Human Latent Metrics: Perceptual and Cognitive Response Corresponds to Distance in GAN Latent Space

Shunichi Kasahara (kasahara@csl.sony.co.jp)
Sony Computer Science Laboratories, Inc
Naoto Ienaga
University of Tsukuba
Kye Shimizu
Sony Computer Science Laboratories, Inc
Kazuma Takada
Sony Computer Science Laboratories, Inc
Maki Sugimoto
Keio University

Research Article

Keywords:

Posted Date: February 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1339104/v1

Human Latent Metrics: Perceptual and Cognitive Response Corresponds to Distance in GAN Latent Space

Naoto Ienaga[1], Kye Shimizu[2], Kazuma Takada[2], Maki Sugimoto[3], and Shunichi Kasahara[2,*]

[1]Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, 305-8573, Japan
[2]Sony Computer Science Laboratories, Inc., Tokyo, 141-0022, Japan
[3]Department of Information and Computer Science, Keio University, Yokohama, 223-8522, Japan
[*]kasahara@csl.sony.co.jp

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, manipulate images using their inherit vector values in the latent space. This interchangeability of latent vector has the potential to calculate not only distance in the latent space, but also human perceptual and cognitive distance toward images, i.e., how humans perceive and recognize images. However, it is still unclear how the distance in the latent space corresponds to human perception and cognition. Our studies investigated the correspondence 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 (CP) task was utilized 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 cognition (FC) task was utilized to examine whether the participants could recognize a face as the same, even after moving an arbitrary distance in the latent space. The results showed that CP and cognition for face images clearly correlates to the distance in the latent space, which can be modeled with a logistic function. We also investigated how the internal layered structure of the latent space correlates to human response by calculating the regression residual error in each layer. As a result, we observed different residual error trends pertaining to CP and FC. Our experiments show that the distance between face images in the latent space corresponds to human perception and cognition for visual changes in face imagery, and additionally indicates that perception and cognition correspond with the latent space differently.
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.

Introduction

By projecting visual information into a vector space, we can computationally manipulate and edit visual information[1–3]. Generative adversarial networks (GANs), including StyleGAN[4], 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. An important property of the latent space is that low-level information such as texture and high-level information such as gender, facial expression, skin color, and posture for face images, can be represented as a vector[5–8]. 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[5,6,8]. Image manipulation by transforming the latent vector enables new image editing and analysis in which various features and attributes are continuously manipulated[9,10] (Fig. 1(A,B)).

However, the degree of correspondence between the generated latent space and human perception and cognition has been poorly investigated. In the training process of GANs[4], 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[11] trained on ImageNet[12] classification. Nevertheless, the correspondence between images generated from a latent space and the 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[13,14], to reflect human perceptual response, Learned Perceptual Image Patch Similarity (LPIPS) was trained from actual human responses through a psychophysical study[15], which outperforms all previous

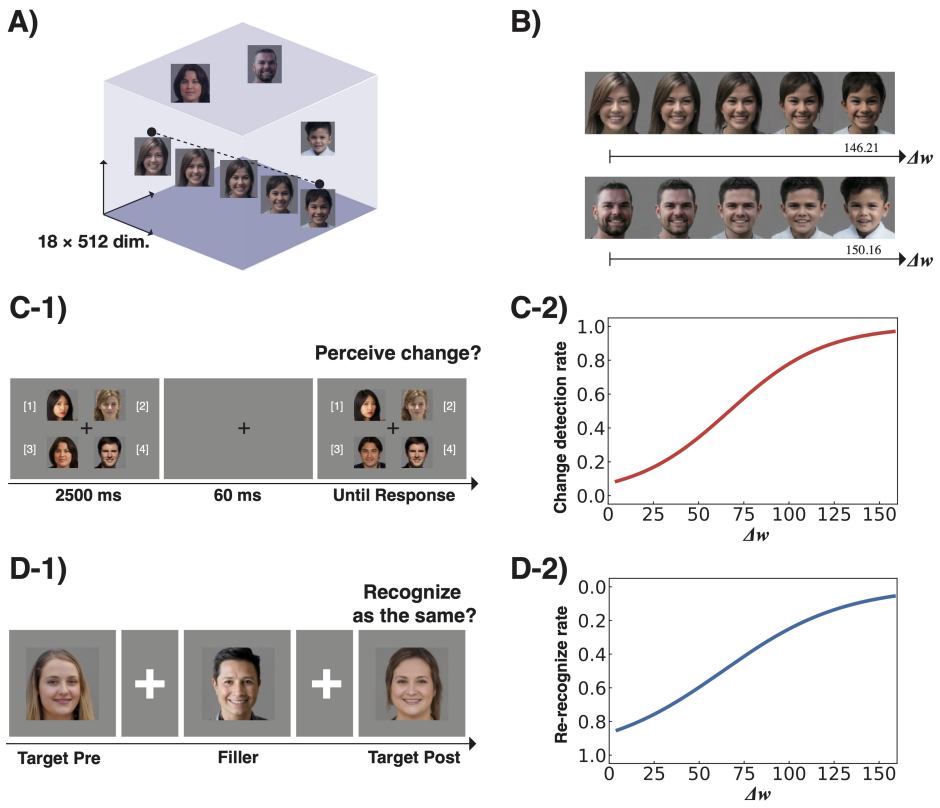


Figure 1. A) A conceptual schematic of a visual information arrangement in the latent space, where a vector in manifold representing a face image is constructed in a latent space of 18×512 dimensions. By interpolating the latent vectors along the vectors connecting arbitrary points in the latent space, 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 $\Delta\mathbf{w}$ increases. C-1) In a CP task, after 66 ms of a visual blank, one of the four face images is changed in accordance with $\Delta\mathbf{w}$. Participants are asked to answer which face was changed. C-2) We investigated the correspondence between the CP rate (participants who perceived the change) and $\Delta\mathbf{w}$. The results of logistic regression in CP clearly show the correlation between human CP and $\Delta\mathbf{w}$. D-1) In a FC 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 $\Delta\mathbf{w}$ 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. D-2) We also investigated the correspondence between face recognition rate (participants who recognized it to be the same face) and $\Delta\mathbf{w}$. The results of logistic regression in face recognition rate clearly shows the correlation between human FC and $\Delta\mathbf{w}$.

metrics. The aforementioned research indicate that acquiring better metrics to represent human perception also benefits various applications. 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[16]. Another study took advantage of GANs to continuously generate realistic images to examine the correlation 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[17]. 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[18]. 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[19,20]. Likewise, we hypothesize that generating continuous infinite face image stimuli (which have more complex structures) would enable us to investigate the relationship between human perception (Fig. 1(C)) and cognitive (Fig. 1(D)) response with a latent space.

Furthermore, building upon previous research that correlates trained networks with human perception and cognitive responses[17,21], we conducted visual psychological experiments using visual stimuli generated from latent vectors of GANs to measure human perceptual and cognitive responses to visual stimuli. The purpose of this study is to model the correspondence between the distance in the latent space to human perceptual and cognitive properties. A new metric, the corresponding

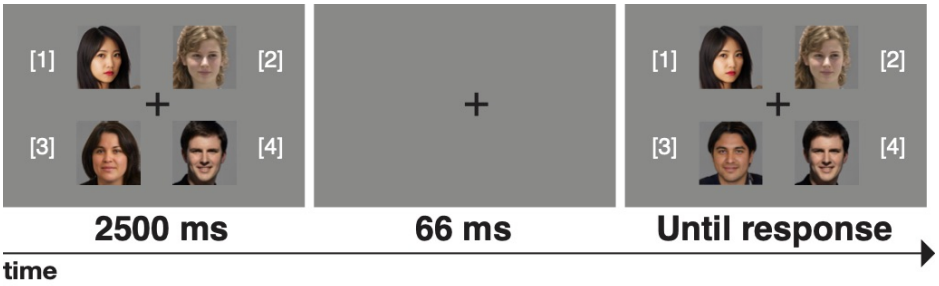


Figure 2. CP task for investigating perceptual behavior. We investigated the perception of face image changes that did not depend on the motion perception that occurred. Four face images were displayed first. After 66 ms of a visual blank, one of the four face images was changed by travelling a certain distance in the latent space. Participants are asked to answer which face was changed by pressing the respective key [1–4] or [0] if they deemed none were changed.

distance in the latent space, which is a common attribute found in all images unlike other properties, can be introduced when expressing the perceptual or cognitive distance of visual stimuli by using the results of this study. This will not only be useful for manipulating images when considering human perceptual and cognitive distances, but will also provide a common attribute for further cognitive science experiments. To investigate the relationship between the distance in the latent space and human visual perceptual and cognitive properties, we focused on face images, conducting both a change perception (CP) task, which is based on the change blindness paradigm[22], to investigate perception and a face cognition (FC) task, which is based on the Exposure Based Face Memory Test[23] and image memory game[24] to investigate cognition.

In the CP task (Fig. 2), 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 CP, we introduced a short visual occlusion as the change blindness paradigm[22], that is, the visual transient is concealed and motion perception is suppressed. 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[25]. After that, the information undergoes pattern recognition processing and is transferred to the visual short-term memory[26]. In the change blindness paradigm, since motion perception is suppressed by the occlusion of the visual transient, which should be a cue for CP[27], 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[28,29]. By exploiting this human perceptual behavior, i.e., perceptual metric, the intensity of face CP independent of motion perception can be determined by examining the CP rate in the CP task relative to the amount of image change. Specifically, four face images were displayed while the participants were gazing 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.

In the FC task (Fig. 3), 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[6] and in addition, the memory game sequence used by Isola et al.[24] with some modifications for short-term memory. In the FC task, participants observed an image sequence in which a large number of face images are presented sequentially. After a target face was presented in the sequence, we examined whether the 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 a new face. Here, unlike the CP task, this task requires the cognitive process for a comparison of recognized facial information. On the basis of 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 $\Delta\mathbf{w}$.

In our tasks, we use realistic face image stimuli including various facial expressions, hairstyles, and postures with a uniform gray background. Note that all of these stimulus images correspond perfectly to their respective latent vectors. This enables us to completely align the results of human responses to image changes with distances 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 $\Delta\mathbf{w}$.

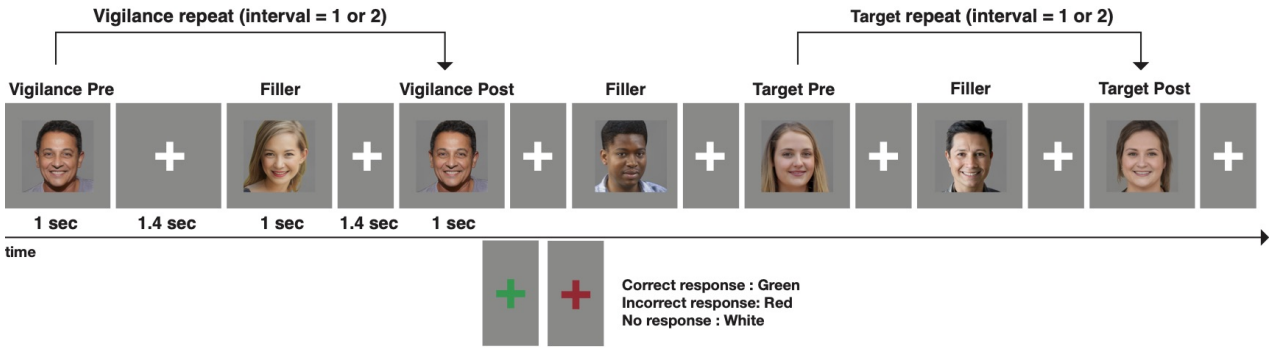


Figure 3. FC task for investigating cognitive behavior. Participants viewed a sequence of face images displayed for 1 second, with a 1.4 second interval. Participants were instructed to press '1' whenever they recognized a face image they had previously seen at any time in the image sequence. Participants received feedback whenever they pressed a key (a green cross for correct detection, and a red cross for an error). In the FC task, there are two type of repeat pairs: vigilance and target. The vigilance repeat, where pre- and post-face images are identical, is for assessing a participant's attention to the task since participants are expected to recognize them as images they had seen before. In target repeat, after a target pre-face is presented in the sequence, the target post-face image is presented again with the corresponding latent vector moved a certain distance in the latent space from the initial presentation. In both repeats, the post-face appears after 1 or 2 intervals of filler faces, which are also all unique.

Results

Change Perception Study

The left of Fig. 4 is a plot where the traveled distance in the latent space $\Delta\mathbf{w}$ of the face image in the latent space is on the x-axis, and the moving average (window size = 50) value of the participants' answers (change perceived = 1, not perceived = 0) as the CP rate on the y-axis in the CP task. As a result of logistic regression, a correlation between $\Delta\mathbf{w}$ and the CP rate was shown. The residual error between the moving average of the measured CP rate and the estimated value by logistic regression was evaluated by the root mean squared error (RMSE), which was 0.0680. Since the CP rate reflects the CP 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 correlation, and the relationship can be expressed by the logistic function. In other words, being able to perceive change more robustly indicates that the perceptual distance is longer.

Face Cognition Study

In the FC task, Fig. 4 - B shows the plot with $\Delta\mathbf{w}$ between the pre- and post-target image on the x-axis and the moving average (window size = 50) 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. 4 - B). As a result, a correlation between $\Delta\mathbf{w}$ and the face recognition rate was shown (RMSE $= 0.0774$). Since the face recognition rate reflects the recognition of whether or not human 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 correlation, and the relationship can be expressed by the logistic function. In addition, the higher the *inability* to recognize, the longer the cognitive distance. To be consistent with the result of the CP task, the face recognition rate is reversed as shown in [1.0–0.0] in the right of Fig. 4 - B.

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

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[4]. Using the same data for both tasks, we performed logistic regression with the $\Delta\mathbf{w}$ of each layer (each respective 512-dimensional vector) and the participants' answers (CP for the CP task and face recognition rate for the FC task). The RMSEs for each of the 18 layers for the CP and FC tasks are shown in Fig. 5. Different properties were observed in the CP and FC tasks. This result suggests that CP and face recognition correspond to different parts of the latent vector even in the same latent space.

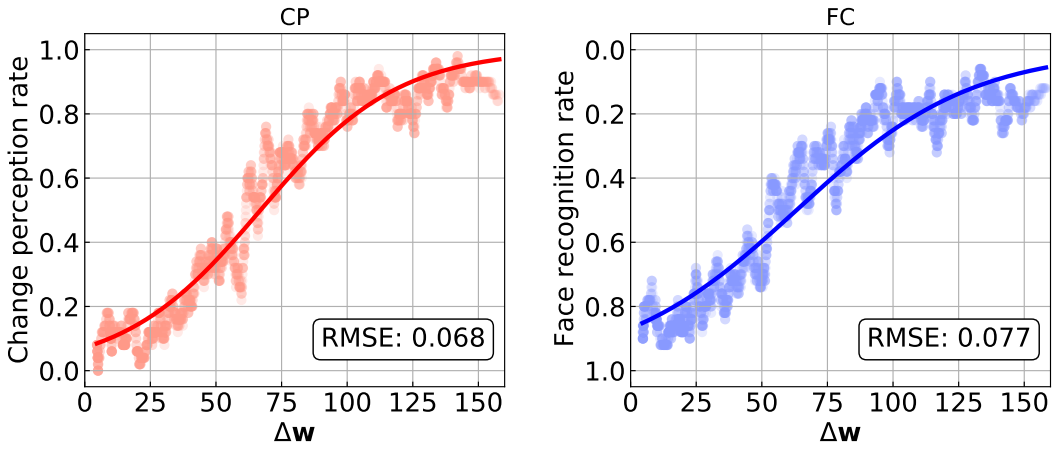


Figure 4. Results of logistic regression in the CP (left) and FC (right) tasks. The x-axis is **Δw** and the y-axis is the moving average of the participants' answers. Participant's answers are the CP rate for CP and the face recognition rate for FC. There is a clear positive correlation between **Δw** and the participants' answers.

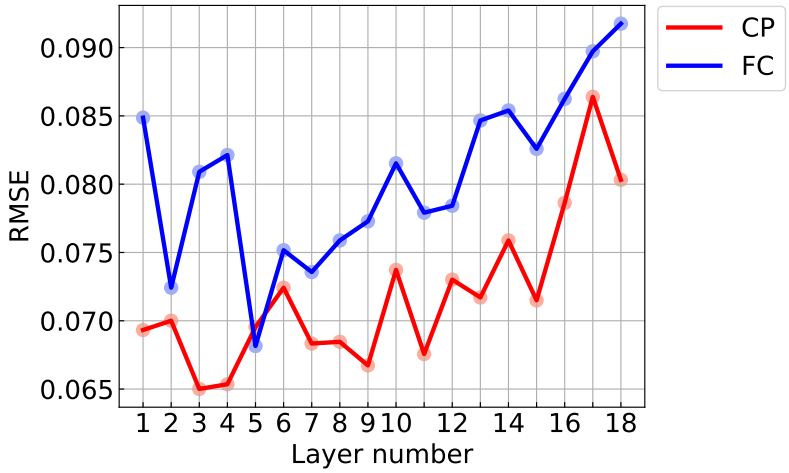


Figure 5. RMSE of logistic regression of each layer in the CP and FC tasks.

Discussion

In the regression error plot for each layer (Fig. 5), different trends are observed between the CP and FC tasks. For the CP task, the overall trend is that the lower the layer, the smaller the RMSE. This means that a vector in a lower layer has a higher correlation with the CP rate, which suggests that lower layers express behavior in human CP better. StyleGAN[4] reported that 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 6) in which we used StyleGAN2. It shows the clear trends that the Coarse layers (1–4) correspond to the 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 RMSE in CP was lower in those layers. Unlike CP, FC has a higher RMSE of the Coarse layers compared with the Middle layers, indicating that the lower layers have a weak correlation with the FC. 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 FC does not have a high correlation with the change in posture but has a high correlation 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.

The paradigm we conducted in this study can be used as a general procedure to investigate the correspondence between the

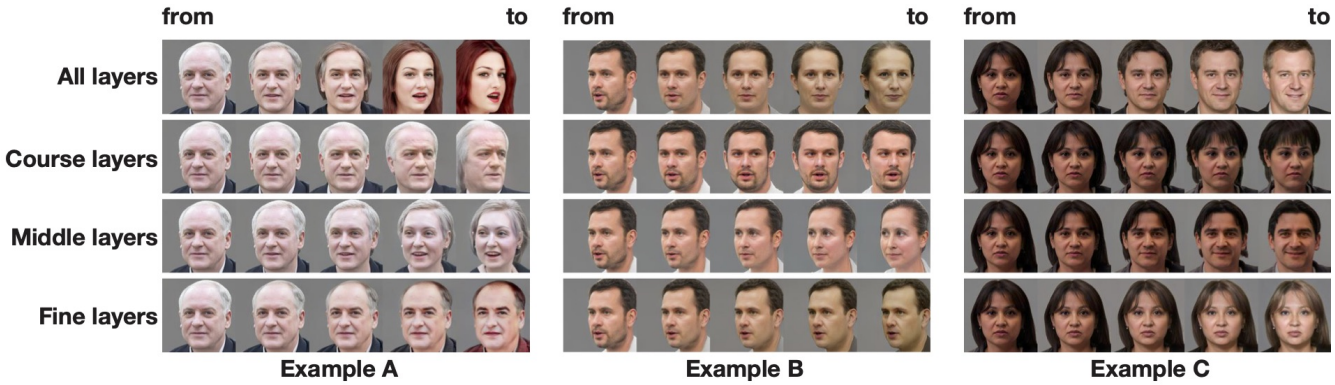


Figure 6. Three examples of image morphing from a "From" image to a "To" image with only specific layers. Each row represents the layers in the latent vector that were interpolated: "All layers:" 1–18, "Coarse layers:" 1–4, "Middle layers:" 5–8, and "Fine layers:" 9–18. There is a trend that different face properties are assigned to each layer; Coarse layers correspond to the shape and posture of the face, Middle layers correspond to physiognomy i.e. facial identity, and Fine layers correspond to color and texture of the face image.

latent space and the response of human perception and cognition. While our study focused only on face images, our proposed method can also be applied to other images such as landscapes, animals, objects, and so on. As a further research agenda, we envision that our methodology will provide the basic means to analyze latent spaces constructed by deep learning methods other than StyleGAN2 from the viewpoint of human visual characteristics. This study focused only on the distance (change) between two image targets. The characteristics of each image themselves, which are the prominence and memorability of the image itself, have not been considered.

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 high correlation 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.

Methods

Generative adversarial network

GANs are image generation models originally developed by Goodfellow et al. in 2014[30]. 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.

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. We used StyleGAN2[31], which is an improved version of StyleGAN[4], for this research. StyleGAN2 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. As evidence, it is known that various image editing is possible by manipulating $\mathbf{w}$ in the latent space[2]. $\mathbf{w}$ 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 higher layers have more detailed information such as color schemes and lighting[4].

We used StyleGAN2 trained on the Flickr-Faces-HQ (FFHQ) dataset[4] for human face image generation. FFHQ is a dataset that consists of 70,000 images of human faces with 1024×1024 resolution.

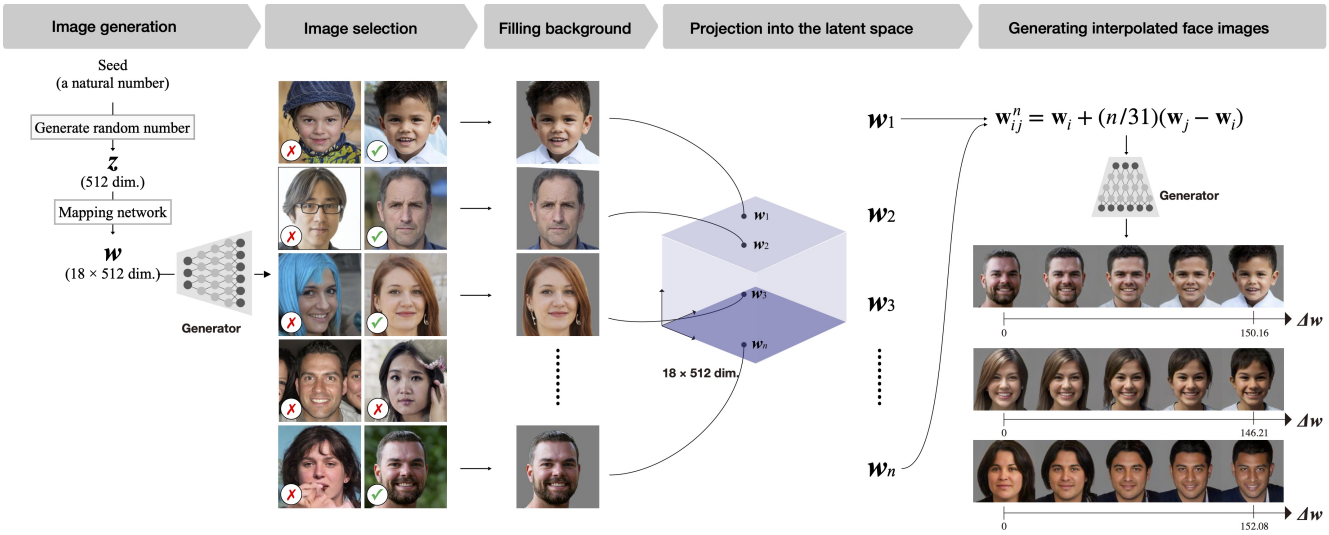


Figure 7. Overview of dataset generation procedure. Firstly, **w** is generated from the seed and images are generated from **w** through the StyleGAN2 generator. Inappropriate images for morphing are excluded. Secondly, after filling the background of the selected images with gray, the images are projected into the latent space (**w** of the image is obtained). Finally, from the pair of **w**s, the intermediate sequence of **w**s is calculated by linearly interpolating between **w**s, in accordance with the formula, and then a morphing image sequence is generated. The distance between the pair of **w**s differs for each pair.

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. 7.

Image generation

Random numbers were generated from a seed (natural number), and **w** was generated from the 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 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.

Image selection

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[32] based on a facial parsing method[33] was used to exclude images including sunglasses and hats. In addition, images that contained multiple faces, faces occluded by hands or microphones, artifacts, flashy hair color and makeup were manually removed.

Filling background

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.

Projection into the latent space

The images were generated from **w** by StyleGAN2, but since the background was filled, it was necessary to project the image into the latent space again. For the projection, we used an open-source library[34]. With this re-projection process of the gray-background-filled image, we can establish the interchangeability between face images and **w**s.

Generating interpolated face images

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. From each pair, 31 **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 **w**s, **w**s at both ends are equal to $\mathbf{w}_i$ and $\mathbf{w}_j$. Within each pair, Δ**w**s among the images were uniform. However, (Δ**w**) changes for each pair.

Experiment Procedure

This study was approved by our local ethics committee (Sony Group Corporation, Application 19F00XX). All online participants provided informed consent before participating in the study.

Change perception task

In the CP task, our aim was to investigate the CP rate with different degrees of visual change depending on Δ**w**. We implemented the CP task in a one-shot paradigm[22] 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 interval duration was set in accordance with previous literature, in which the detection change rate was reported to converge with a visual blank of 66 ms or longer[35]. 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 **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 undetectable stimulus can affect subsequent decisions in forced-choice tasks[36], to avoid any subliminal effects of blinded images, we also asked participants to press '0' when they could not detect any change among the four images.

Face cognition task

In the FC task, our aim was to investigate how humans recognize an nearly identical face even with slight visual changes depending on Δ**w**. We implemented the FC task with different degrees of visual change on the basis of the memory game by[37]. Participants viewed a sequence of images, each of which was displayed for 1 second, with a 1.4 second interval in between image presentations (Figure 3). Then, participants were instructed to press '1' whenever they recognized a face image they had seen previously, 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 for an error). The 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, 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 correlation between Δ**w** and whether the POST image is reaffirmed as being "recognized previously" in the image sequence.

To determine whether participants were attentive to the task, we also included vigilance tasks using vigilance images. The vigilance image pairs 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 images, filler images are used to generate the PRE-POST spacing between each corresponding pair. Note that target, vigilance, and filler images are visually indistinguishable as they are generated through the exact same procedure previously described. The 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 had seen previously.

Participants

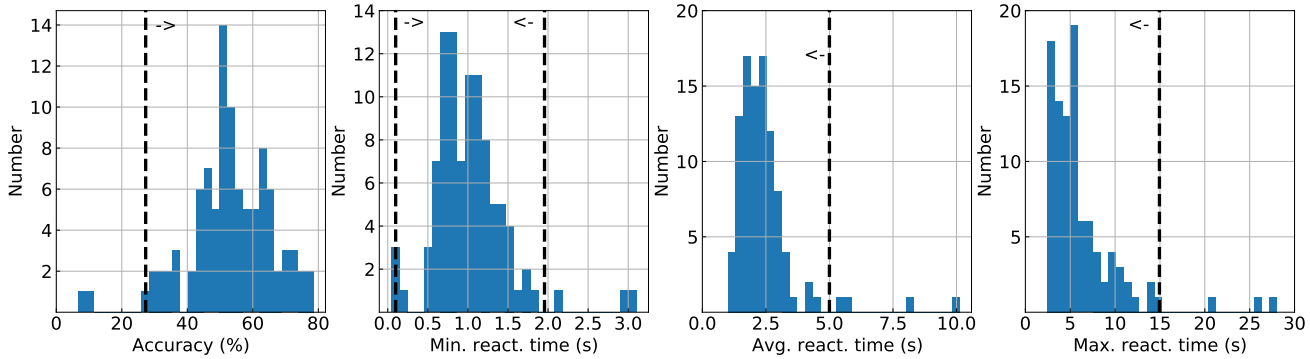
We recruited 98 participants for the CP task and 152 participants for the FC task through an online subject recruitment platform (https://prolific.co/). We invited participants who were 18 to 40 years old, had 90% or higher task approved rates in other online tasks, and experienced at least ten online tasks, but less than 10,000. Participants joined either the CP or FC tasks. The demographics of all participants are as follows: 42.15% self-reported female and 57.85% self-reported male. Age was within

Table 1. Top 10 nationalities. The total number of nationalities was 31.

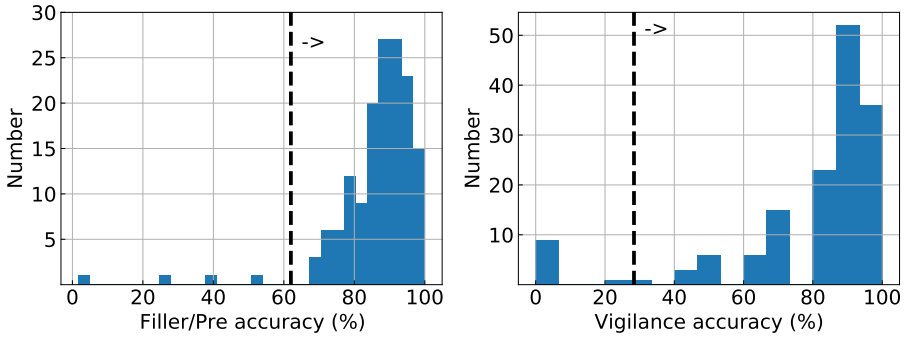
1	United States	22.42%
2	Poland	10.76%
3	Mexico / United Kingdom	9.87%
5	Portugal	8.97%
6	South Africa	7.17%
7	Italy	4.93%
8	Netherlands	3.14%
9	Canada / Hungary	2.69%

Table 2. Top 10 current countries of residence. The total number of countries was 24.

1	United States	21.52%
2	United Kingdom	13.00%
3	Poland	10.31%
4	Mexico / Portugal	9.87%
6	South Africa	7.62%
7	Italy	5.83%
8	Canada / Netherlands	2.69%
10	France / Hungary	2.24%



(a) CP



(b) FC

Figure 8. Exclusion of outliers from experimental data. Each figure shows the histogram of each evaluation index, and the dotted vertical line shows the range of 2σ. The data shown in the direction of the arrow was used for the analysis. Each y-axis represents the number of participants.

the range of 26.3 ± 5.6. Nationalities and current countries of residence are shown in Table 1 and 2, 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).

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 CP and FC tasks are as follows.

CP task (Fig. 8a): 2940 data entries (98 participants, 30 trials for each participant) were collected. Eighty-eight data entries of $\Delta \mathbf{w} = 0$ were excluded, because when $\Delta \mathbf{w} = 0$, the correct answer is never known. All participant trials that were out of the range of (2 σ) (the mean ± 2 times of the standard deviation) in any one of the accuracies (correct answer rate for each

participant's entire sequence) were excluded. The minimum/mean/maximum reaction time were 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.

FC task (Fig. 8b): 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 σ (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.

To analyze the general trends of the CP or face recognition rates depending on $\Delta\mathbf{w}$, the data was smoothed using a simple moving average method with a window size of 50 and then fitted with the logistic function.

Ethical approval

All procedures were approved by the Sony Group Corporation,Ethics Committee and in accordance with the 1964 Helsinki declaration. We did not collect any information to identify participants.

Data availability

The datasets generated and analysed during the current study are available at https://github.com/SuperceptionLab/Human-Latent-Metrics .

References

1. Pan, X. *et al.* Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis Mach. Intell. (Early Access)* 1–1, DOI: 10.1109/TPAMI.2021.3115428 (2021).
2. Abdal, R., Qin, Y. & Wonka, P. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8296–8305 (2020).
3. Viazovetskyi, Y., Ivashkin, V. & Kashin, E. Stylegan2 distillation for feed-forward image manipulation. In Vedaldi, A., Bischof, H., Brox, T. & Frahm, J.-M. (eds.) *Computer Vision – ECCV 2020*, 170–186 (Springer International Publishing, Cham, 2020).
4. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410 (2019).
5. Ververas, E. & Zafeiriou, S. SliderGAN: Synthesizing expressive face images by sliding 3D blendshape parameters. *Int. J. Comput. Vis.* **128**, 2629–2650 (2020).
6. Shen, Y., Gu, J., Tang, X. & Zhou, B. Interpreting the latent space of GANs for semantic face editing. *arXiv* (2019). 1907.10786.
7. Alaluf, Y., Patashnik, O. & Cohen-Or, D. Only a matter of style: age transformation using a style-based regression model. *ACM Trans. Graph.* **40**, 1–12 (2021).
8. Geng, Z., Cao, C. & Tulyakov, S. Towards Photo-Realistic facial expression manipulation. *Int. J. Comput. Vis.* **128**, 2744–2761 (2020).
9. Jiang, W. *et al.* 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 (2020).
10. Goetschalckx, L., Andonian, A., Oliva, A. & Isola, P. GANalyze: Toward visual definitions of cognitive image properties. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5743–5752 (2019).
11. Simonyan, K. & Zisserman, A. Very deep convolutional networks for Large-Scale image recognition. *arXiv* (2014). 1409.1556.
12. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).
13. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for Real-Time style transfer and Super-Resolution. *arXiv* (2016). 1603.08155.
14. Dosovitskiy, A. & Brox, T. Generating images with perceptual similarity metrics based on deep networks. *arXiv* (2016). 1602.02644.

15. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595 (2018).

16. Bau, D. *et al.* Gan dissection: Visualizing and understanding generative adversarial networks (2018). 1811.10597.

17. Son, G., Walther, D. B. & Mack, M. L. Scene wheels: Measuring perception and memory of real-world scenes with a continuous stimulus space. *bioRxiv* DOI: 10.1101/2020.10.09.333708 (2021). https://www.biorxiv.org/content/early/2021/04/01/2020.10.09.333708.full.pdf.

18. Goetschalckx, L., Andonian, A. & Wagemans, J. Generative adversarial networks unlock new methods for cognitive science. *Trends Cogn. Sci.* **25**, 788–801, DOI: https://doi.org/10.1016/j.tics.2021.06.006 (2021).

19. Schurgin, M. W., Wixted, J. T. & Brady, T. F. Psychophysical scaling reveals a unified theory of visual memory strength. *Nat Hum Behav* **4**, 1156–1172 (2020).

20. Luck, S. J. & Vogel, E. K. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends Cogn. Sci.* **17**, 391–400 (2013).

21. Morgenstern, Y. *et al.* An image-computable model of human visual shape similarity (2020).

22. Rensink, R. A. CHAPTER 13 - change blindness. In Itti, L., Rees, G. & Tsotsos, J. K. (eds.) *Neurobiology of Attention*, 76–81 (Academic Press, Burlington, 2005).

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

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

25. Haber, R. N. & Standing, L. G. Direct measures of short-term visual storage. *Q. J. Exp. Psychol.* **21**, 43–54, DOI: 10.1080/14640746908400193 (1969). https://doi.org/10.1080/14640746908400193.

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

27. Kanai, R. & Verstraten, F. A. J. Visual transients without feature changes are sufficient for the percept of a change. *Vis. Res.* **44**, 2233–2240 (2004).

28. Rensink, R. A., O'Regan, J. K. & Clark, J. J. To see or not to see: The need for attention to perceive changes in scenes. *Psychol. Sci.* **8**, 368–373 (1997).

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

30. Goodfellow, I. *et al.* Generative adversarial nets. *Adv. neural information processing systems* **27** (2014).

31. Karras, T. *et al.* Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119 (2020).

32. zllrunning. face-parsing.pytorch. https://github.com/zllrunning/face-parsing.PyTorch.

33. Yu, C. *et al.* Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341 (2018).

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

35. Kasahara, S. & Takada, K. Stealth updates of visual information by leveraging change blindness and computational visual morphing. *ACM Trans. Appl. Percept.* **18**, 1–17 (2021).

36. Laloyaux, C., Devue, C., Doyen, S., David, E. & Cleeremans, A. Undetected changes in visible stimuli influence subsequent decisions. *Conscious. Cogn.* **17**, 646–656 (2008).

37. Isola, P., Xiao, J., Parikh, D., Torralba, A. & Oliva, A. What makes a photograph memorable? *IEEE transactions on pattern analysis machine intelligence* **36**, DOI: 10.1109/TPAMI.2013.200 (2013).

Acknowledgements

Author contributions statement

N.I. wrote the machine learning source code and performed the data analysis. K.S. and K.T. wrote the online experiment source code. M.S. was involved in the discussion and reviewed the manuscript. S.K. designed the study and provided the funding. N.I. and S.K. prepared all the figures. N.I., K.S., and S.K. wrote the main manuscript text. All authors read and approved the final manuscript.

Additional information

The author(s) declare no competing interests.