameter collected along with gaze data, or EEG could be leveraged to gain further insights. Future work could also investigate accuracy in the CG/photo task using millisecond exposure of the stimuli images.

Our study visualizes how humans view photographic and computer-generated faces. Further work can try to better determine the speed-accuracy relationship, by showing different groups of participants the images for different durations. Our results suggest that people adopt a combination of holistic and configural approaches while trying to classify faces as computer-generated or photographic. We believe that while attention to improving primary features, most importantly eyes, is crucial, attention must also be paid to granting CG faces a holistically "real" impression.

[**Answer Key for Figure 1(Left):** Computer-generated images in top panel, matching photographic images in bottom panel.]

References

[1] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich, "The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 14, no. 4, pp. 379–393, 2005.

[2] S. Bentin, T. Allison, A. Puce, E. Perez, and G. McCarthy, "Electrophysiological studies of face perception in humans," *Journal of cognitive neuroscience*, vol. 8, no. 6, pp. 551–565, 1996.

[3] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *Journal of neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.

[4] D. Maurer and R. E. Young, "Newborn's following of natural and distorted arrangements of facial features," *Infant Behavior and Development*, vol. 6, no. 1, pp. 127–131, 1983.

[5] E. Valenza, F. Simion, V. M. Cassia, and C. Umiltà, "Face preference at birth." *Journal of experimental psychology: Human Perception and Performance*, vol. 22, no. 4, p. 892, 1996.

[6] J. W. Shepherd, "Studies of cue saliency," *Perceiving and remembering faces*, pp. 105–131, 1981.

[7] C. J. Mondloch, R. Le Grand, and D. Maurer, "Configural face processing develops more slowly than featural face processing," *Perception*, vol. 31, no. 5, pp. 553–566, 2002.

[8] G. J. Hole, P. A. George, and V. Dunsmore, "Evidence for holistic processing of faces viewed as photographic negatives," *Perception*, vol. 28, no. 3, pp. 341–359, 1999.

[9] M. J. Farah, K. D. Wilson, M. Drain, and J. N. Tanaka, "What is" special" about face perception?" *Psychological review*, vol. 105, no. 3, p. 482, 1998.

[10] M. G. Calvo, A. Gutiérrez-García, P. Avero, and D. Lundqvist, "Attentional mechanisms in judging genuine and fake smiles: Eye-movement patterns." *Emotion*, vol. 13, no. 4, p. 792, 2013.

[11] J. Morton and M. H. Johnson, "Conspec and conlern: a two-process theory of infant face recognition." *Psychological review*, vol. 98, no. 2, p. 164, 1991.

[12] A. Schwaninger, S. Ryf, and F. Hofer, "Configural information is processed differently in perception and recognition of faces," *Vision Research*, vol. 43, no. 14, pp. 1501–1505, 2003.

[13] J. Sergent, "An investigation into component and configural processes underlying face perception," *British Journal of Psychology*, vol. 75, no. 2, pp. 221–242, 1984.

[14] I. M. Santos and A. W. Young, "Inferring social attributes from different face regions: Evidence for holistic processing," *Quarterly Journal of Experimental Psychology*, vol. 64, no. 4, pp. 751–766, 2011.

[15] V. Goffaux and B. Rossion, "Faces are" spatial"–holistic face perception is supported by low spatial frequencies." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 32, no. 4, p. 1023, 2006.

[16] S. Fan, T.-T. Ng, J. S. Herberg, B. L. Koenig, and S. Xin, "Real or fake?: human judgments about photographs and computer-generated images of faces," in *SIGGRAPH Asia 2012 Technical Briefs*. ACM, 2012, p. 17.

[17] O. Holmes, M. S. Banks, and H. Farid, "Assessing and improving the identification of computer-generated portraits," *ACM Transactions on Applied Perception (TAP)*, vol. 13, no. 2, p. 7, 2016.

[18] B. Mader, M. S. Banks, and H. Farid, "Identifying computer-generated portraits: The importance of training and incentives," *Perception*, vol. 46, no. 9, pp. 1062–1076, 2017.

[19] G. Van Belle, P. De Graef, K. Verfaillie, B. Rossion, and P. Lefèvre, "Face inversion impairs holistic perception: Evidence from gaze-contingent stimulation," *Journal of Vision*, vol. 10, no. 5, pp. 10–10, 2010.

[20] H. Farid and M. J. Bravo, "Perceptual discrimination of computer generated and photographic faces," *Digital Investigation*, vol. 8, no. 3, pp. 226–235, 2012.

[21] S. Fan, R. Wang, T.-T. Ng, C. Y.-C. Tan, J. S. Herberg, and B. L. Koenig, "Human perception of visual realism for photo and computer-generated face images," *ACM Transactions on Applied Perception (TAP)*, vol. 11, no. 2, p. 7, 2014.

[22] C. E. Looser and T. Wheatley, "The tipping point of animacy: How, when, and where we perceive life in a face," *Psychological science*, vol. 21, no. 12, pp. 1854–1862, 2010.

[23] K. F. MacDorman, R. D. Green, C.-C. Ho, and C. T. Koch, "Too real for comfort? uncanny responses to computer generated faces," *Computers in human behavior*, vol. 25, no. 3, pp. 695–710, 2009.

[24] J. M. Henderson, C. C. Williams, and R. J. Falk, "Eye movements are functional during face learning," *Memory & cognition*, vol. 33, no. 1, pp. 98–106, 2005.

[25] G. J. Walker-Smith, A. G. Gale, and J. M. Findlay, "Eye movement strategies involved in face perception," *Perception*, vol. 42, no. 11, pp. 1120–1133, 2013.

[26] S. W. Janik, A. R. Wellens, M. L. Goldberg, and L. F. Dell'Osso, "Eyes as the center of focus in the visual examination of human faces," *Perceptual and motor skills*, vol. 47, no. 3, pp. 857–858, 1978.

[27] J. H.-w. Hsiao and G. Cottrell, "Two fixations suffice in face recognition," *Psychological Science*, vol. 19, no. 10, pp. 998–1006, 2008.

[28] E. J. Carter, M. Mahler, and J. K. Hodgins, "Unpleasantness of animated characters corresponds to increased viewer attention to faces," in *Proceedings of the ACM Symposium on Applied Perception*. ACM, 2013, pp. 35–40.

[29] O. B. Holmes, "How realistic is photorealistic?" 2015.

[30] D. Green and J. Sweats, "Signal detection theory and psychophysics," 1988.

[31] A. Vision, "The psychology of looking and seeing," 2003.

[32] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM, 2000, pp. 71–78.

[33] E. Jain, Y. Sheikh, and J. Hodgins, "Inferring artistic intention in comic art through viewer gaze," in *Proceedings of the ACM Symposium on Applied Perception*. ACM, 2012, pp. 55–62.

[34] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior research methods*, vol. 45, no. 1, pp. 251–266, 2013.

[35] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.

[36] M. F. Peterson and M. P. Eckstein, "Looking just below the eyes is optimal across face recognition tasks," *Proceedings of the National Academy of Sciences*, vol. 109, no. 48, pp. E3314–E3323, 2012.

[37] L. Sharan, R. Rosenholtz, and E. H. Adelson, "Accuracy and speed of material categorization in real-world images," *Journal of vision*, vol. 14, no. 9, pp. 12–12, 2014.

[38] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–36, 2006.

[39] R. A. Johnston and A. J. Edmonds, "Familiar and unfamiliar face recognition: A review," *Memory*, vol. 17, no. 5, pp. 577–596, 2009.

[40] D. Piepers and R. Robbins, "A review and clarification of the terms "holistic","configural", and "relational" in the face perception literature," *Frontiers in psychology*, vol. 3, p. 559, 2012.