Human Latent Metrics: Perceptual and Cognitive Response Correlates to Distance in GAN Latent Space for Facial Images

Kye Shimizu*
KyeShimizu@gmail.com
Sony Computer Science Laboratories, Inc.
Tokyo, Japan

Naoto Ienaga*
ienaga@iit.tsukuba.ac.jp
University of Tsukuba
Tsukuba, Japan

Kazuma Takada
k-takada@csl.sony.co.jp
Sony Computer Science Laboratories, Inc.
Tokyo, Japan
Okinawa Institute of Science and Technology Graduate University
Okinawa, Japan

Maki Sugimoto
maki.sugimoto@keio.jp
Keio University
Yokohama, Japan

Shunichi Kasahara
kasahara@csl.sony.co.jp
Sony Computer Science Laboratories, Inc.
Tokyo, Japan

ABSTRACT

Generative adversarial networks (GANs) generate high-dimensional vector spaces (latent spaces) that can interchangeably represent vectors as images. Advancements have extended their ability to computationally generate images indistinguishable from real images such as faces, and more importantly, to manipulate images using their inherit vector values in the latent space. This interchangeability of latent vectors has the potential to calculate not only the distance in the latent space, but also the human perceptual and cognitive distance toward images, that is, how humans perceive and recognize images. However, it is still unclear how the distance in the latent space correlates with human perception and cognition. Our studies investigated the relationship between latent vectors and human perception or cognition through psycho-visual experiments that manipulates the latent vectors of face images. In the perception study, a change perception task was used to examine whether participants could perceive visual changes in face images before and after moving an arbitrary distance in the latent space. In the cognition study, a face recognition task was utilized to examine whether participants could recognize a face as the same, even after moving an arbitrary distance in the latent space. Our experiments show that the distance between face images in the latent space correlates with human perception and cognition for visual changes in face imagery, which can be modeled with a logistic function. By utilizing our methodology, it will be possible to interchangeably convert between the distance in the latent space and the metric of human perception and cognition, potentially leading to image processing that better reflects human perception and cognition.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Interaction devices*; *Displays and imagers.*

KEYWORDS

generative adversarial networks, change perception, face cognition

ACM Reference Format:
Kye Shimizu, Naoto Ienaga, Kazuma Takada, Maki Sugimoto, and Shunichi Kasahara. 2022. Human Latent Metrics: Perceptual and Cognitive Response Correlates to Distance in GAN Latent Space for Facial Images. In *ACM Symposium on Applied Perception 2022 (SAP '22), September 22–23, 2022, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3548814.3551460

1 INTRODUCTION

Face images are comprised of wide range of complicated parameters and are significant visual information for humans[Valentine et al. 2016]. To understand the mechanism of facial perception, the concept of a multidimensional psychological space of face was investigated by using computer generated illustrations and artificial 3D faces [Blanz and Vetter 1999; Egger et al. 2019; Jozwik et al. 2022; Valentine et al. 2016]. Recent emergence of generative adversarial networks (GANs), including StyleGAN [Karras et al. 2019], enable us to represent various levels of visual information features in a high-dimensional vector space (latent space) by training data, even with a complex distribution. By projecting visual information into a latent vector space, we can computationally manipulate and edit visual information [Abdal et al. 2020; Pan et al. 2021; Viazovetskyi et al. 2020]. Among the myriad of GANs, the StyleGAN series stands out for its ability to generate photo-realistic images that are indistinguishable from real photographs at first glance. An important property of latent space is that low-level information such as texture and high-level information such as gender, facial expression,

*Both authors contributed equally to this research.

SAP '22, September 22–23, 2022, Virtual Event, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9455-0/22/09…$15.00
https://doi.org/10.1145/3548814.3551460

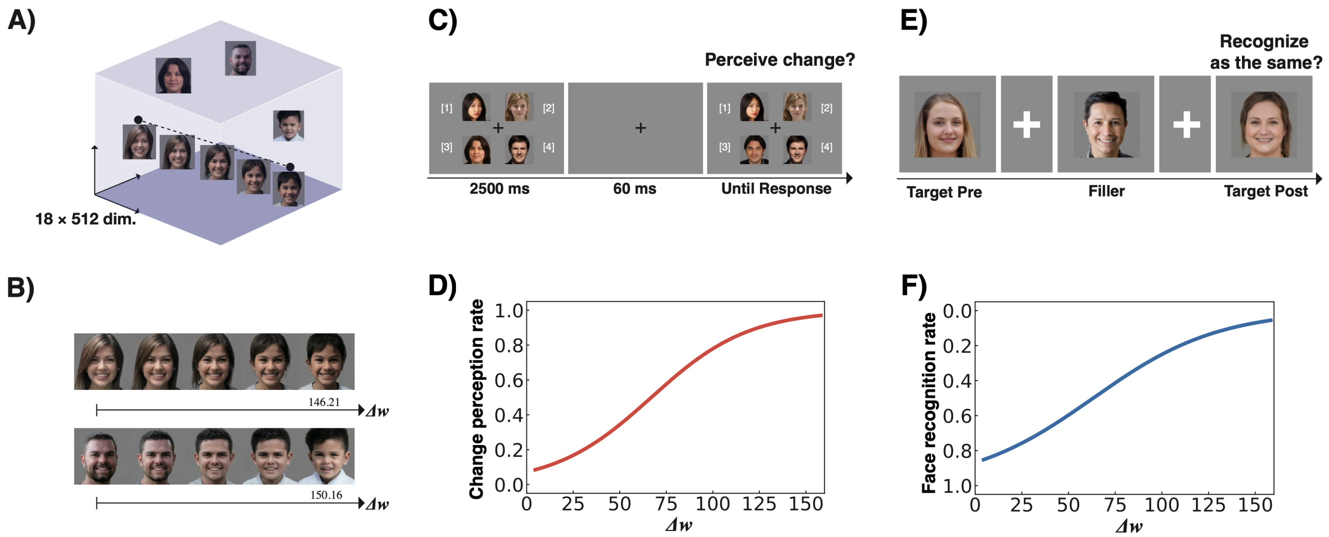


Figure 1: A) A conceptual schematic of a visual information arrangement in the latent space, where a vector in manifold representing a face image is constructed in a latent space of 18×512 dimensions. By interpolating the latent vectors along the vectors connecting arbitrary points in the latent space, we can generate an intermediate face of certain faces with arbitrary steps. B) Examples of the series of face image stimuli generated for the experiments. A face gradually morphs into another face as the difference in the latent vector Δw increases. C) In a change perception task, after 66 ms of a visual blank, one of the four face images is changed in accordance with Δw. Participants are asked to answer which face was changed. D) We investigated the correspondence between the change perception (CP) rate (how many participants perceived the change) and Δw. The results of logistic regression in CP rate clearly show the relationship between human change perception and Δw. E) In a face recognition task, after a target pre-face is presented in a sequence, the target post-face image was presented again with the corresponding latent vector traveled Δw from the initial presentation, i.e., the post-face would be some level of a similar face with pre-face. Participants were instructed to press a key when they recognized a face they had seen before. F) We also investigated the correspondence between face recognition (FR) rate (how many participants recognized it to be the same face) and Δw. The results of logistic regression in FR rate clearly shows the relationship between human face recognition and Δw.

skin color, and posture for face images, can be represented as a vector [Alaluf et al. 2021; Geng et al. 2020; Shen et al. 2019; Ververas and Zafeiriou 2020]. Image manipulation by transforming the latent vector enables new image editing in which various features and attributes are continuously manipulated [Goetschalckx et al. 2019; Jiang et al. 2020].

However, the relationship between latent vector manipulation and the characteristics of human perception has been reported only qualitatively. Then, we hypothesized that latent vector manipulation correlates with human perceptual and cognitive responses. The purpose of this study is to model the correspondence between the distance in the latent space with the perceptual and cognitive properties of humans. This interchangeable connection enables the ability to estimate how another person will perceive changes and generate images which have been altered with human perception in mind.

To investigate our hypothesis and building upon previous research that correlates trained networks with human perception and cognitive responses [Morgenstern et al. 2020; Son et al. 2021], we conducted visual psychological experiments using face images generated from latent vectors of GANs to measure human perceptual and cognitive responses to visual stimuli. The experiments are as follows; a change perception task, which is based on the change blindness paradigm [Rensink 2005], to investigate perception and a face recognition task, which is based on the Exposure Based Face Memory Test [Gillian Rhodes , Andy Calder, Mark Johnson, and James V. Haxby 2011] and image memory game [Isola et al. 2011].

Our results showed that the human response to face change perception and cognition clearly correlates with the distance in the latent space, which can be modeled with a logistic function. These results contribute to providing a new metric, the corresponding distance in the latent space of a visual stimuli, which can also be represented as the perceptual and cognitive distance. In addition, and more importantly, our findings are a step towards an interchangeable connection between a generative model with GAN latent vectors and human perceptual and cognitive responses. This will not only be useful for manipulating images while considering human perceptual and cognitive distances, but will also provide a common attribute for further cognitive science experiments.

2 RELATED WORK

2.1 GAN as generative visual stimuli

Previous studies have revealed the relationship between human perception and cognitive response to the distance in stimulus space with continuously generated visual stimulus such as colors [Luck and Vogel 2013; Schurgin et al. 2020]. By transforming a vector in the space, the features in an interpretable image can be smoothly changed with a virtually infinite number of parametric properties[Geng et al. 2020; Shen et al. 2019; Ververas and Zafeiriou 2020].

Another study took advantage of GANs to continuously generate realistic images to examine the relationship between generated images and mental representations of visual experiences in terms of perceptual similarity and memory properties, and reported similarities with previous studies obtained with simpler visual stimuli [Son et al. 2021]. In addition to improving the quality of the images generated by GANs, they presented a new experimental approach to cognitive science by generating continuously changing visual stimuli in a latent space composed of meaningful information structures [Goetschalckx et al. 2021]. The results obtained from visual experiments based on GANs are expected to be applied to the design of visual displays. Kasahara et al. investigate how visual change blindness happens with GAN-based morphing continuous images, which contributes covert visual updates without consuming users’ attention [Kasahara and Takada 2021]. However, the results were expressed only as percentages and not as distances on the latent space.

2.2 Latent space for human psychophysics

Previous research has proposed a methodology to show how visual information is internally represented in a GAN generator and discovered interpretable units in the network [Bau et al. 2018]. In the training process of GANs[Karras et al. 2019], perceptual smoothness is taken into account such as perceptual path length (PPL), which is defined from other pre-trained Visual Geometry Group (VGG) architecture networks [Simonyan and Zisserman 2014] trained on ImageNet [Deng et al. 2009] classification. However, the correspondence between images generated from a latent space and human perception and cognition has been confirmed only heuristically. While the distance in a VGG feature space was used for “perceptual loss” for image regression problems [Dosovitskiy and Brox 2016; Johnson et al. 2016], to reflect the human perceptual response, Learned Perceptual Image Patch Similarity (LPIPS) was trained from actual human responses through a psychophysical study [Zhang et al. 2018], which outperforms all previous metrics. The aforementioned research indicates that acquiring better metrics to represent human perception also benefits various applications.

3 METHOD

3.1 Generative adversarial network

GANs are image generation models originally developed by Goodfellow et al. in 2014 [Goodfellow et al. 2014]. A GAN can generate quite realistic images by training a generator and discriminator adversarially. The number of studies of GANs is increasing year by year, and various GAN frameworks have been proposed for a wide variety of purposes. We used StyleGAN2, which is an improved version of StyleGAN [Karras et al. 2019], for this research to not only utilize the ability to manipulate $\mathbf{w}$ in the latent space for image editing, but also generate images that are not easily identifiable as changed or altered.

StyleGAN2, which is used in this research, first obtains a latent vector $\mathbf{w}$, which consists of 18×512 dimensional real numbers, by passing the latent code $\mathbf{z}$, which consists of 512 dimensional real numbers, through a nonlinear mapping network. Then, by inputting $\mathbf{w}$ into a synthetic network, an image is generated. Since the synthetic network can generate an image from $\mathbf{w}$, $\mathbf{w}$ potentially holds all the information of the image. $\mathbf{w}$ is 18×512 dimensions because the StyleGAN2 synthetic network consists of 18 layers. It is known that the image information that $\mathbf{w}$ differs depending on the layer. The lower layers have more global information, such as face shape and orientation, hairstyles, and facial expressions, while the higher layers have more detailed information, such as color schemes and lighting [Karras et al. 2019].

We used StyleGAN2 trained on the Flickr-Faces-HQ (FFHQ) dataset [Karras et al. 2019] for the generation of human face images. FFHQ is a dataset that consists of 70,000 images of human faces with 1024×1024 resolution.

3.2 Measurement of human perceptual behavior

In the change perception task (Fig. 1 - C), which investigated human perception, we focused on the perceptual visual comparison ability to perceive visual changes after movement in the latent space as an evaluation index. To investigate this change perception, we introduced a short visual mask, which suppresses motion perception, as a paradigm of change blindness [Rensink 2005]. In the initial process of visual information processing, almost all visual information is first stored in the iconic memory register, which has the characteristic of being retained for a short period of time, approximately 300 ms [Haber and Standing 1969]. After that, the information undergoes pattern recognition processing and is transferred to the visual short-term memory [Atkinson and Shiffrin 1968]. In the change blindness paradigm, since motion perception is suppressed by the mask, which should be a cue for change perception [Kanai and Verstraten 2004], the iconic memory can be overwritten with the target visual information (since there is no retention gain from attention), and thus, the change can be hardly founded [Rensink 2002; Rensink et al. 1997]. By exploiting this human perceptual behavior, i.e., perceptual metric, the intensity of face change perception independent of motion perception can be determined by examining the CP rate (how many participants could perceive the change) in the change perception task relative to the amount of image change. We note here that although the flicker paradigm has been also employed for change blindness research [Rensink et al. 1997], we employed the one-shot paradigm [Rensink 2005] for our experiment to get a percent correct score from human responses of participants, and to eliminate the potential strategy of directing attention, such as ordering attention to potential options.

In our perception experiment, four face images were displayed while participants gazed at the center of the display. After 66 ms of a visual blank, one of the four face images is changed corresponding to a certain distance in the latent space. Participants are asked to answer which face was changed by pressing the respective key. Then, we investigated the correspondence between the change detection rate, in which humans perceive the change, and the distance $\Delta\mathbf{w}$ traveled in the latent space.

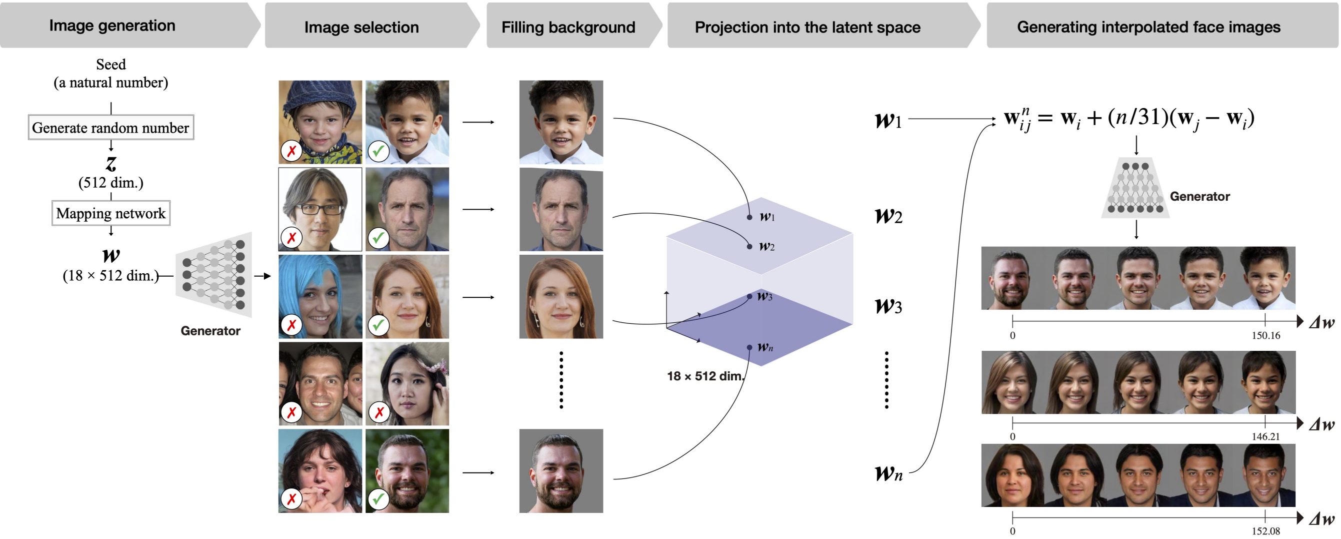


Figure 2: Overview of dataset generation procedure. Firstly, w is generated from the seed and images are generated from w through the StyleGAN2 generator. Inappropriate images for morphing are excluded. Secondly, after filling the background of the selected images with gray, the images are projected into the latent space (w of the image is obtained). Finally, from the pair of ws, the intermediate sequence of ws is calculated by linearly interpolating between ws, in accordance with the formula, and then a morphing image sequence is generated. The distance between the pair of ws differs for each pair.

3.3 Measurement of human cognitive behavior

In the face recognition task (Fig. 1 - E), we investigated the cognitive behavior and whether the participants would recognize a certain face as the same face after it has moved in the latent space. Our experimental procedure was based on the Exposure-Based Face Memory Test [Shen et al. 2019] and, in addition, the memory game sequence used by Isola et al. [Isola et al. 2011] with some modifications for short-term memory. In the face recognition task, the participants observed an image sequence in which a large number of face images are presented sequentially. After presenting a target face in the sequence, we examined whether participants could recognize the face image presented again after a temporal interval of more than one second, with the presentation of a different face as a destructive stimulus. When the target face image was presented again, the corresponding latent vector moved $\Delta\mathbf{w}$ from the initial presentation. We then investigated whether the face was recognized as known or as a new face. Here, unlike the change perception task, this task requires the cognitive process for a comparison of recognized facial information. Based on this human cognitive behavior, the intensity of face recognition, including cognitive processing, i.e., cognitive metric, can be obtained by examining the face recognition rate against the amount of image change that corresponds with $\Delta\mathbf{w}$.

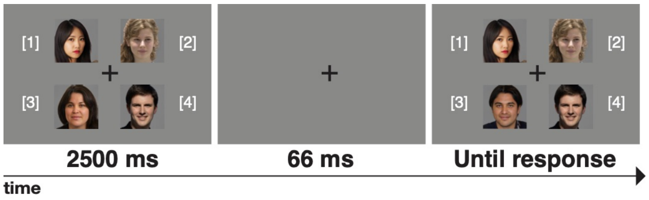


Figure 3: Change perception task for investigating perceptual behavior.

Because each of the stimulus images cooorespond to their respective latent vectors, it allows us to align the results of the human responses to the image changing to the distance in the latent space. The two tasks investigate the corresponding human perceptual and cognitive metrics with machine-learned latent space for face images. In our study, although the two tasks have different experimental paradigms, we performed them using the same face image dataset derived from the same latent space. This enables us to interpret the correspondence of both the CP rate and face recognition rate by $\Delta\mathbf{w}$.

3.4 Dataset

For complete alignment of the results of human responses to image changes with distances in the latent space, in our study, we used realistic face image stimuli with a uniform gray background, which corresponds perfectly to their respective **w**s. The overview of the dataset generation is shown in Fig. 2.

First, random numbers were generated from a seed (natural number), and **w** was generated from random numbers used as **z**. StyleGAN2 uses a technique called the truncation trick to improve the quality of images that exist in a sparse area in the latent space.

$$\mathbf{w}' = \bar{\mathbf{w}} + \varphi(\mathbf{w} - \bar{\mathbf{w}})$$

$\bar{\mathbf{w}}$ indicates the center of gravity of the latent space, and the style scale φ indicates how close **w** is to $\bar{\mathbf{w}}$ (**w** is exactly the same as $\bar{\mathbf{w}}$ at $\varphi = 0$). This trick can improve the quality of the image generated from (**w** at the expense of the variety of imagery. This is because the

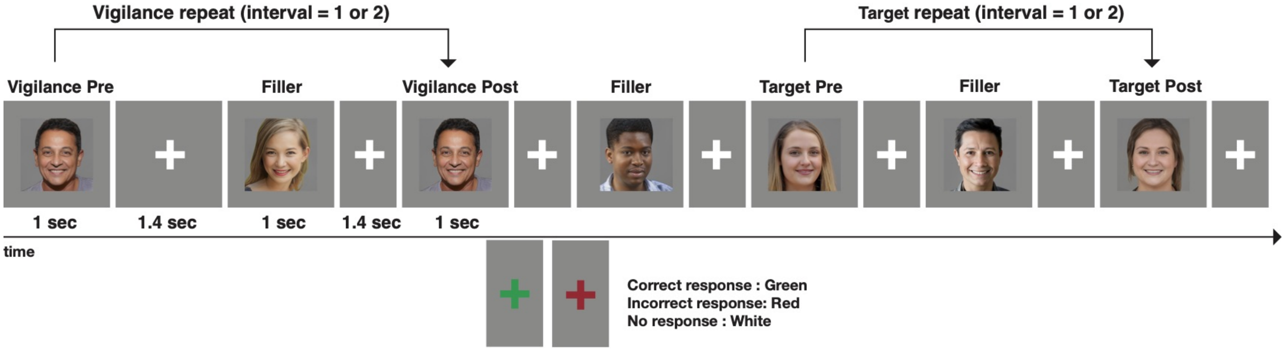


Figure 4: Face recognition task for investigating cognitive behavior.

center of gravity of the latent space is dense (well learned). We set $\varphi = 0.8$. An image was generated by inputting $\mathbf{w}'$ into the synthetic network.

The FFHQ dataset contains a myriad of images of people, including those wearing accessories such as hats, eyeglasses, and earrings. If these images are included in the morphing process, the accessories turn out deformed, generating unnatural images. Therefore, an open-source library [zllrunning 2019] based on a facial parsing method [Yu et al. 2018] was used to exclude images including sunglasses and hats. Additionally, images that contained multiple faces, faces occluded by hands or microphones, artifacts, flashy hair color, and makeup were manually removed.

To prevent the participants using background imagery as context clues for detecting changes or memorizing face images, the facial parsing method was used again to create a uniform gray background. At this time, if the background or face area was divided into multiple areas, the image was discarded.

The images were generated from $\mathbf{w}$ by StyleGAN2, but since the background was filled, it was necessary to project the image back into the latent space. For the projection, we used an open-source library [rolux 2019]. With this re-projection process of the gray background filled image, we can establish the interchangeability between face images and $\mathbf{w}$s.

500 images that passed the selection process were used for morphing. 250 pairs were randomly selected and interpolated face images were generated for each pair. Of each pair, 31 $\mathbf{w}$s was calculated in accordance with the following formula, and 31 images were generated.

$$\mathbf{w}_{ij}^{n} = \mathbf{w}_i + \frac{n}{30}(\mathbf{w}_j - \mathbf{w}_i)$$

$\mathbf{w}_i$ and $\mathbf{w}_j$ are a pair, and n is a natural number from 0 to 30. Of the 31 $\mathbf{w}$s, $\mathbf{w}$s at both ends are equal to $\mathbf{w}_i$ and $\mathbf{w}_j$. Within each pair, $\Delta\mathbf{w}$s among the images were uniform. However, $(\Delta\mathbf{w})$ changes for each pair.

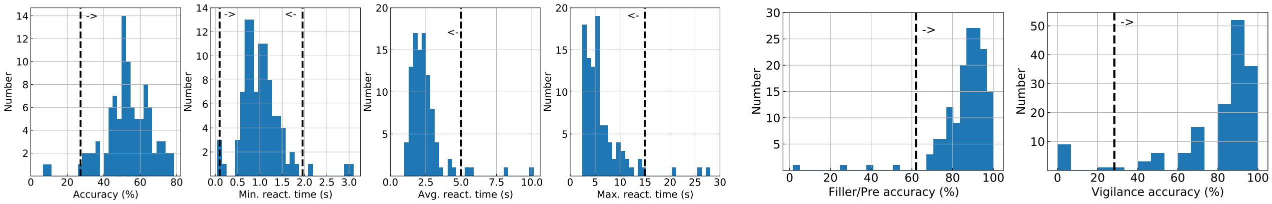
3.5 Experiment Procedure

This study was approved by our local ethics committee (Sony Group Corporation, Application 21-F-0039). All online participants gave their informed consent prior to participating in the study.

3.5.1 Change perception task. In the change perception task, our objective was to investigate the CP rate (how many participants could perceive the change) with different degrees of visual change depending on $\Delta\mathbf{w}$. We implemented the change perception task in a one-shot paradigm [Rensink 2005] where participants were instructed to report whether they noticed a change between sequential images. Participants started a trial by pressing '9' on the keyboard, and then four images (PRE) were displayed on the screen with a 50 % gray background. A 50 % gray blank interval was inserted for 66 ms, followed by another display of four images, which included a target image that had been visually changed (POST). The position of the target image is randomized between the four images to ensure that the participant is engaged in the task as designed. The duration of the interval was set according to previous literature, in which the detection change rate was reported to converge with a visual blank of 66 ms or longer [Kasahara and Takada 2021]. Each participant in 1 experiment received the same $\Delta\mathbf{w}$, but with different generated faces. From the aforementioned generated images, we used $\mathbf{w}_0$ for the target image in PRE and $\mathbf{w}_n (1 \leq n \leq 30)$ for the target image in POST. The PRE and POST pairs of all target face images and filler face images were synthesized with generated $\mathbf{w}$s from unique random seed values.

Participants were instructed to press the corresponding key '1'–'4' when they perceived a change. They were also asked to focus on the fixation cross mark throughout all trials. Here, our main focus here was the explicit perception of change. Considering that previous studies have shown that an undetectable stimulus can affect subsequent decisions in forced choice tasks [Laloyaux et al. 2008], to avoid subliminal effects of blinded images, we also asked participants to press '0' when they could not detect any change among the four images.

3.5.2 Face recognition task. In the face recognition task, our aim was to investigate how humans recognize an nearly identical face even with slight visual changes depending on $\Delta\mathbf{w}$. We implemented the face recognition task with different degrees of visual change based on the memory game by [Isola et al. 2013]. The participants viewed a sequence of images, each of which was displayed for 1 second, with a 1.4 second interval between image presentations (Figure 4). The participants were then instructed to press '1' whenever they recognized a face image they had previously seen, anytime in the image sequence. Participants received feedback whenever they pressed a key (a green cross shown at the center of the screen for correct detection, and a red cross is shown for an error). The



(a) change perception task

(b) face recognition

Figure 5: Exclusion of outliers from the experimental data. Each figure shows the histogram of each evaluation index, and the vertical dotted line shows the range of 2σ. Data shown in the direction of the arrow in each graph was used for the analysis. Each y-axis represents the number of participants.

sequence of face images comprised 'targets' (40 images) and 'vigilance' (20 images) and 'fillers' (120+ images).

The target images are a pair of PRE and POST images (target repeat), where the PRE image is the generated image from $\mathbf{w}_0$ and the POST image is the generated image from $\mathbf{w}_n (1 \leq n \leq 30)$. In the image sequence, after the PRE image is displayed, the POST image is displayed one or two images later. Our main interest is the relationship between $\Delta\mathbf{w}$ and whether the POST image is reaffirmed as being "recognized previously" in the image sequence. Each participant in 1 experiment received the same $\Delta\mathbf{w}$, but with different generated faces.

To determine whether participants were attentive to the task, we also included vigilance tasks using vigilance images. The vigilance image pairs (vigilance repeats) each use the same face image for both PRE and POST to ensure that participants are attentive to the experiment by giving a recognition response to the POST vigilance image. In each experiment, ten pairs of vigilance images were added. We excluded data of participants who were not attentive to the task.

In addition to the target and vigilance image repeats, filler images are used to generate the PRE-POST spacing between each corresponding pair. In both repeats, the post-face appears after 1 or 2 intervals of filler faces, which are also all unique. Note that the target, vigilance, and filler images are visually indistinguishable, since they are generated by the exact same procedure previously described. Participants are not informed of the rules and mechanisms of the experiment and are simply instructed to press '1' when presented with an image they feel to have seen previously.

3.6 Participants

We recruited 98 participants for the change perception task and 152 participants for the face recognition task through an online subject recruitment platform (https://prolific.co/). We invited participants who were 18 to 40 years old, had 90% or higher task approved rates in other online tasks, and experienced at least ten online tasks, but less than 10,000. Participants participated in change perception or face recognition tasks. The demographics of all participants are as follows: 42.15% self-reported females and 57.85% self-reported males. Age was within the range of 26.3 ± 5.6. The nationalities and current countries of residence are shown in Tables **??** and **??**, respectively.

Before beginning the main experiment, participants performed practice trials to verify their understanding of the trial procedure. We excluded 24 of the participants from the main analysis on the basis of the criterion explained in the "Data preprocessing and analysis" section (screening rate = 9.6%). All participants received monetary compensation for their participation (1 euro).

3.7 Data preprocessing and analysis

Since we recruited participants via an online subject recruitment platform, we strictly excluded outliers to prevent any effect from participants with short attention spans and local optimization for the tasks. Data selection criteria for the change perception and face recognition tasks are as follows. change perception task (Fig. 5a): 2940 data entries (98 participants, 30 trials for each participant) were collected. Eighty-eight data entries of $\Delta\mathbf{w} = 0$ were excluded. All participant trials that were outside the range of ($2\ \sigma$) (the mean ±2 times of the standard deviation) in any one of the accuracies (correct answer rate for the entire sequence of each participant) were excluded. The minimum / mean / maximum reaction time was also excluded because the accuracy criteria were applied only to the lower limit. These criteria removed ten participants. Finally, 2560 data entries (88 participants) were obtained.

Face recognition task (Fig. 5b): 3040 data entries (152 participants, 20 trials for each participant) were collected. We excluded all data of participants whose accuracy for filler/vigilance PREs/target PREs or for vigilance POSTs were not within the range of $2\ \sigma$ (both have only the lower limit), as we deem a potential issue in the participant's ability to perform the experiment or the subject's attention to the task. Fourteen participants were excluded, and, 2760 data entries (138 participants) were acquired.

4 RESULTS

The binary dot plots in Fig. 6 show the raw data of human responses in the change perception (left) and face recognition (right) tasks in a binary format. Note that all (2560 for change perception, 2760 for face recognition) data entries are displayed in each graph. To produce the general trends of change perception and facial recognition depending on $\Delta\mathbf{w}$, we performed the analysis with the fitting logistic function to model the relationship between human responses and $\Delta\mathbf{w}$. To analyze the trend of the large number of response data with a continuous scale of $\Delta\mathbf{w}$, we discretized a continuous $\Delta\mathbf{w}$ value into 13 bin based on the Sturges rule.

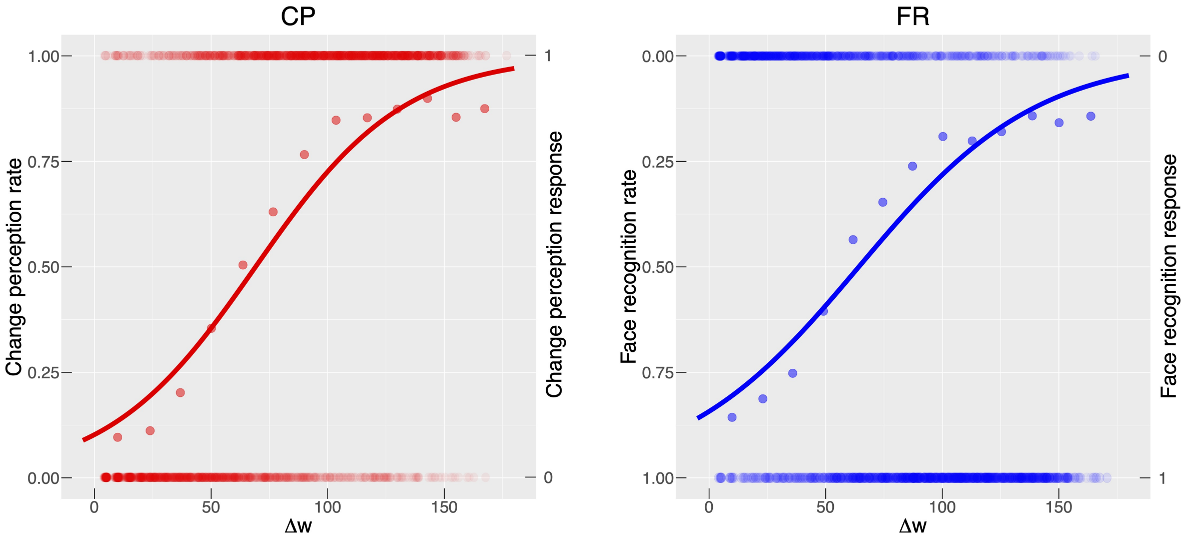


Figure 6: Raw data from the human response data in the change perception (left) and the face recognition (right) tasks (CP response: change perceived = 1, not perceived = 0, and FR response: face recognized as seen = 1, not recognized = 0). The curve shows results of logistic regression in the CP rate (left) and FR rate (right). The x-axis is the discretized Δw. The Dot plots represent the averaged value of the change perception rate in each discretized bin value. The solid line curve shows the logistic regression curve to model the human response $R^2 = 0.9185$ in the CP rate (left) and $R^2 = 0.9183$ in FR rate (right).

4.1 Change Perception Study

The left of Fig. 6 is a plot where the distance traveled in the latent space $\Delta\mathbf{w}$ of the face image in the latent space is on the x axis, and the averaged value of the discretized bin of the responses of the participants (perceived change = 1, not perception = 0) is the CP rate on the y-axis in the change perception task. As a result of logistic regression (solid line in Fig. 6), we can observe a clear relationship between $\Delta\mathbf{w}$ and the change perception rate. McFadden's R-Squared value, which is to assess how well a model fits the data, was $R^2 = 0.9185$ Since the CP rate reflects the change perception in the state where the motion perception is suppressed by the change blindness paradigm, it was shown that $\Delta\mathbf{w}$ and the human perceptual distance have a clear positive relationship that can be expressed by the logistic function. In other words, being able to perceive change more robustly indicates that the perceptual distance is longer.

4.2 Face Cognition Study

In the face recognition task, Fig. 6 (right) shows the plot with $\Delta\mathbf{w}$ between the pre- and post-target image on the x-axis and the discretized bin averaged value of the participants' answers (face recognized as seen = 1, not recognized = 0) as the face recognition rate on the y-axis. We also performed logistic regression (solid line in Fig. 6). As a result, we can observe a clear relationship between $\Delta\mathbf{w}$ and the face recognition rate. McFadden's R-Squared value was $R^2 = 0.9183$. Since the face recognition rate reflects the recognition of whether or not humans can recognize the same person in the face recognition paradigm, it shows that $\Delta\mathbf{w}$ and the human cognitive distance have a clear positive relationship that can be expressed by the logistic function. Furthermore, the higher the *inability* to recognize, the longer the cognitive distance. To be consistent with the result of the change perception task, the face recognition rate is reversed as shown in [1.0–0.0] in Fig. 6 (right).

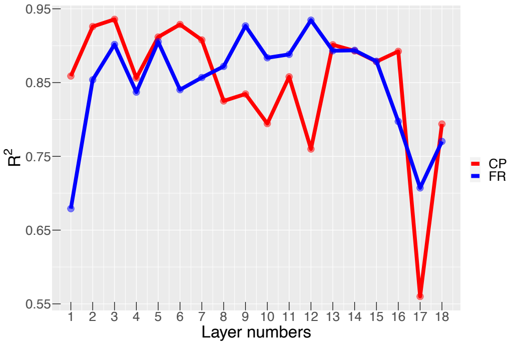


Figure 7: R^2 values of logistic regression of each layer in face recognition and change perception tasks.

Our result revealed that the perceptual distance of the face image defined as the change perception rate and the cognitive distance of the face image defined as the face recognition rate can be modeled by the logistic function with the distance $\Delta\mathbf{w}$ in the same latent space.

4.3 Additional analysis

Furthermore, we investigated the relationship between the internal structure of the GAN's latent vector and the human perception and cognitive response. The latent vector is a matrix with a size of 18 × 512. Each layer has been found to contribute different visual information [Karras et al. 2019]. Using the same data for both tasks,

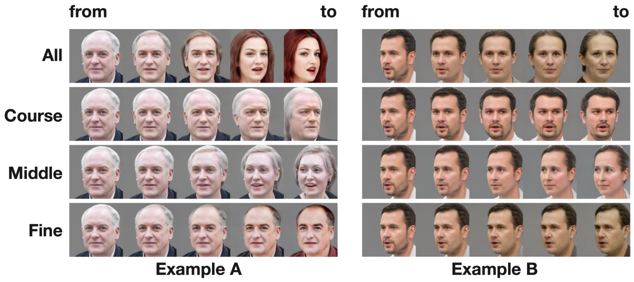


Figure 8: Two examples of image morphing from a "From" image to a "To" image with only specific layers. Each row represents the layers in the latent vector that were interpolated: "All layers:" 1–18, "Coarse layers:" 1–4, "Middle layers:" 5–8, and "Fine layers:" 9–18.

we performed logistic regression with the $\Delta\mathbf{w}$ of each layer (each respective 512-dimensional vector) and the participants' answers (CP rate for the change perception task and FR rate for the face recognition task). The R^2 value for each of the 18 layers for face recognition and change perception tasks is shown in Fig. 7. Different properties were observed in the change perception and face recognition tasks. This result suggests that change perception and face recognition correspond to different parts of the latent vector, even in the same latent space.

5 DISCUSSION

In the regression error plot for each layer (Fig. 7), different trends are observed between the change perception and face recognition tasks. For the change perception task, the general trend is that the lower the layer, the larger the R^2 value. This means that a vector in a lower layer has a stronger relationship with the CP rate, suggesting that lower layers better express the behavior in human change perception. StyleGAN [Karras et al. 2019] reported that the images generated from each layer inherited different subsets of visual aspects, such as pose or shape. We investigated the trend of visual changes with each layer in our material (Fig 8) in which we used StyleGAN2. It shows the clear trends that the Coarse layers (1–4) correspond to shape and posture, the Middle layers (5–8) correspond to physiognomy, and the Fine layers (9–18) correspond to the color and texture of the face image. The fact that the Coarse and Middle layers have latent image features that bring about greater visual changes is consistent with the result that the R^2 value in CP was larger in those layers. Unlike CP, FR has a smaller R^2 value of the Coarse layers compared with the Middle layers, indicating that the lower layers have a weak relationship with the FR. Since the change in the Coarse layers contributes to the change in posture, the change in the image is large. In addition, the Middle layers contribute to facial identity. From these facts, it is shown that FR does not have a strong relationship with the change in posture but has a stronger relationship with changes in facial impression (physiognomy). The results suggest the possibility of using the same latent space to explain two different human visual processes (perception and cognition). By investigating these in more detail, we can potentially describe various human visual processes with one common latent space.

6 LIMITATION AND FUTURE WORK

The findings of our studies focus on facial images from StyleGAN2. However, we expect that the paradigm we conducted in this study can be used as a general procedure to investigate the correspondence between the various face generation model with latent space and the response of human perception and cognition. As previous research have pointed out limitations in the StyleGAN2 [Bermano et al. 2022], where it fails in domains where strong structure is not present, we envision that our methodology will provide the basic means to analyze latent spaces constructed by utilizing deep learning methods other than StyleGAN2, such as StyleGAN3 [Karras et al. 2021], and expand to related areas including 3D avatar creation such as MetaHuman Creator Tool by Epic Games [Siddiqui 2022]. Although the face recognition task was particularly relevant to short-term visual memory in this experiment, it is expected to investigate the cognitive distance in long-term memory by presenting the same procedure with a longer interval. In addition, while our studies are performed with face images, our proposed method can also be applied to other images such as landscapes, animals, objects, etc., in the case when those visual images can be described as a latent vector.

7 APPLICATION

When image processing is applied using latent space, using our model, it is possible to predict how humans are likely to perceive and recognize changes depending on the amount of modification. This allows the facial image to be modified to take into account how it will be perceived by others. At the same time, it also allows for accounting for how a facial image is perceived as unmodified, and designing images utilizing this property. For example, an image can be designed to be recognized as the same 60 percent of the time, while changing its age or gender using the latent space of GANs [Shen et al. 2020]. The proposed method also allows for protecting one's information and identity, as it allows for masking one's true facial identity, while still being recognizable. Recent research has shown the usage of GAN generated face imagery to mask facial biometrics from video conference applications that collect data, such as Zoom [Elsden et al. 2022].

8 CONCLUSION

In this study, we investigated the correspondence between the latent space generated by a GAN model and human perception and cognition through two psycho-visual tasks. In our tasks, we measured whether a human could perceive changes in or recognize face images. As a result, our model showed a strong relationship between the distance between two target face images in the latent space and human perception/cognition. The paradigm conducted in this study can be applied as a general procedure for investigating the correspondence between the latent space computationally generated by deep learning models and human perception/cognitive responses.

ACKNOWLEDGMENTS

This work was supported by JST Moonshot R&D Program (Grant Number JPMJMS2013). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agencies.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2stylegan++: How to edit the embedded images?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8296–8305.

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021. Only a matter of style: age transformation using a style-based regression model. *ACM Trans. Graph.* 40, 4 (July 2021), 1–12.

R C Atkinson and R M Shiffrin. 1968. Human memory: A proposed system and its control processes. *The psychology of learning and motivation: II.* 249 (1968).

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. 2018. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. arXiv:1811.10597 [cs.CV]

Amit H. Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Or Patashnik, and Daniel Cohen-Or. 2022. State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. https://doi.org/10.48550/ARXIV.2202.14020

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH ’99).* ACM Press/Addison-Wesley Publishing Co., USA, 187–194.

J Deng, W Dong, R Socher, L Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition.* 248–255.

Alexey Dosovitskiy and Thomas Brox. 2016. Generating Images with Perceptual Similarity Metrics based on Deep Networks. *arXiv* (Feb. 2016). arXiv:1602.02644 [cs.LG]

Bernhard Egger, William A P Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2019. 3D Morphable Face Models – Past, Present and Future. (Sept. 2019). arXiv:1909.01815 [cs.CV]

Chris Elsden, David Chatting, Michael Duggan, Andrew Carl Dwyer, and Pip Thornton. 2022. Zoom Obscura: Counterfunctional Design for Video-Conferencing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI ’22).* Association for Computing Machinery, New York, NY, USA, Article 143, 17 pages. https://doi.org/10.1145/3491102.3501973

Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 2020. Towards Photo-Realistic Facial Expression Manipulation. *Int. J. Comput. Vis.* 128, 10 (Nov. 2020), 2744–2761.

Gillian Rhodes , Andy Calder, Mark Johnson, and James V. Haxby. 2011. Oxford Handbook of Face Perception. In *Oxford Handbook of Face Perception* (1 ed.). Oxford University Press.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. 2019. GANalyze: Toward Visual Definitions of Cognitive Image Properties. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* 5743–5752.

Lore Goetschalckx, Alex Andonian, and Johan Wagemans. 2021. Generative adversarial networks unlock new methods for cognitive science. *Trends in Cognitive Sciences* 25, 9 (2021), 788–801. https://doi.org/10.1016/j.tics.2021.06.006

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Ralph Norman Haber and L. G. Standing. 1969. Direct Measures of Short-Term Visual Storage. *Quarterly Journal of Experimental Psychology* 21, 1 (1969), 43–54. https://doi.org/10.1080/14640746908400193 arXiv:https://doi.org/10.1080/14640746908400193

Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2013. What Makes a Photograph Memorable? *IEEE transactions on pattern analysis and machine intelligence* 36 (10 2013). https://doi.org/10.1109/TPAMI.2013.200

P Isola, J Xiao, A Torralba, and A Oliva. 2011. What makes an image memorable? *CVPR 2011* (2011).

Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. 2020. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5194–5202.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *arXiv* (March 2016). arXiv:1603.08155 [cs.CV]

Kamila M Jozwik, Jonathan O’Keeffe, Katherine R Storrs, Wenxuan Guo, Tal Golan, and Nikolaus Kriegeskorte. 2022. Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proc. Natl. Acad. Sci. U. S. A.* 119, 27 (July 2022), e2115047119.

Ryota Kanai and Frans A J Verstraten. 2004. Visual transients without feature changes are sufficient for the percept of a change. *Vision Res.* 44, 19 (2004), 2233–2240.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. In *Proc. NeurIPS.*

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4401–4410.

Shunichi Kasahara and Kazuma Takada. 2021. Stealth Updates of Visual Information by Leveraging Change Blindness and Computational Visual Morphing. *ACM Trans. Appl. Percept.* 18, 4 (Oct. 2021), 1–17.

Cédric Laloyaux, Christel Devue, Stéphane Doyen, Elodie David, and Axel Cleeremans. 2008. Undetected changes in visible stimuli influence subsequent decisions. *Conscious. Cogn.* 17, 3 (Sept. 2008), 646–656.

Steven J Luck and Edward K Vogel. 2013. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends Cogn. Sci.* 17, 8 (Aug. 2013), 391–400.

Yaniv Morgenstern, Frieder Hartmann, Filipp Schmidt, Henning Tiedemann, Eugen Prokott, Guido Maiello, and Roland W Fleming. 2020. An image-computable model of human visual shape similarity. (Jan. 2020), 2020.01.10.901876 pages.

Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. 2021. Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access)* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3115428

Ronald A Rensink. 2002. Change detection. *Annu. Rev. Psychol.* 53 (2002), 245–277.

Ronald A Rensink. 2005. CHAPTER 13 - Change Blindness. In *Neurobiology of Attention*, Laurent Itti, Geraint Rees, and John K Tsotsos (Eds.). Academic Press, Burlington, 76–81.

Ronald A Rensink, J Kevin O’Regan, and James J Clark. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychol. Sci.* 8, 5 (Sept. 1997), 368–373.

rolux. 2019. stylegan2encoder. https://github.com/rolux/stylegan2encoder.

Mark W Schurgin, John T Wixted, and Timothy F Brady. 2020. Psychophysical scaling reveals a unified theory of visual memory strength. *Nat Hum Behav* 4, 11 (Nov. 2020), 1156–1172.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2019. Interpreting the Latent Space of GANs for Semantic Face Editing. *arXiv* (July 2019). arXiv:1907.10786 [cs.CV]

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the Latent Space of GANs for Semantic Face Editing. In *CVPR.*

J. Rafid Siddiqui. 2022. FExGAN-Meta: Facial Expression Generation with Meta Humans. https://doi.org/10.48550/ARXIV.2203.05975

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* (Sept. 2014). arXiv:1409.1556 [cs.CV]

Gaeun Son, Dirk B. Walther, and Michael L. Mack. 2021. Scene wheels: Measuring perception and memory of real-world scenes with a continuous stimulus space. *bioRxiv* (2021). https://doi.org/10.1101/2020.10.09.333708 arXiv:https://www.biorxiv.org/content/early/2021/04/01/2020.10.09.333708.full.pdf

Tim Valentine, Michael B Lewis, and Peter J Hills. 2016. Face-space: A unifying concept in face recognition research. *Q. J. Exp. Psychol.* 69, 10 (Oct. 2016), 1996–2019.

Evangelos Ververas and Stefanos Zafeiriou. 2020. SliderGAN: Synthesizing Expressive Face Images by Sliding 3D Blendshape Parameters. *Int. J. Comput. Vis.* 128, 10 (Nov. 2020), 2629–2650.

Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. 2020. StyleGAN2 Distillation for Feed-Forward Image Manipulation. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 170–186.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV).* 325–341.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 586–595.

zllrunning. 2019. face-parsing.PyTorch. https://github.com/zllrunning/face-parsing.PyTorch.

9 APPENDIX

The dataset used for the experiments are available online [1].

[1] https://github.com/SuperceptionLab/HumanLatentMetrics