Using Deep Neural Networks to Disentangle Visual and Semantic Information in Human Perception and Memory.

Adva Shoham[1,*], Idan Grosbard[1,2,3,*], Or Patashnik[3], Daniel Cohen-Or[3], Galit Yovel[1,2]

[1]School of Psychological Sciences
[2]Sagol School of Neuroscience
[3]The Blavatnik School of Computer Science
Tel Aviv University, Tel Aviv, Israel
[*]Equal contribution

Corresponding authors:
Adva Shoham (advashoham@mail.tau.ac.il)
Galit Yovel (gality@tauex.tau.ac.il)

Abstract

Mental representations of familiar categories are composed of visual and semantic information. Disentangling the contributions of visual and semantic information is challenging as they are intermixed in human mental representations. Deep neural networks (DNNs) that are trained either on images or on text or by pairing images and text enable us now to disentangle human mental representations into their visual, visual-semantic and semantic components. Here we used these DNNs to uncover the content of human mental representations of familiar faces and objects when they are viewed or recalled from memory. Results show a larger visual than semantic contribution when images are viewed and a reversed pattern when they are recalled. We further revealed a previously unknown unique contribution of an integrated visual-semantic representation in both perception and memory. We propose a new framework in which visual and semantic information contribute independently and interactively to mental representations in perception and memory.

Introduction

An essential function of the human mind is to generate mental representations that enable recognition of people and objects in our environment, so we can effectively interact with them[1,2]. Successful recognition relies on the ability to match the representations that are generated by the perceptual system when stimuli are viewed to their stored representations in memory. These mental representations, however, are not exact replicas of the external world but are reconstructions of the mind. Unravelling how the external world is reconstructed by the mind, is a long standing, challenging quest in Cognitive Sciences, as we have no direct access to the content of these mental representations[3].

Prominent theories in cognitive sciences have debated for decades whether mental representations of stimuli in the external world are primarily perceptual[4–7], semantic [8,9], or both [10]. Despite numerous behavioral and neuroimaging studies that have explored the contribution of perceptual and semantic information to mental representations [11–22], fundamental questions about the content of these representations have remained unanswered: what is the relative contribution of perceptual and semantic information to these mental representations, do they contribute independently or interactively, and how are they manifested when images are viewed or recalled from memory? These questions are hard to answer with current cognitive and neural measures, as perceptual and semantic information are typically intermixed in these measures and therefore difficult to disentangle.

One way to disentangle these different types of information is by using computational models that can generate pure visual or semantic representations. The success of this approach depends on the extent to which these algorithms can represent complex, naturalistic visual and semantic information that is similar to the human mind. Recent deep neural networks (DNNs) that are trained either on images or text offer a way to address this challenge, by generating distinct visual and semantic representations for the same stimuli (Figure 1). Although DNNs differ from the human mind in many ways, including architecture, computational operations and training experience[23–26], many recent studies have shown that these algorithms account for a significant proportion of variance in human representations of faces and objects and are by far better than any previous computational models [25,27–32]. In addition, a multi-modal algorithm that learns to classify images by pairing images and their semantically meaningful captions from webpages (CLIP-contrastive image-language pre-training)[33] enables us now to explore the existence of an integrated visual-semantic representation that has not been explored previously in humans' mental representations. Current models of face and object recognition presume that visual and semantic information are processed by distinct systems and are linked in long-term memory[34–38]. However, semantic information is naturally associated with familiar stimuli during the process of learning already in infancy[39–41]. This learning may shape the visual representation of faces and objects, generating an integrated visual-semantic representation that we can now explore with multi-modal DNNs.

Accordingly, in the current study we used these DNNs to disentangle the unique contributions of visual, visual-semantic and semantic information to the representations generated by humans for the same visual stimuli when they are viewed or recalled from memory. This was possible by measuring the representational geometry of the same stimuli based on their DNNs embeddings and using them as predictors of human visual similarity ratings of the same images when they are viewed (perception) or recalled based on their names (memory). Recent studies that have examined the similarity between the representations generated by humans and DNNs for faces and objects have focused on visual DNNs and on human representations during perception when the images are viewed [(e.g. 42–48)]. The contribution of visual-semantic and semantic DNNs to these mental representations and the nature of the representations during recall, have not been explored so far.

To quantitatively assess the contributions of visual, visual-semantic and semantic information to the mental representations of familiar stimuli, we used visual (VGG-16)[49], visual-semantic (CLIP)[33] and semantic (SGPT)[50] DNNs to model human representations of familiar faces in perception and memory. To study human face representations, human participants were asked to rate the visual similarity of famous faces when presented with their images or when presented with their names, and therefore had to reconstruct their visual appearance from memory (Figure 2A). These similarity measures were used to construct the representational geometry of familiar faces in perception and memory (Figure 2B-C). We used the representational geometries of the same identities based on visual (VGG-16), visual-semantic (CLIP) and semantic (SGPT) DNNs (see Figure 1) to predict human representations in perception and memory. This enabled us to assess whether visual-semantic and semantic DNNs improve predictions of human visual representations beyond the commonly used visual DNN [(e.g. 42–48)] and quantify their unique contributions in perception and memory. The same method was used to assess human semantic representations. Finally, we used the same approach to extend our findings from faces to the representation of objects.

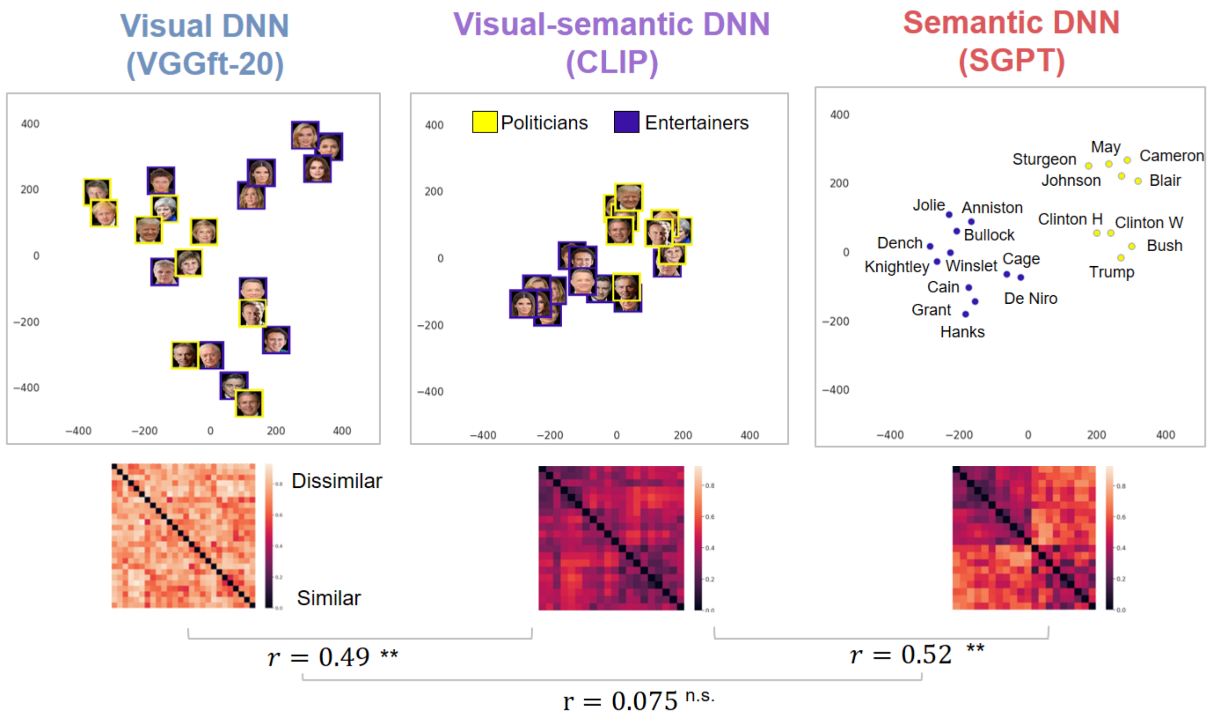


Figure 1: **The representational geometry of familiar faces based on visual, visual-semantic and semantic DNNs.** A t-SNE visualization of the representational geometry of familiar faces based on the RDMs of their embeddings in visual (VGGft-20: face pre-trained VGG fine-tuned to the 20 familiar identities), visual-semantic (CLIP) and semantic (SGPT) DNNs. VGGft-20 and CLIP representations are based on images and SGPT representation is based on the first paragraph from Wikipedia of the same familiar identities. The CLIP visual-semantic RDM was correlated with both the pure visual (VGGft-20) (r(188) = 0.49, p < 001, two-sided, CI= 0.37, 0.59) and pure semantic (SGPT) representations (r(188)=0.52, p < .001, two-sided, CI = 0.4, 0.61), whereas the visual (VGGft-20) and semantic (SGPT) representations were not significantly correlated (r(188)=0.09, p=0.300, two-sided CI = -.07, 0.22) (N=190).

Results

Experiment 1A: Representations of familiar faces in perception and memory

We examined the representational geometry of 20 internationally famous identities, 9 politicians and 11 entertainers. We selected identities that were included in the training set of CLIP (see methods section and Extended data Figure 1). We then fine-tuned a visual, face-trained DNN (VGG-16) to classify these 20 identities. This DNN is named, VGGft-20, ft= fine tuned; 20: for the 20 familiar identities. The results reported here are based on the representations in the penultimate layer that is used for the classification.

We measured the cosine distance between the embeddings of the face images in the visual (VGGft-20) and visual-semantic (CLIP) DNNs and the semantic (SGPT) DNN's embeddings of the first paragraph of their Wikipedia text (see Supplementary Table 1 for the Wikipedia text of each identity). Figure 1 shows the representational dissimilarity matrices (RDMs) and a t-SNE visualization [51] of the geometry of the identities according to each of the three DNNs. The representations of the identities are clustered by occupation by the semantic DNN (SGPT), and the visual-semantic DNN (CLIP), but not by the visual DNN (VGGft-20).

We computed the correlations between the RDMs of the three DNNs (N = 190). To test the significance of each correlation, we performed a two-sided, one sample t-test. The correlations between the RDMs of the different algorithms show that the pure visual (VGGft-20) and pure semantic (SGPT) representations were not correlated (r(188)=0.09, p=0.300, CI = -.07, 0.22). The RDM of the visual-semantic DNN (CLIP) was correlated with the RDMs of both the pure visual representation of VGGft-20 (r(188) = 0.49, p < 001, CI= 0.37, 0.59) and the pure semantic representation of SGPT (r(188)=0.52, p < .001, CI = 0.4, 0.61), which reflects the visual-semantic nature of CLIP's representation (see supplementary results of Experiment 1A and Extended data Figure 2 for the correlations of each of the DNNs with Gender, Occupation and Age).

To assess the nature of the representations of familiar faces in perception and memory, human participants were asked to rate the visual similarity of the same 20 familiar identities when presented with their images (perception) or by recalling their facial

appearance from memory when presented with their names (memory) (Figure 2A). We first computed the correlation between the RDM of each participant with the average RDM across all other participants (lower bound noise ceiling). The lower bound noise ceiling were r=0.51 for the Perception task and r= 0.51 for the Memory task (see Figure 2D). We then computed the correlation between the RDMs of the visual similarity ratings in

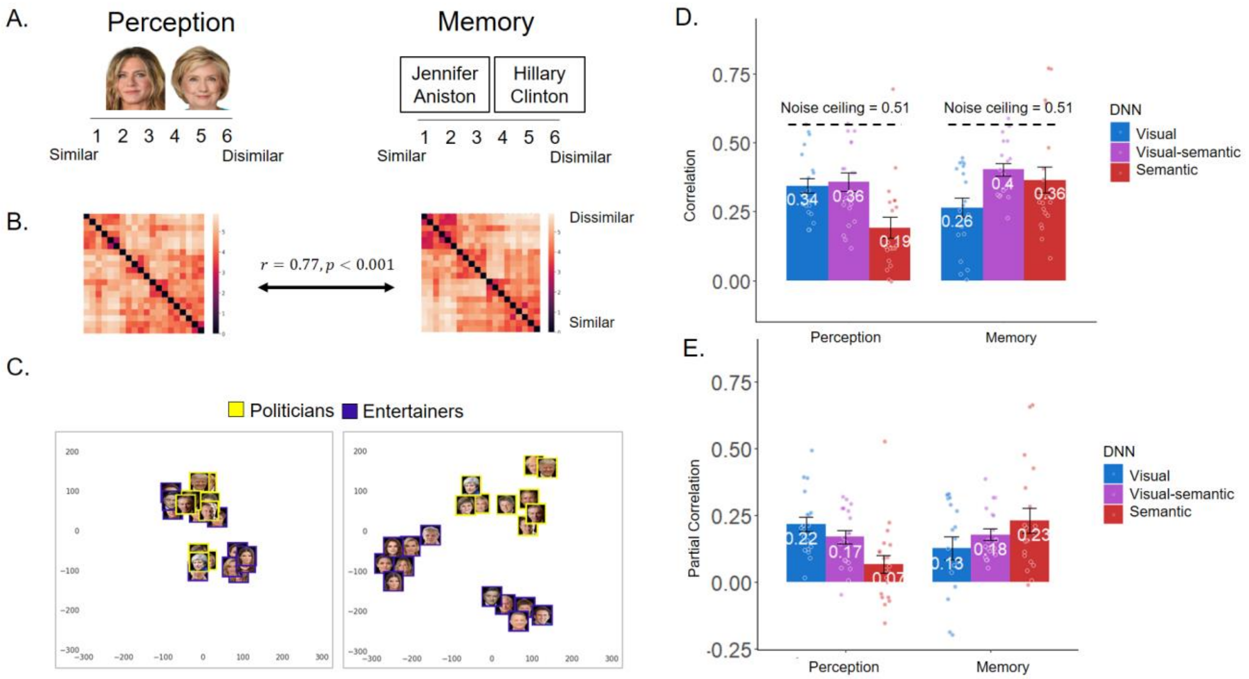


Figure 2: **The contribution of visual, visual-semantic and semantic DNNs to human representations of familiar faces in perception and memory**: A. Participants rated the visual similarity of face images (perception) or the reconstruction of their appearance from memory based on their names (memory). B. RDMs based on human visual ratings in perception (left) (N=20) and memory (right) (N=19). The representations of familiar faces in perception and memory were highly correlated (r(188) = 0.77, p < .001). C. A t-SNE visualization of the representational geometry of familiar faces in perception (left) and memory (right). The labels indicate the last name of each familiar identity. The full name can be found in Supplementary Table 1. D. The mean values +/- SEM of the correlations between the RDMs of the same identities in visual (VGG-16ft-20), visual-semantic (CLIP) and semantic (SGPT) DNNs with human representations in perception and memory across participants. Horizontal lines indicate the lower bound noise ceiling. A one-sample two-sided t-test was used on Fisher's z transformed values to test the statistical significance of each of the correlations. ANOVA and post-hoc comparisons (FDR corrected) were used on Fisher's z transformed correlations to compare the contribution of the DNNs to human representations in perception and memory. E. The mean values +/- SEM of the partial correlations of each DNN with human representations in perception and memory across participants when the other two DNNs are held out. A one-sample two-sided t-test was used on Fisher's z transformed values to test the statistical significance of each partial correlation.

perception and memory averaged across participants (Figure 2B). The correlation between the average visual similarity ratings in perception and memory was very high (r(188) = 0.77, p < .001, CI = 0.7, 0.82, two-sided), indicating that participants generated a visual image of the familiar faces in memory. We used t-SNE to visualize the representational geometry of the faces in perception and memory based on their RDMs.

As can be seen in Figure 2C, the identities are clustered based on their occupation (politicians or entertainers) in memory, but not in perception. Thus, although the visual representations in perception and memory are highly correlated, they appear to differ in the relative contribution of visual and semantic information.

Visual and semantic contributions to face representations: To quantify the contribution of visual and semantic information to humans' representations of familiar faces, we examined whether the representations of familiar faces in human perception and memory are correlated with the representations of the same identities in visual (VGG-16), visual-semantic (CLIP), and semantic (SGPT) DNNs (in the penultimate layer; see Extended data Figure 3 for VGG and CLIP across all the layers). We calculated the correlation between the RDM of each of the participants and each of the three DNNs (see Figure 2D for raw Pearson correlations and reliabilities of human similarity ratings). To test the statistical significance of the correlations they were Fisher's z transformed. To test the significance of each correlation we performed a one sample, two-sided t-test, and FDR corrected for multiple comparisons. Results show that all DNNs were significantly correlated with human mental representations in perception and memory (The values reported in the text are the mean of the Fisher's z transformed correlations). Perception: VGGft-20 (r= 0.37, t(19) = 11.5, p < .001, CI =0.30,0.43, Cohen's d = 2.57) ; CLIP (r = 0.39, t(19) = 9.4, p < .001, CI = 0.30,0.47, Cohen's d = 2.12); SGPT (r= 0.20, t(19) = 4.6, p < .001, CI =0.11,0.30, Cohen's d = 1.04). Memory: VGGft-20 (r = 0.28, t(18) = 7.3, p < .001, CI: 0.20,0.36, Cohen's d = 1.68); CLIP (r= 0.43, t(18) = 15.0, p < .001, CI =0.37,0.49, Cohen's d = 3.44) and SGPT (r= 0.41, t(18)= 6.42, p < .001, CI = 0.55,0.28, Cohen's d = 1.47). However, they showed different patterns, with higher correlation with visual than semantic DNN in perception and a reversed pattern in memory. A mixed ANOVA with DNN (visual, visual-semantic, semantic) and Task (Perception, Memory) on the Fisher's z transformed correlations across participants revealed a significant interaction of DNN and Task F(1.12,41.27) = 7.72, p= 0.007, η_p^2 =0.17. The Greenhouse-Geisser correction was used to adjust for lack of sphericity. Post-hoc comparison (two-sided, FDR corrected) revealed a lower correlation of the semantic DNN than the visual-semantic DNN t(111) =3.04, p = 0.008) and the visual DNN t(111) =2.71, p = .011) in visual perception and a lower correlation of the visual DNN than the visual-semantic t(111) =2.55, p = .036) and the semantic DNN t(111) =2.21, p = .043) in visual memory.

Given that the visual-semantic DNN is correlated with both the visual and the semantic DNNs (see Figure 1), we next assessed whether the visual-semantic DNN accounts for any unique variance in human representations in perception and memory beyond the pure visual and semantic DNNs. To that end, we calculated for each participant the partial correlations of each DNN with human similarity ratings of faces in perception or memory, when the two other DNNs are held out. To test the significance of each partial correlation, the correlations with Fisher's z transformed. We then performed a one sample two-sided, t-test, FDR corrected for multiple comparisons. The values reported in the text are the mean of the Fisher's z transformed partial correlations (see Figure 2E for the raw correlation values). Results show a significantly unique contribution to visual perception of the visual DNN (VGGft-20: r = 0.22, t(19)=7.9, p < .001. CI= 0.17,0.28, Cohen's d = 1.77) and visual-semantic DNN (CLIP: r= 0.17, t(19)= 6.7, p < .001. CI= 0.12,0.23, Cohen's d = 1.51) but not the semantic DNN (SGPT: r = 0.07, t(19) = 1.99, p = 0.061. CI= -0.003,0.14, Cohen's d = 0.44). Results show a significant unique contribution to visual memory of each of the three DNNs- VGGft-20 (r = 0.13, t(18) = 2.91, p =0.011. CI: 0.04, 0.22, Cohen's d = 0.67); CLIP (r = 0.18, t(18) = 7.9, p < .001. CI: 0.13, 0.23, Cohen's d = 1.82); SGPT (r = 0.25, t(18) = 4.58, p < .001. CI: 0.14, 0.37, Cohen's d = 1.05) (Figure 2E). See Extended Data Figure 4 and Extended Data Table 1, for the contributions of the DNNs to human similarity ratings, when the variance of gender, age and occupation is held out.

Finally, to compute the proportion of variance that visual-semantic and semantic DNNs explain beyond the visual DNN in human representations in perception and in memory, we performed a linear regression with the three models as predictors of the average similarity ratings in perception and memory. Given that previous studies used only visual DNNs to model face representations[46,52,53], we first used only the visual model as predictor of the behavioral data, then we used both the visual and visual-semantic as predictors, and finally we included all three models as predictors (See Table 1 for the regression analysis). We found that the proportion of variance increased from 35% to 51% when a visual-semantic DNN was added to the visual DNN, but no further improvement when the semantic DNN was added. With respect to human memory, the visual DNN accounted for 16% of the variance. The proportion of variance increased to 51% of variance when the visual-semantic DNN was added and to 67% when the semantic DNN was added to the regression analysis.

Taken together, our findings show that visual-semantic and semantic DNNs significantly improve the prediction of human representations of familiar faces, beyond the pure visual algorithm that has been so far used to model human face representations [27,31,45,46,54–59]. They also reveal a reversed contribution of visual and semantic information to perception and memory with a stronger visual contribution in perception and a stronger semantic contribution in memory.

	Perception				**Memory**			
Model	**VGG**				**VGG**			
Predictors	Estimates (Standardized)	CI	t-value	p	Estimates (Standardized)	CI	t-value	p
(Intercept)	0.57	-0.25 – 1.39	1.36	0.174	1.36	0.29 – 2.43	2.5	0.013
VGG	5.00 (0.38)	4.03 – 5.98	10.14	<0.001	3.99 (0.30)	2.71 – 5.26	6.18	<0.001
R2 / R2 adjusted:	0.353 / 0.350, F(1,188) = 102.8, p < 0.001				0.168 / 0.164, F(1,188) = 38.2, p < 0.001			
Model	**VGG and CLIP**				**VGG and CLIP**			
Predictors	Estimates (Standardized)	CI	t-value	p	Estimates (Standardized)	CI	t-value	p
(Intercept)	0.99	0.27 – 1.72	2.71	0.007	2.09	1.26-2.92	4.97	<0.001
VGG	3.15 (0.24)	2.18 – 4.12	6.4	<0.001	0.82 (0.06)	0.30 – 1.93	1.45	0.149
CLIP	3.01 (0.29)	2.25 – 3.78	7.78	<0.001	5.15 (0.50)	4.27 – 6.03	11.58	<0.001
R2 / R2 adjusted:	0.511 / 0.506, F(2,187) = 97.93, p < 0.001				0.516 / 0.511, F(2,187) = 99.74, p < 0.001			
Model	**VGG, CLIP and SGPT**				**VGG, CLIP and SGPT**			
Predictors	Estimates (Standardized)	CI	t-value	p	Estimates (Standardized)	CI	t-value	p
(Intercept)	0.74	-0.02 – 1.49	1.92	0.056	1.01	0.28 – 1.74	2.17	0.031
VGG	3.37 (0.26)	2.39 – 4.35	6.76	<0.001	1.74 (0.13)	0.79 – 2.69	3.61	<0.001
CLIP	2.49 (0.24)	1.59 – 3.39	5.45	<0.001	2.97 (0.29)	2.10 – 3.84	6.73	<0.001
SGPT	0.56 (0.08)	0.04 – 1.08	2.11	0.036	2.32 (0.34)	1.81 – 2.82	9.08	<0.001
R2 / R2 adjusted:	0.523 / 0.515, F(3,186) = 67.97, p < 0.001				0.664 / 0.659, F(3,186) = 123, p < 0.001			

Table 1: A linear regression model comparison in which only pre-trained face-VGG-ft20 was used as a predictor (top), when CLIP was added as an additional predictor (middle) and when SGPT was added as a third predictor (bottom) of the representations of faces in human perception (left) and human memory (right). Statistical significance was estimated with two-sided tests.

Experiment 1B: AI-generated faces are similar to human face representations.

Results so far show correlations between human similarity ratings and the similarity of the embeddings of the face images with VGG and CLIP. This method has been commonly used to assess the correspondence between DNNs and human mental/neural representations [31,44,56,60]. Generative adversarial networks (GANs) can generate a visual image for each of these identities based on their VGG or CLIP embeddings. This offers us a more direct way to assess the similarity between humans and DNNs by asking human participants to rate the similarity of the DNN generated images and assess their correlations with human similarity ratings of the original faces. We used StyleGAN, a generative adversarial network [61,62], to generate faces based on their embeddings in VGG and CLIP (see Methods and supplementary methods of Experiment 1B for the procedure

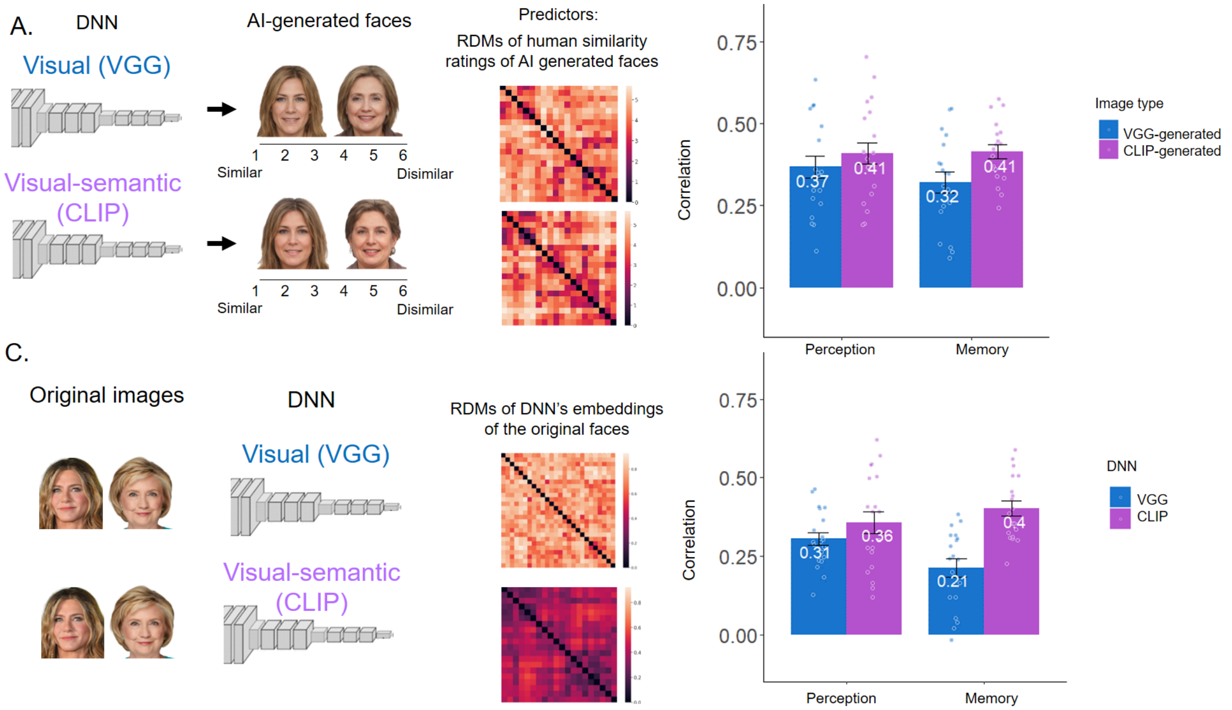


Figure 3: **Human similarity ratings of AI-generated faces.** A. RDMs of human similarity ratings of VGG-generated faces (top) and CLIP-generated faces (bottom). The AI-generated faces are not copyrighted and can be obtained by contacting the authors. The mean values +/- SEM of the correlations between averaged human similarity ratings of VGG-generated and CLIP-generated faces with the similarity rating of the original faces of the same identities in perception (N = 20) or memory (N = 19) across participants. B. The RDM of VGG (top) and CLIP (bottom) based on the embeddings of the original images. The mean values +/- SEM of the correlations with the RDMs of the embeddings of VGG and CLIP. ANOVA and post-hoc comparisons (FDR corrected) were used to compare the results shown in panels A and B.

used to generate the faces; See supplementary results of Experiment 1B for human recognition level of VGG and CLIP-generated faces). These AI-generated images offer us a way to assess the similarity between human perception and memory and the representations generated by VGG and CLIP. To that effect, we asked a new group of human participants to rate the visual similarity of the VGG- or CLIP-generated faces. We averaged human similarity ratings for VGG-generated and CLIP-generated faces and computed their RDMs (Figure 3A). We computed the correlations between these RDMs and the RDMs of human similarity ratings in perception and memory for each participant. Because the VGG-generated faces were created with a VGG algorithm that was not fine-tuned to the 20 famous faces, in this analysis we correlated human behavior RDMs with the RDM that is based on the embeddings of the original VGG, which yielded correlations that were slightly lower (Figure 3B) than the correlations in Experiment 1A that were computed with VGGft-20 (Figure 2D).

To test the similarity between these patterns of results we performed a 3-way mixed ANOVA with Task (Perception, Memory) as a between participants factor and DNN Type (VGG, CLIP) and Similarity Type (DNN embedding, GAN-image similarity) as repeated measures on the correlations. The similarity data based on DNN embeddings are the data collected for Experiment 1A. The interaction between the three factors was not significant ($F(1,37) = 2.77$, $p = 0.105$, $\eta_p^2 = .070$), indicating that the similarity rating of the AI-generated faces yielded a similar pattern of correlations with human similarity ratings of the original images in perception and memory as the correlations with the DNN embeddings of the original faces we examined in Experiment 1 (see Extended data Figure 5 for RDM and t-SNE of human similarity ratings and additional analysis). Overall, these findings further support the similarity between the representations of humans and DNNs and offer a complementary way to investigate the correspondence between human and DNNs representations.

Experiment 2: Visual-semantic representation of unfamiliar faces.

Our findings so far show that the visual-semantic DNN (CLIP) was a strong predictor of human mental representations of familiar faces in both visual perception and visual memory. But what is the origin of this visual-semantic representation? One possibility is that it is generated from the semantic information that is associated with familiar faces. Another possibility is that visual-semantic learning of familiar faces shapes the visual features that are used for face perception beyond the information that is learned from pure visual experience. In that case, we expect that the visual-semantic DNN will be correlated also with human representations of unfamiliar faces. To test this possibility, we ran the same perception task we described above, with 20 faces that were unfamiliar to both humans and the DNNs. Similar to the analysis reported in Experiment 1, we computed the correlations and partial correlations between the RDM of human visual similarity ratings of unfamiliar faces with the RDMs of their embeddings of visual (VGG) and visual-semantic (CLIP) DNNs. For this analysis we used the pretrained VGG without fine-tuning to the 20 familiar faces. For statistical analysis the correlations were Fisher's z transformed. A two-sided, one sample t-test was used to test the significance of the correlations, FDR corrected for multiple comparisons. Raw correlations are reported in Figure 4 and the Fisher's z transformed correlations in the text. Results show that both the visual (VGG: r = 0.23, t(19) = 6.82, p < .001, CI= 0.16,0.30, Cohen's d = 1.52) and visual-semantic (CLIP: r = 0.31, t(19) = 5.49 p < .001, CI=0.19,0.42, Cohen's d = 1.23) DNNs were significantly correlated with the representation of unfamiliar faces in human perception (Figure 4B). Partial correlations further revealed that the visual (VGG: r = 0.09, t(19) = 5.3 p < .001 CI= 0.05,0.12, Cohen's d = 1.19) and visual-semantic (CLIP: r= 0.21, t(19) = 4.63 p < .001 CI= 0.12,0.31, Cohen's d = 1.03) representations uniquely contributed to human visual representation of unfamiliar faces (Figure 4C).

We next assessed whether the pattern of correlations that we found for unfamiliar faces is different from the results that we found for familiar faces reported in Experiment 1A (Figure 4B). We performed a mixed ANOVA with DNN (VGG, CLIP) as a repeated measure factor and face familiarity (Unfamiliar, Familiar) as a between-participants factor on the Fisher's z transformed correlations. The analysis revealed a main effect of DNN (F(1,38) = 10.52, p = 0.002, η_p^2 = .217) indicating that CLIP was a better predictor than VGG of human perceptual representations. We found no effect of face familiarity (F(1,38) = 2.64, p = 0.112, η_p^2 = .07) and no significant interaction between DNN and Face familiarity (F(1,38) = 0.04, p = 0.835, η_p^2 = .001).

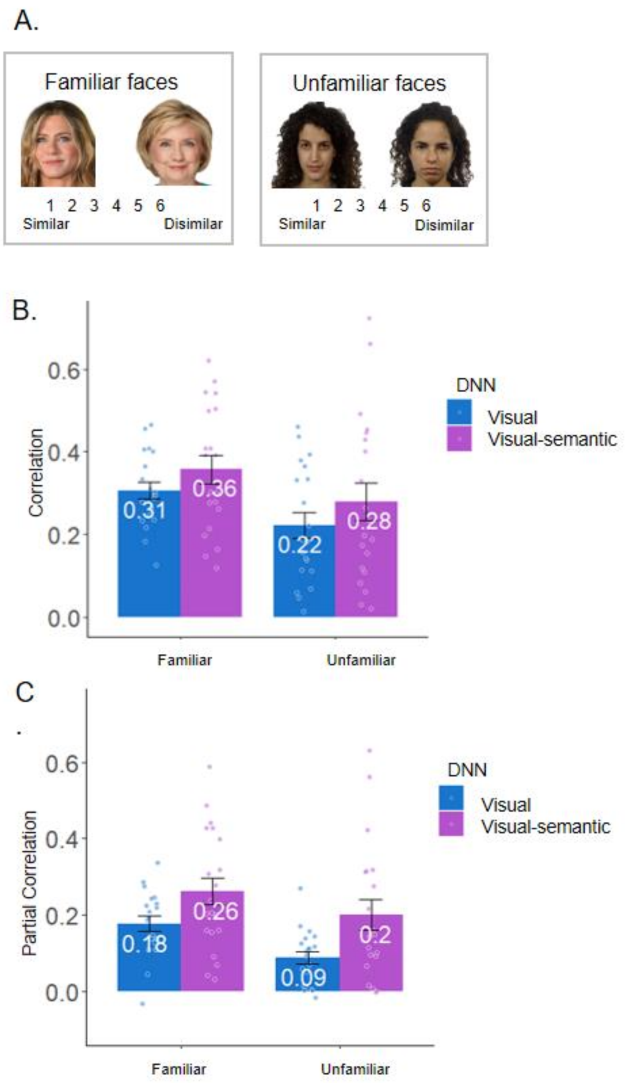


Figure 4: **The contribution of visual and visual-semantic information to the perceptual representation of unfamiliar faces** A. Human participants rated the visual similarity of unfamiliar faces using the same procedure we used for familiar faces in Experiment 1. The original familiar face images used in the experiment are not copyrighted. The face images shown are licensed drawings of famous and unfamiliar identities from freepik.com. These images were not used in the experiment. B. Data are presented as mean values +/- SEM of the correlations between visual (VGG) and visual-semantic (CLIP) with visual similarity ratings of familiar (N = 20) and with unfamiliar (N = 20) faces. C. The mean values +/- SEM of the partial correlations between visual (VGG) and visual-semantic (CLIP) with visual similarity ratings of familiar and with unfamiliar faces. ANOVA and post-hoc comparisons (FDR corrected) on the Fisher's Z transformed correlations were used to assess the contribution of each DNN to human representations of familiar and unfamiliar faces.

These findings suggest that the visual-semantic DNN was a better predictor than the purely visual DNN for both familiar and unfamiliar faces (Figure 4). Note that the correlations of the visual DNN (VGG) with familiar faces are lower than in the results reported in Experiment 1, because in Experiment 1 we used VGGft-20, which was fine-tuned to the familiar identities and therefore a better fit of their representations.

Taken together, our findings show that visual-semantic learning of familiar faces generates a visual representation that uniquely contributes to the representations of both familiar and unfamiliar faces beyond the contribution of pure visual information. See also supplementary results of Experiment 2 and supplementary Figure 1 for CLIP and VGG classification performance for familiar and unfamiliar faces. We will discuss the implications of these findings on current models of face recognition in the Discussion section.

Experiment 3: Visual and semantic contribution to semantic representations

Human mental representations of familiar stimuli can be also represented by their semantic meaning. We can therefore apply the same approach to assess the extent to which visual, visual-semantic and semantic DNNs account for these semantic representations (see methods). This can also further validate the extent to which DNNs' semantic representations are similar to human semantic representations. To that effect, participants were asked to ignore the visual appearance of the identities and judge them only based on their biographical information. A new group of participants made semantic similarity judgments when presented with the images or the names of the identities (Figure 5A). Given that the semantic information about the identities is retrieved from memory both in the image and the name tasks, we expected that the two tasks would generate similar findings. The correlations between the rating of each participant with the average ratings of all other participants (lower bound noise ceiling) of the image task was r = 0.55 and for name task was r = 0.74. Indeed, the correlation between the RDMs of semantic similarity judgments of images and names (averaged across participants) was very high (r=0.94, p < .001, CI=0.92,0.95, two-sided) (Figure 5B). Figure 5C shows that in both tasks the identities were clustered by their occupation. We then computed for each participant the correlations (Figure 5D) and the partial correlations (Figure 5E) with the three DNNs. Correlations were converted to Fisher's z transformed for statistical analysis. We performed a one sample, two-sided, t-test, FDR corrected for multiple comparisons. The values reported in the text are the mean of the Fisher's z transformed correlations (see Figure 5D for the raw correlation values). Results reveal significant correlations between the visual, visual-semantic and semantic DNNs with semantic similarity judgments when images were presented (VGGft-20: r = 0.22, t(18) = 5.9, p < .001. CI= 0.14,0.30, Cohen's d = 1.36; CLIP: r = 0.46, t(18) = 10.4, p < .001. CI= 0.37,0.55, Cohen's d = 2.39: SGPT r = 0.54, t(18) = 5.6, p < .001. CI= 0.34,0.75, Cohen's d = 1.29) and when names were presented (visual VGGft-20: r = 0.2, t(19) = 7.1, p < .001. CI= 0.14,0.27, Cohen's d = 1.59; visual-semantic CLIP: r = 0.60, t(19) = 17.8, p < .001. CI= 0.53,0.67, Cohen's d = 3.98: semantic SGPT r = 0.81, t(19) = 11.89, p < .001. CI=0.67,0.95, Cohen's d = 2.66) (Figure 5D). A 2-way ANOVA of DNN (VGGft-20, CLIP, SGPT) and Task

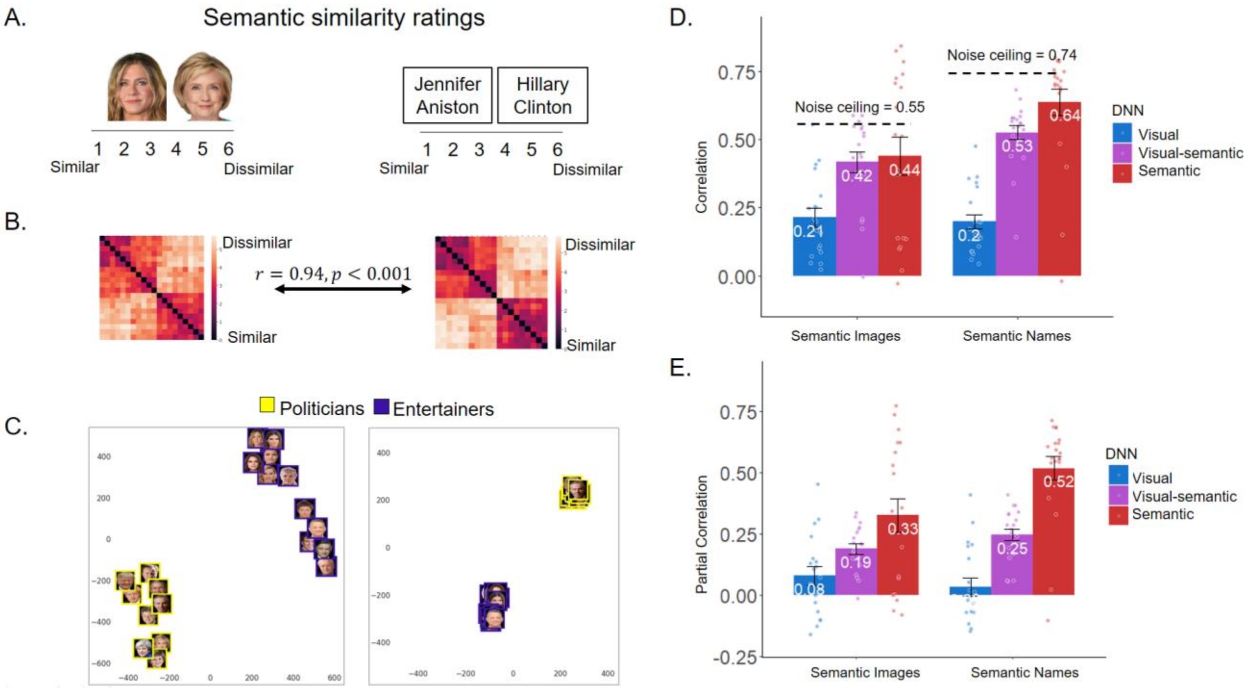


Figure 5: **The contribution of visual, visual-semantic and semantic DNNs to human semantic representations**. A. Participants rated the semantic similarity between familiar identities when they were presented with pictures (N=19) or their names (N=20). The original familiar face images used in the experiment are not copyrighted. The face images shown are licensed drawings of famous identities from freepik.com that were not used in the experiment. B. The RDMs based on human semantic similarity ratings were highly correlated. C. A t-SNE visualization of the RDMs, showing the clustering of the identities based on their occupations. The labels indicate the last name of each familiar identity. The full name can be found in Supplementary Table 1. D. The mean values +/- SEM of the correlations between the RDMs based on embeddings of the same identities in visual (VGG), visual-semantic (CLIP) and semantic (SGPT) DNNs with human semantic representations. A one-sample two-sided t-test was used on Fisher's z transformed values to test the statistical significance of each of the correlations. ANOVA and post-hoc comparisons (FDR corrected) were used on Fisher's z transformed correlations to assess the contribution of each DNN to human representations in perception and memory. E. The mean values and +/- SEM of partial correlations for each DNN with semantic rating based on images or names when the other two DNNs are held constant. A one-sample, two-sided t-test was used on Fisher's z transformed correlations to test the statistical significance of each of the partial correlations.

(Perception, Memory) (Greenhouse-Geisser corrected) revealed a main effect of DNN F(1.06,39.2) = 42.98, p < 001, η_p^2 = 0.53 and a marginally significant interaction between DNN and task (F(1.06,39.2) = 3.84, p = 0.055, η_p^2 = 0.09). Post-hoc comparisons (two-sided, FDR corrected) revealed a significantly larger correlation of human semantic representations with the semantic DNN than the visual-semantic DNN (t(114) = 2.57, p = 0.011) and the visual DNN (t(114) = 7.91, p < .001) and a larger correlation with the visual-semantic DNN than the visual DNN (t(114) = 5.33, p < .001).

To test the unique contribution of each type of information we measured the significance of each partial correlation, using a two-sided, one sample t-test, FDR corrected for multiple comparisons. The partial correlations with semantic judgements when images were presented, were significant for visual-semantic DNN (CLIP: r = 0.19, t(19) = 8.59, p < .001. CI: 0.15-0.24, Cohen's d = 1.97) and semantic DNN (SGPT: r = 0.38, t(19) = 4.51, p < .001. CI: 0.2,0.56, Cohen's d = 1.04) but not with visual DNN (VGGft-20: r = 0.08, t(19) = 2.08, p = 0.052, Cohen's d = 0.48). The same pattern was found for semantic judgements when names were presented: a high correlation with visual-semantic DNN (CLIP: r = 0.25, t(19) = 10.13, p < .001. CI: 0.2,0.31, Cohen's d = 10.13) and with semantic DNN (SGPT r = 0.6, t(19) = 10.45, p < .001. CI: 0.48,0.6, Cohen's d = 2.34) but not with visual DNN (VGGft-20: r = 0.04, t(19) = 0.96, p = 0.350, Cohen's d = 0.96). Thus, both semantic and visual-semantic DNNs uniquely contribute to semantic ratings of familiar faces. These findings show a strong correspondence between humans and large language model representations of semantic information. They also indicate that human semantic judgements are not purely semantic but also include a visual-semantic component.

Experiment 4: Representations of objects in perception and memory.

In a final experiment, we extended our findings on familiar faces to familiar objects using the same approach. We measured the distance between the embeddings of 20 images of familiar objects (see Extended Data Figure 6 for the images and names) based on VGG-16 pre-trained on ImageNet, CLIP and the embedding of their Wikipedia definition based on SGPT (see Supplementary Table 3). Because visual and semantic information are correlated for some categories of objects (e.g. animals tend to be curvier than man-made objects), to assess the isolated contributions of visual and semantic information, we pre-selected a subset of 20 objects from different categories that showed the lowest correlation between their visual DNN (VGG-objects) and semantic DNN (SPGT) representations (r = 0.17, p = .013, two-sided, CI=0.036,0.31) (see Methods). We then measured the correlations between the RDMs of the embeddings of the visual, visual-semantic, and semantic DNNs (see Extended Data Figure 7 for RDMs and t-SNE visualization). The RDM of the visual-semantic (CLIP) DNN was correlated with the RDMs of both the visual DNN (VGG-objects, r = 0.56, p < .001, two-sided, CI=0.45,0.65) and

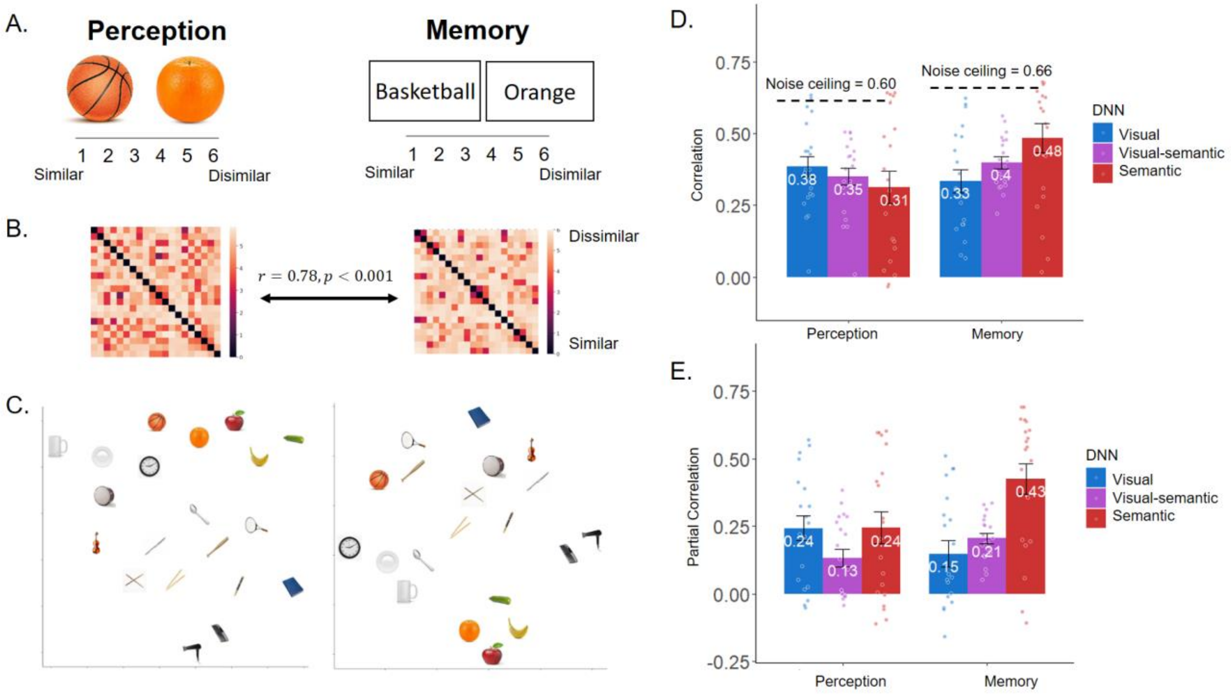


Figure 6: **The contribution of visual, visual-semantic and semantic DNNs to human representations of objects in perception and memory** A. An example of two objects that were used in the visual similarity task, when images are viewed (Perception, N=20) or recalled based on their names (Memory, N=20). B. The correlations between visual (object-VGG), visual-semantic (CLIP) and semantic (SGPT) DNNs and human perception and memory of objects. C. The t-SNE visualization shows the representational geometry of objects based their representations in perception and memory. D. The mean values +/- SEM of the correlations between the RDMs based on embeddings of the same objects in visual (object VGG), visual-semantic (CLIP) and semantic (SGPT) DNNs with human representations in perception and memory across participants. A one-sample, two-sided t-test was used on Fisher's z transformed values to test the statistical significance of each of the correlations. ANOVA and post-hoc comparisons (FDR corrected) were used on Fisher's z transformed correlations to assess the contribution of each DNN to human representations in perception and memory. E. The mean values and +/- SEM of the partial correlations of each DNN with human representations in perception and memory across participants while the other two DNNs are held out. A one-sample two-sided t-test was used on Fisher's z transformed values to test the statistical significance of each of the partial correlations.

semantic DNN (SGPT, r= 0.29, p < .001, two-sided, CI=0.15,0.41), consistent with the visual-semantic nature of its representation.

We then asked human participants to judge the visual similarity of the objects when presented with their images (perception) or to recall their visual appearance from memory (Figure 6A). The correlation of the ratings of each participant with the average ratings of all other participants (lower bound noise ceiling) was r= 0.60 for the image task and r= 0.66, for the name task. Similar to the findings with familiar faces, the correlations

	Perception				**Memory**			
Model	**VGG**				**VGG**			
Predictors	Estimates (Standardized)	CI	t-value	p	Estimates (Standardized)	CI	t-value	p
(Intercept)	1.99 (4.97)	1.44 – 2.55	7.05	<0.001	3.12 (5.33)	2.51 – 3.72	10.11	<0.001
VGG	4.22 (0.50)	3.44 – 4.99	10.68	<0.001	3.13 (0.37)	2.28 – 3.98	7.28	<0.001
R2 / R2 adjusted:	0.38/ 0.37, F(1,188) = 114.17, p < 0.001				0.22 / 0..22, F(1,188) = 52.92, p < 0.001			
Model	**VGG and CLIP**				**VGG and CLIP**			
Predictors	Estimates (Standardized)	CI	t-value	p	Estimates (Standardized)	CI	t-value	p
(Intercept)	1.69 (4.97)	1.14 – 2.25	6.05	<0.001	2.65 (5.33)	2.08 – 3.23	9.14	<0.001
VGG	3.13 (0.37)	2.23 – 4.03	6.84	<0.001	1.46 (0.17)	0.52 – 2.40	3.07	0.002
CLIP	3.58 (0.23)	1.90 – 5.25	4.22	<0.001	5.51 (0.35)	3.77 – 7.25	6.25	<0.001
R2 / R2 adjusted:	0.43 / 0.43, F(2,187) = 71.06, p < 0.001				0.35 / 0.35, F(2,187) =51.38, p < 0.001			
Model	**VGG, CLIP and SGPT**				**VGG, CLIP and SGPT**			
Predictors	Estimates (Standardized)	CI	t-value	p	Estimates (Standardized)	CI	t-value	p
(Intercept)	0.88 (4.97)	0.32 – 1.45	3.07	0.002	1.21 (5.33)	0.76 – 1.67	5.26	<0.001
VGG	3.09	2.27 – 3.91	7.4	<0.001	1.39 (0.16)	0.73 – 2.05	4.14	<0.001
CLIP	2.4	0.82 – 3.97	3	0.003	3.41 (0.22)	2.14 – 4.67	5.32	<0.001
SGPT	1.68	1.15 – 2.21	6.21	<0.001	2.99 (0.47)	2.56 – 3.42	13.76	<0.001
R2 / R2 adjusted:	0.53 / 0.52, F(3,186) =69.77 p < 0.001				0.68 / 0.67, F(3,186) = 131.8, p < 0.001			

Table 2: A linear regression model comparison in which only pre-trained face-VGG-ft20 was used as a predictor (top), when CLIP was added as an additional predictor (middle) and when SGPT was added as a third predictor (bottom) of the representations of objects in human perception (left) and human memory (right). Statistical significance was estimated with two-sided tests.

between the representations of the images when they were viewed or recalled from memory, averaged across participants, was very high (r = 0.78, p < 0.001, two-sided, CI= 0.72,0.83).

Visual and semantic contributions to object representations: We then examined the correlations of the three DNNs with human similarity ratings of objects and their recall from memory (Figure 6D). To test the significance of each correlation we performed a one sample t-test (two-tailed) on Fisher's z transformed correlations, FDR corrected for multiple comparisons. Figure 6 displays the raw correlations and the lower bound noise ceiling. In the text we report the Fisher's z transformed correlations. The representation in perception was correlated with visual (VGG-objects: r= 0.42, t(19) = 9.73, p < .001. CI= 0.33,0.51, Cohen's d = 2.18), visual-semantic (CLIP: r = 0.37, t(19) = 11.11, p < .001.

CI= 0.30,0.44, Cohen's d = 2.48) and semantic (SGPT: r = 0.35, t(19) = 5.3, p < .001. CI: 0.21-0.49, Cohen's d = 1.19) DNNs. The representation in memory was correlated with visual (VGG-objects: r = 0.36, t(19) = 7.6, p < .001. CI: 0.26-0.46, Cohen's d = 1.71), visual-semantic (CLIP: r = 0.42, t(19) = 17.14, p < .001. CI= 0.37,0.48, Cohen's d = 3.83) and semantic (SGPT: r = 0.56, t(19) = 8.81, p < .001. CI: 0.43-0.70, Cohen's d = 1.97) DNNs. To examine whether the contributions of the three DNNs were different in perception and memory, we performed a mixed ANOVA with DNN (visual, visual-semantic, semantic) as repeated measures and Task (Perception, Memory) as a between-subjects factor on the Fisher's z transformed correlations across participants. The Greenhouse-Geisser correction was used to adjust for the lack of sphericity. Results revealed a marginally significant interaction of DNN and Task F(1.06, 40.23) = 3.9, p= 0.053, η^2=0.09. Post hoc comparison (two-sided, FDR corrected) revealed a lower correlation of the visual DNN relative to the semantic DNN with human representation in memory (t(114) = 2.91, p =.031), but no difference between the three DNNs in perception. To examine the unique contribution of each DNN, we computed the partial correlations of each DNN with visual perception and visual memory when holding the other two DNNs constant for each participant. All partial correlations were Fisher's z transformed and statistical significance was assessed with a two-sided, one sample t-test, FDR corrected for multiple comparisons. Results show a significant unique contribution of each of the three DNNs in perception and memory. Visual Perception: visual DNN (r = 0.24, t(19) = 5.25, p < .001. CI=0.15,0.34, Cohen's d = 1.17), visual-semantic DNN, CLIP (r = 0.13, t(19) = 4.12, p < .001. CI= 0.07,0.2, Cohen's d = 0.92) and semantic DNN, SGPT (r = 0.24, t(19) = 4.17, p < .001. CI= 0.12,0.37, Cohen's d = 0.93). Visual memory: VGG (r = 0.15, t(19) = 3.01, p < .007. CI= 0.05,0.27, Cohen's d = 0.67); CLIP (r = 0.21, t(19) = 10.9, p < .001. CI= 0.17,0.24, Cohen's d = 2.46); SGPT (r = 0.43, t(19) = 7.43, p < .001, CI=. 0.31,0.54, Cohen's d = 1.66).

Finally, to measure the proportion of variance that the visual-semantic and the semantic DNNs accounted for beyond the typically used visual DNN, we performed a linear regression with the three models as predictors of the average similarity ratings in perception and memory. Results for the representation in perception show that the visual DNN alone accounts for 38% of variance, together with the visual-semantic DNN they accounted for 43% of the total variance and all three algorithms accounted for 53% of the variance. Results for the representation in memory show that the visual DNN alone accounts for 22% of variance, together with the visual-semantic DNN they accounted for 35% of the total variance and all three algorithms accounted for 68% of the variance (see Table 2 for regression analysis). Overall, these findings show that by adding the visual-semantic and semantic DNNs to the pure visual DNN, we can better account for human representations of objects in perception and memory. In supplementary results of Experiment 4 and Supplementary Figure 2, we report the correlations of the DNNs with human semantic similarity ratings of objects, which overall showed high correlations with semantic and visual-semantic DNNs as was also the case for faces.

Discussion

The goal of the current study was to uncover the content of human mental representations in perception and memory by quantitatively assess the independent as well as integrated contributions of visual and semantic information. This was enabled by using the representations that are generated for the images of the same stimuli in visual (VGG) and visual-semantic (CLIP) DNNs and for their textual description by language (SGPT) DNNs, to model their representations when images are presented (perception) or when they are recalled from memory. Our findings reveal that the integration of visual, visual-semantic, and semantic DNNs explains a considerable amount of variance (>60%) in human mental representations of faces and objects. Moreover, their relative contributions to mental representations in perception and memory were reversed. We found a larger visual contribution in perception and a larger semantic contribution during recall. Notably, an integrated visual-semantic representation accounted for additional unique variance in both perception and memory beyond the pure contributions of perceptual and semantic information.

The representation that is generated by the visual-semantic DNN (CLIP) offers us a way to explore a distinct, integrated visual-semantic representation in perception and memory that has not been considered so far. Current models of face and object recognition posit that perceptual and semantic information are processed by separate cognitive and neural mechanisms [12,34,38]. However, during the process of learning and interacting with faces and objects, visual and semantic information are naturally associated [39,41,63]. Our findings show that a visual-semantic algorithm (CLIP), which learns to classify images by linking them to meaningful semantic information, generates a distinct visual representation that accounts for unique variance in the representations of faces and objects in human perception and memory. Moreover, the visual-semantic DNN uniquely contributed to the representation of unfamiliar faces, for which humans and DNNs do not have semantic information, beyond the visual DNN. These findings imply that contrary to current models of face and object recognition that regard semantic information as a mere supplement that is linked to visual information in long-term memory [17,34,36,38,64], our findings propose a framework in which semantic information actively shapes the perceptual representation during the learning process. It further suggests that the recognition advantage that is reported in humans for familiar than unfamiliar faces [65–68] is not only due to the extensive visual experience with familiar faces [66,69,70] but may be also due to the contribution of visual-semantic experience to the representation of face identity, which is biased for familiar identities (see supplementary results of Experiment 1B and supplementary Figure 1 for CLIP's performance for familiar and unfamiliar faces).

The many recent studies that have examined the similarity between human and DNNs representations of faces and objects have focused on visual DNNs and the representations of faces and objects in human perception [25,27,32,56,71,72]. Our study goes beyond these studies by examining how semantic algorithms may account for face and object representations[37]. Moreover, we examined their contributions to the nature of the representation that is generated during recall. Our findings show that even pure semantic information, extracted by NLP algorithms, accounts for the representation of familiar faces and objects in memory, beyond visual and visual-semantic information. These non-visual, semantic representations were based on textual descriptions of non-visual, semantic information about the familiar identities in Wikipedia and were indeed uncorrelated with their visual representations generated by the visual DNN. The correlations between these semantic representations and human semantic ratings of the same stimuli were very high, indicating their validity as measures of human semantic judgements (Figure 5). It is noteworthy, that the semantic definitions of objects often include visual descriptions about the color and shape of the object. In addition, different object categories often inherently differ not only in their semantics but also in their visual appearance and are therefore difficult to dissociate[14]. Nevertheless, this is not the case for faces, where the textual description included no information on their visual appearance. Still, even for faces, semantic information influenced visual similarity judgments of faces recalled from memory. Thus, even though participants were instructed to judge the visual similarity of the faces or objects, pure semantic information still dominated the representation in memory. Overall, these findings show how DNNs enable us to uncover concealed semantic biases in the representation of familiar stimuli in memory.

The similarity between human mental representations and DNN's representations is typically measured by the distance between the embeddings of the images [29,44,72–74]. Current face generator algorithms (StyleGAN) enable us to visualize these embeddings. This procedure offers us a way to assess the correspondence between humans' and DNNs' representations, by asking human participants to rate the similarity between AI-generated faces and using these similarity ratings as predictors of human representations of the original identities in perception and memory. The results of this analysis were remarkably similar to the predictions based on VGG and CLIP embeddings of the original images (Figure 3). Furthermore, the pattern of correlations of age, gender and occupation with similarity ratings of the generated images (Extended data Figure 5) was similar to the pattern that was found with the embedding of the original images (Extended data Figure 2). These findings further demonstrate the close correspondence between human and DNNs representational geometries, indicating that they can be used as valid models of human face representations.

Our findings show that by combining visual, visual-semantic and semantic DNNs we can account for 50-60% of the variance in human mental representations in perception and memory, suggesting the usefulness of these algorithms to model human representations. Nevertheless, several limitations should be acknowledged. First, the visual-semantic (CLIP) and semantic DNNs used in our study were pre-trained on a massive amount of data, the content of which we had neither access to nor knowledge about. Still, the visualization of their representations and the correlations among them (Figure 1) enabled us to validate their visual, visual-semantic and semantic content. In addition, the high correlation between human semantic rating and the SGPT embeddings, indicates that it is a valid computational model of human semantic representations (Figure 5, Supplementary Figure 2). Second, our findings show that the visual-semantic information accounts for additional variance beyond the pure visual and pure semantic information. However, it is not clear how this distinct visual-semantic representation is generated during learning, how early in visual processing it emerges and what is its functional significance. Finally, when using behavioral judgments to measure perceptual and/or semantic similarity, we cannot tell the extent to which they reflect a pure representation or are also affected decision processes[4]. In addition, our approach should be used in future neuroimaging studies, which solve this possible effect of response bias, by measuring the representational geometry based on the distance between the neural response to the stimuli, eliminating the need for an explicit similarity judgment. These neuroimaging studies can also reveal at what stage of processing and by which neural systems these different representations emerge.

The reversed contributions of visual and semantic information in perception and memory are consistent with a recent study that measured reaction time and EEG during visual and semantic judgments of objects while they were presented or recalled from memory[21]. Their findings showed reverse information flow between perception and memory with semantic to visual processing in memory and vice versa in perception. These and our findings are in line with studies that showed that reconstruction of information in memory prioritizes the semantic meaning of the stimulus[75]. These reconstructions account for recognition errors that are based on the meaning rather than the perceptual features of the recalled events[76,77]. The dominance of semantic information in the representation in memory is also in line with the benefit of semantic relative to visual information during encoding for recognition memory [78–80]. The high correlation between semantic DNN and human memory indicates that NLP algorithms can be used to predict the way semantic information may enhance[12] or interfere[77] with performance on recognition tasks.

Our findings are also relevant to studies that investigated the degree of similarity between the representations generated for the same stimuli in perception and memory. Many studies have revealed that the neural representations generated during recall elicit similar neural responses to the representations that are generated in perception [81–83]. Here we addressed this central question by measuring the correlations between the representational geometries of the same stimuli when they are viewed or recalled from memory. Consistent with findings that emphasize the similarity between perception and recall/imagery, we reveal high correlations between the representational geometries when images are viewed and when they are recalled from memory. However, by decomposing these representations into their visual and semantic components, we also revealed significant differences between them as reflected by the reversed dominance of perceptual and semantic information in perception and memory.

In conclusion, human intelligence relies on cognitive operations that integrate different codes into a unified representation. Our findings show that even face recognition, which has been traditionally perceived as a task resolved by the visual system, is best predicted by a combination of visual, visual-semantic and semantic representations. Moreover, the semantic information that is naturally associated with visual categories during learning may play an even greater role than previously considered in human perception, by shaping the visual features of learned and unlearned categories. Deep learning algorithms enable us to examine the nature of these visual, visual-semantic, and semantic representations and their unique contributions to perception and memory. The approach we used here is not limited to faces or objects and can be similarly applied to model human representations of other categories and domains such as familiar places as well as sounds and voices. It can be similarly applied to study other measures of human mental representations as well as in populations who suffer from perceptual and memory disorders. Our findings also inform machine intelligence in that they highlight the contribution of multi-model systems that integrate sensory and semantic information. Future AI models of human cognition may also integrate emotional, and motor representations as well as attentional and motivational factors that select behaviorally relevant information. This type of multi-system operation may underlie the efficient learning and adaptive behavior that is required to further close the gap between computer and human general intelligence. Overall, our findings show how human and machine intelligence may inform and advance each other by offering theoretical insights and methodological approaches that are less evident when each system is studied alone.

Methods

The study confirms with the ethical regulations and was approved by the ethics committee of Tel Aviv University (Approval no. 0005357_2).

Pre-registration of the familiar face experiments is available here https://aspredicted.org/N3T_M3N. 04/25/2022. Pre-registration of the object experiments is available here https://aspredicted.org/5F1_BVG. 01/23/2023.

As we indicated in the pre-registration, prior to collecting data for the experiments reported here, we performed a pilot study in which we ran visual and semantic similarity judgements tasks for all possible pairs of 26 celebrity faces and examined the reliability of these tasks and the correlations between the visual and semantic DNNs with visual and semantic judgements. Analysis of these data across different image samples indicated that similar results and reliability measures can be obtained with 20 images (190 pairs), which is the number of images we used in the experiments reported in this article. The data reported in this paper were collected after the study was pre-registered and do not include the data collected in the pilot study.

Familiar Face images: We selected 20 identities of international celebrities, 9 politicians and 11 entertainers, that were included in the CLIP training set (see Extended Data Figure 1 for a complete list of the identities) and generate a visual-semantic representations of their face images. To select identities that CLIP was trained with, we examined whether face images of the selected identities could be correctly classified by CLIP. We measured the cosine similarity between the embeddings of names and images of these identities based on CLIP and selected only identities that the similarity between their image and their name was the highest relative to any of the other names (see Extended data Figure 1).

We then selected face images of these identities from a Google images search. The face images were colored images that included the face of each identity. The background of the images was removed, and all images were aligned to the same size. The name stimuli were printed names of the identities in black font on a white background (see Figure 2A).

DNN generated faces: To generate faces, we used a generator for human faces – StyleGAN [61,62]. Our goal was to transform VGG and CLIP embeddings to StyleGAN's embeddings. To do so, we first trained a model to generate StyleGAN's embeddings based on the DNN's embeddings; the full training procedure is described in the supplementary methods of Experiment 1B. We used this model to generate images of the identities used in the experiment based on VGG and CLIP representations. To reduce the effect of noise found in the DNNs' embeddings of a single image, which can be transferred to the generated images, we used the DNNs' representations of 20 different images of each identity and calculated their mean StyleGAN's embedding. Using this mean we generated each image.

Unfamiliar faces: We selected a set of 20 faces from an in-house face database that we created by taking photos of Tel Aviv University students and their friends in a studio by a professional photographer. Since their face images do not appear on webpages in English, they are unlikely to be familiar to CLIP.

Object images: Twenty familiar objects were downloaded from Google images. We first measured the embeddings of a set of 80 images based on VGG-16 and their textual definition based on SGPT. We selected 20 objects that had the lowest correlations between the visual and semantic embeddings with all other images, so we can measure their independent contributions.

Deep neural networks:

Face-trained VGG-16: To obtain a pure visual representation of the face images we used the VGG-16 algorithm [49] and trained it to classify 8749 identities of faces from the VGGFace2 dataset [84]. All images were first aligned using the MTCNN algorithm [85]. Training and image preprocessing followed the procedure from [86] with the following changes: Images were normalized according to $\mu = [0.5, 0.5, 0.5], \sigma = [0.5, 0.5, 0.5]$, the training was done on batches of 128 images, for 120 epochs of 1000 batches and after 100 epochs, the learning rate was reduced to 1e-3. The network's performance was measured on a face verification task consisting of 6000 pairs of face images from the Labeled Faces in the Wild benchmark [87]. The network's verification performance was 97%.

To familiarize VGG with the face stimuli, we finetuned the pre-trained VGG-16 on the same 20 identities used during the test (VGGft-20). We collected 20 face images for each of the 20 identities from Google images. 75% of images (15) were used to train the network, and the other 25% of images (5) were used to validate its classification performance. All face images were preprocessed using the same methodology used for pre-training. To fit the network, we replaced its classification layer (FC8) with a new classification layer of 20 units only. To reduce the forgetting of the identity distinguishing facial features, and overfitting the specific features of these specific images, we froze the weights of all convolutional layers, and only trained the fully connected layers. Optimization was done during 10 epochs, using the Adam optimization algorithm (Kingma & Ba, 2014) with a batch size of 1, a learning rate of 1e-5, with all other parameters having the default values offered by the PyTorch framework [88]. The model reached 100% classification accuracy on the validation set. The correlation between the penultimate layer representation of the pre-trained and the fine-tuned VGG was 0.95.

To extract the representation of each face image of the study stimuli, the images were first aligned using the MTCNN algorithm [85]. We then extracted the embeddings based on the feature vector representation in the penultimate (fc7) layer of the network and computed the similarity between each face pair based on the cosine distance between these feature vectors. In Extended data Figure 3 we also show these correlations with all other layers of the network.

Object-trained VGG-16: To examine the similarities between the objects' images, we used the VGG16 architecture[49] trained on the ImageNet classification dataset [89]. We used the implementation and pre-trained weights provided by the PyTorch framework [88].

CLIP (Contrastive Language-Image pre-training): CLIP is trained to create similar representations for images and their text caption based on 400M images from the internet [33]. We extracted the embeddings of each face image based on the output layer of trained ViT-B/32 architecture. In Extended data Figure 3 we show these correlations with the RDMs of all layers of the network. We computed the similarity between each face pair based on the cosine distance between these representations. All face images were aligned using the MTCNN algorithm [85], and then pre-processed according to the values supplied by OpenAI's implementation [33].

SGPT: GPT Sentence Embeddings for Semantic Search is a recent natural language processing (NLP) algorithm that is first pre-trained to predict the next word in a sentence similar to other NLP algorithms and use contrastive fine-tuning to create similar representations for pairs of sentences that describe the same content[50]. We extracted the embeddings of the text of the first paragraph in Wikipedia of each identity based on the 1.3B parameters bi-encoder's output layer and computed the similarity between each identity pair based on the cosine distance between these representations. Supplementary Table 1 shows the first paragraph in Wikipedia that was used for each identity.

Human similarity ratings:

Participants

Familiar Face tasks: A total of 80 participants were recruited for this study from the Prolific platform. 20 participants were assigned to each of the four experimental conditions: Visual similarity based on images (mean age 30, 19 females) or names (mean age 29, 14 females) and semantic similarity based on images (mean age 30, 13 females) or names (mean age 29, 16 females). Two participants were excluded from the analysis (1 participant from the visual memory condition, and 1 participant from the semantic (images) condition) because they were not familiar with 30% or more of the presented identities, which resulted in a total of 78 participants. The participants were paid 4 GBP for their participation in the experiment (8 GBP/hour). They gave informed consent prior to the experiment. The study was approved by the ethics committee of Tel Aviv University.

AI-generated face tasks: A total of 40 participants were recruited from the prolific platform. 20 participants were assigned to each of the two experimental conditions: visual similarity ratings according to VGG generated images (mean age 31, 15 females) or CLIP generated images (mean age 30, 17 females). The participants were paid 4 GBP for their participation in the experiment (8 GBP/hour). They gave informed consent prior to the experiment. The study was approved by the ethics committee of Tel Aviv University.

Unfamiliar faces: A total of 20 participants were recruited (mean age = 29.5,13 females) from the prolific platform to rate the visual similarity of a set of unfamiliar faces. The participants were paid 4 GBP for their participation in the experiment (8 GBP/hour). They gave informed consent prior to the experiment. The study was approved by the ethics committee of Tel Aviv University.

Object tasks: 80 participants were recruited for this study from the Prolific platform 20 participants were assigned to each of the four experimental conditions: Visual similarity based on images (mean age 31, 11 females) or names (mean age 32, 8 females) and semantic similarity based on images (mean age 30, 13 females) or names (mean age 31, 15 females). The participants were paid 4 GBP for their participation in the experiment (8 GBP/hour). They gave informed consent prior to the experiment. The study was approved by the ethics committee of Tel Aviv University.

Procedure

Participants rated the visual or semantic similarity of all possible pairs of the 20 identities (190 pairs).

Visual similarity rating: Each trial presented one pair of images or names of two different identities, and the participants were asked to rate the visual similarity of their faces either based on the images (perception condition) or the reconstruction of their faces from memory, based on their name (memory condition). In the memory condition, we emphasized in the instructions that similarity should be based on the visual appearance of the face based on their memory. The image/name pairs were presented on the screen one at a time, above a similarity scale (1 (very similar) - 6 (very dissimilar)) until response. The participants selected the similarity score with the mouse. The next pair was presented 1 second after their response. The participants had a forced break for a minimum of 10 seconds after 80 pairs were presented, and again after 160 pairs were presented. After the completion of the ratings of all 190 pairs, the participants were asked to indicate for each face/name whether they were familiar with it before the experiment. The experiment lasted about 30 minutes.

The same procedure was used to collect human similarity ratings of the DNN-generated faces and unfamiliar faces.

Semantic similarity rating: Participants were asked to rate the similarity of the identities based on biographical or any other semantic information they know about them. We emphasized that the similarity should not be based on visual appearance but only on semantic information.

The same procedures were used to collect visual and semantic similarity ratings with object images and object names (see Extended Data Figure 6 for the list of images and names).

Data analysis:

Representational similarity matrices (RDMs)

Human RDMs: We generated RDMs for the 190 pairs of faces/objects (all paired combinations of 20 stimuli) based on human similarity ratings for each participant. We also generated RDMs based on human similarity ratings averaged across all participants of the VGG and CLIP-generated images. For familiar faces and objects, we generated RDMs based on visual or semantic similarity of the images and their names. For unfamiliar faces, we generated RDMs based on visual similarity ratings of images.

We computed the inter-rater reliability of each task, by the correlations of each participant rating with the average similarity rating of the other participants in that task, which is the lower bound noise ceiling.

Trials' exclusion: We excluded trials if the participant was unfamiliar with one of the familiar identities. Participants who were unfamiliar with 30% (or more) of the identities were excluded from the analysis. We excluded trials with similarity rating response that were shorter than 200 ms and longer than 30,000 ms (30 seconds), based on the assumption that participants did not perform the task well on such trials. In the face tasks, 5.8% of the trials were excluded (3 trials due to RT longer than 30 sec and the rest based on the familiarity criterion). In the object tasks, 0.4% trials were excluded (due to RT longer than 30 sec).

Deep neural network (DNN) RDMs: We measured the cosine similarity between the embedding of the same identities/objects in the penultimate layer of VGG and in the output layer of CLIP, based on their images and in the output layer of SGPT based on the first paragraph of their Wikipedia textual description/dictionary definition respectively.

Correlation and regression analyses

We computed the correlations between human similarity ratings and DNN distance scores. To assess the unique contribution of the different algorithms to the variance in human similarity ratings we computed the partial correlations of each DNN while holding the other two DNNs constant. ANOVA, post hoc comparisons and one-sample t-tests were performed on the Fisher's z transformed correlations using two-sided null-hypothesis tests. FDR was applied to correct for multiple comparisons.

To assess the proportion of variance that the DNN explained in the averaged human similarity ratings of faces and objects in perception and memory we performed a multiple linear regression in which VGG was the first predictor. We then added CLIP and then also

SGPT to assess the additional proportion of variance that each algorithm explained in the human data.

To assess the similarity of human visual representations of the original stimuli and the AI-generated images, we used the average human visual similarity ratings of VGG- and CLIP-generated images as the predictors of human visual similarity ratings of the original face stimuli. We calculated the correlations of each participant's ratings based on memory and perception with RDMs of human visual similarity ratings of VGG- and CLIP-generated images. Then we performed ANOVA to compare this pattern of correlations with the pattern found when using the DNNs embeddings of the original faces as predictors of the humans' ratings.

Representational geometry visualization: We used t-SNE, a nonlinear dimensionality reduction technique [90] for visualization purposes only. The correlation and regression analyses were based on the similarity measures (RDMs).

Data Availability Statement

Data ware analyzed using R [91].

The datasets are available in this link

https://osf.io/3hwmy/?view_only=584ab8985520411183321008a2fb1a60

The following Datasets were used for DNN training:

1. ImageNet - https://www.image-net.org/download.php
2. VGGFace2 - Currently there is no official download link - need to contact the dataset's original publishers.
3. CelebA-HQ - https://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html

Code Availability Statement

The code for data analysis is available in this link:

https://osf.io/3hwmy/?view_only=584ab8985520411183321008a2fb1a60

The deep learning algorithms are open-source codes that can be obtained from the references cited in the text.

Acknowledgement

We would like to thank Rafi Malach, Michael Gilead and Idan Blank for discussions of this manuscript. We thank Guy Levy and Jenan Hajajra for their help with stimuli collection and initial analysis of Experiment 4. This work was funded by an Israeli Science Foundation grant (ISF 917/2021) to GY. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions Statement

A.S. and G.Y. conceptualization and experimental design. A.S and I.G wrote the code and performed the analysis of the deep learning algorithms. A.S collected data and performed analysis of human behavioral data. I.G, O.P and D.C-O designed the experiment and wrote the codes for Experiment 1B. A.S and G.Y wrote the manuscript.

Competing interests Statement

The authors declare no competing interests.

References

1. Gibson, J. J. The ecological approach to visual perception: classic edition. J. Broadcast. (1979).

2. Sperry, R. W. Neurology and the mind-body problem. Am Sci 40, 291–312 (1952).

3. Miller, G. A. The cognitive revolution: a historical perspective. Trends Cogn. Sci. 7, 141–144 (2003).

4. Firestone, C. & Scholl, B. J. Cognition does not affect perception : Evaluating the evidence for “ top-down ” effects. (2017)

5. Barsalou, L. W. Perceptual symbol systems. 577–660 (1999).

6. Kosslyn, S. M. Image and Brain : The Resolution of the Imagery Debate. (2014).

7. Tversky, A. Features of similarity. Psychol Rev 84, 327–352 (1977).

8. Leshinskaya, A. & Caramazza, A. For a cognitive neuroscience of concepts : Moving beyond the grounding issue. Psychon Bull Rev 991–1001, (2016).

9. Pylyshyn, Z. W. Mental imagery : In search of a theory. 157–237 (2002).

10. Clark, J. M. & Paivio, A. A Dual Coding Perspective on Encoding Processes. in Imagery and Related Mnemonic Processes (eds. McDaniel, M. A. & Pressley, M.) 5–33 (Springer New York, 1987). doi:10.1007/978-1-4612-4676-3_1.

11. Bankson, B. B., Hebart, M. N., Groen, I. I. A. & Baker, C. I. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. Neuroimage 178, 172–182 (2018).

12. Bar, M. Visual objects in context. Nat. Rev. Neurosci. 5, 617–629 (2004).

13. Barense, M. D., Henson, R. N. A. & Graham, K. S. Perception and Conception: Temporal Lobe Activity during Complex Discriminations of Familiar and Novel Faces and Objects. J. Cogn. Neurosci. 23, 3052–3067 (2011).

14. Bonnen, T. When the ventral visual stream is not enough : A deep learning account of medial temporal lobe involvement in perception ll ll Article When the ventral visual stream is not enough : A deep learning account of medial temporal lobe involvement in perception. Neuron 2755–2766 (2021)

15. Bracci, S. & Op de Beeck, H. Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. J. Neurosci. 36, 432–444 (2016).

16. Capitani, E., Caramazza, A. & Borgo, F. What Are the Facts of Semantic Category-Specific Deficits ? Cognitive Neuropsychology, 20(3-6), 213-261 (2003).

17. Clarke, A. & Tyler, L. K. Understanding What We See: How We Derive Meaning From Vision. Trends Cogn. Sci. 19, 677–687 (2015).

18. Visconti di Oleggio Castello, M., Haxby, J. V. & Gobbini, M. I. Shared neural codes for visual and semantic information about familiar faces in a common representational space. Proc. Natl. Acad. Sci. 118, e2110474118 (2021).

19. Hasantash, M. & Afraz, A. Richer color vocabulary is associated with better color memory but not color perception. Proc. Natl. Acad. Sci. 117, 31046–31052 (2020).

20. Inho, M. C. Understanding perirhinal contributions to perception and memory : Evidence through the lens of selective perirhinal damage. vol. 124 9–18 (2019).

21. Linde-Domingo, J., Treder, M. S., Kerrén, C. & Wimber, M. Evidence that neural information flow is reversed between object perception and object reconstruction from memory. Nat. Commun. 10, (2019).

22. Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L. Y. & Barense, M. D. Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. eLife 7, 1–29 (2018).

23. Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron 98, 630-644 16 (2018).

24. Geirhos, R. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. in 7th International Conference on Learning Representations, ICLR 2019 1–22 (2019).

25. Kriegeskorte, N. Deep neural networks: a new framework for modelling biological vision and brain information processing. Annual review of vision science, 1, 417-446. (2015)

26. Marcus, G. Deep Learning: A Critical Appraisal. 1–27 (2018).

27. Dobs, K., Martinez, J., Kell, A. J. E. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. Sci Adv 8, eabl8913, (2022).

28. Grand, G., Blank, I. A., Pereira, F. & Fedorenko, E. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. Nat. Hum. Behav. 6, 975–987 (2022).

29. Groen, I. I. A. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 1–26, (2018).

30. Hasson, U., Nastase, S. A. & Goldstein, A. Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Neuron 105, 416–434 (2020).

31. Abudarham, N., Grosbard, I. & Yovel, G. Face Recognition Depends on Specialized Mechanisms Tuned to View-Invariant Facial Features: Insights from Deep Neural Networks Optimized for Face or Object Recognition. Cogn Sci 45, (2021).

32. Jacobs, R. A. & Bates, C. J. Comparing the Visual Representations and Performance of Humans and Deep Neural Networks. Curr Dir Psychol Sci 28, 34–39 (2019).

33. Radford, A. Learning Transferable Visual Models From Natural Language Supervision. International conference on machine learning (2021)

34. Bruce, V. & Young, A. Understanding face recognition. Br. J. Psychol. 77, 305–327 (1986).

35. Clarke, A., Taylor, K. I., Devereux, B., Randall, B. & Tyler, L. K. From perception to conception: How meaningful objects are processed over time. Cereb. Cortex 23, (2013).

36. Clarke, A. & Tyler, L. K. Object-Specific Semantic Coding in Human Perirhinal Cortex. J. Neurosci. 34, 4766–4775 (2014).

37. Devereux, B. J., Clarke, A. & Tyler, L. K. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Sci. Rep. 8, 10636 (2018).

38. Gobbini, M. I. & Haxby, J. V. Neural systems for recognition of familiar faces. Neuropsychologia 45, 32–41 (2007).

39. Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C. & Smith, L. B. Real-world visual statistics and infants' first-learned object names. Philos. Trans. R. Soc. B Biol. Sci. 372, (2017).

40. Hall, D. G., Corrigall, K., Rhemtulla, M., Donegan, E. & Xu, F. Infants' use of lexical-category-to-meaning links in object individuation. Child Dev 79, (2008).

41. Yee, M., Jones, S. S. & Smith, L. B. Changes in visual object recognition precede the shape bias in early noun learning. Front Psychol 3, (2012).

42. Carlin, J. D. & Kriegeskorte, N. Adjudicating between face-coding models with individual-face fMRI responses. PLoS Comput Biol 13, 1–28 (2017).

43. Dobs, K., Martinez, J., Kell, A. J. E. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. 8913, (2021).

44. Kubilius, J., Bracci, S. & Beeck, H. P. Deep Neural Networks as a Computational Model for Human Shape Sensitivity. PLoS Comput Biol (2016)

45. O’Toole, A. J. & Castillo, C. D. Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning. Annu Rev Vis Sci 7, 543–570 (2021).

46. O’Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q. & Chellappa, R. Face Space Representations in Deep Convolutional Neural Networks. Trends Cogn Sci 22, 794–809 (2018).

47. Schyns, P. G., Snoek, L. & Daube, C. Degrees of algorithmic equivalence between the brain and its DNN models. Trends Cogn Sci 26, 1090–1102 (2022).

48. Tsantani, M., Kriegeskorte, N., McGettigan, C. & Garrido, L. Faces and voices in the brain: A modality-general person-identity representation in superior temporal sulcus. Neuroimage (2019).

49. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. ArXiv Prepr. ArXiv14091556 (2014).

50. Muennighoff, N. SGPT: GPT Sentence Embeddings for Semantic Search. ArXiv Prepr. ArXiv220208904 (2022).

51. Maaten, L. & Hinton, G. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, (2008).

52. Abudarham, N., Bate, S., Duchaine, B. & Yovel, G. Developmental prosopagnosics and super recognizers rely on the same facial features used by individuals with normal face recognition abilities for face identification. Neuropsychologia 160, 107963 (2021).

53. Dobs, K., Kell, A. J., Martinez, J., Cohen, M. & Kanwisher, N. Using task-optimized neural networks to understand why brains have specialized processing for faces. J Vis 20, (2020).

54. Cavazos, J. G., Jeckeln, G., Hu, Y. & O’Toole, A. J. Strategies of Face Recognition by Humans and Machines. in Deep Learning-Based Face Analytics 361–379. (Springer, 2021).

55. Jacob, G., Pramod, R. T., Katti, H. & Arun, S. P. Qualitative similarities and differences in visual object representations between brains and deep networks. Nat Commun 12, 1–14 (2021).

56. Jozwik, K. M. Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. Proc Natl Acad Sci U A 119, 1–11 (2022).

57. Song, Y., Qu, Y., Xu, S. & Liu, J. Implementation-independent representation for deep convolutional neural networks and humans in processing faces. Front Comput Neurosci 14, 601314 (2021).

58. Tian, F., Xie, H., Song, Y., Hu, S. & Liu, J. The Face Inversion Effect in Deep Convolutional Neural Networks. Front Comput Neurosci 16, 1–8 (2022).

59. Yildirim, I., Belledonne, M., Freiwald, W. & Tenenbaum, J. Efficient inverse graphics in biological face processing. Sci Adv (2020).

60. Mur, M., Bandettini, P. A. & Kriegeskorte, N. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci (2008).

61. Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Trans Pattern Anal Mach Intell 43, 4217–4228 (2018).

62. Karras, T. Analyzing and Improving the Image Quality of StyleGAN. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 8107–8116 (2019). doi:10.1109/CVPR42600.2020.00813.

63. Slone, L. K., Smith, L. B. & Yu, C. Self-generated variability in object images predicts vocabulary growth. Dev Sci 22, (2019).

64. Young, A. W. & Bruce, V. Understanding person perception. Br. J. Psychol. 102, 959–974 (2011).

65. Burton, A. M., Jenkins, R. & Schweinberger, S. R. Mental representations of familiar faces. Br. J. Psychol. 102, 943–958 (2011).

66. Jenkins, R., White, D., Montfort, X. & Burton, A. M. Variability in photos of the same face. Cognition 121, 313–323 (2011).

67. Kramer, R. S. S., Young, A. W. & Burton, A. M. Understanding face familiarity. Cognition (2018).

68. Young, A. W. & Burton, A. M. Are we face experts? Trends Cogn Sci 22(2), 1–11 (2017).

69. Burton, M. A. Why has research in face recognition progressed so slowly? The importance of variability. Q J Exp Psychol Hove 66, 1467–1485 (2013).

70. Ritchie, K. L. & Burton, A. M. Learning faces from variability. Q. J. Exp. Psychol. 70, 897–905 (2017).

71. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: Pitting neural networks against each other as models of human cognition. Proc. Natl. Acad. Sci. 117, 29330–29337 (2020).

72. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. (2016) doi:10.1038/nn.4244.

73. Kaniuth, P. & Hebart, M. N. Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. Neuroimage 257, (2022).

74. Khaligh-Razavi, S. M., Kriegeskorte, N. D. S. & Unsupervised, N. Models May Explain IT Cortical Representation. PLoS Comput Biol (2014).

75. Schacter, D. L., Norman, K. A. & Koutstaal, W. The cognitive neuroscience of constructive memory. Annual review of psychology. Annu Rev Psychol 49, 289–318 (1998).

76. Schacter, D. L. Insights From Psychology and Cognitive Neuroscience. Am. Psychol. (1999).

77. Schacter, D. L., Guerin, S. A. & St. Jacques, P. L. Memory distortion: an adaptive perspective. Trends Cogn. Sci. 15, 467–474 (2011).

78. Bower, G. H. & Karlin, M. B. Depth of processing pictures of faces and recognition memory. J. Exp. Psychol. 103, 751–757 (1974).

79. Craik, F. I. M. & Lockhart, R. S. Levels of processing: A framework for memory research. J. Verbal Learn. Verbal Behav. 11, 671–684 (1972).

80. Schwartz, L. & Yovel, G. Social Judgements Improve Face Recognition More Than Perceptual Judgements. J Vis 17, 1001 (2017).

81. Ganis, G., Thompson, W. L. & Kosslyn, S. M. Brain areas underlying visual mental imagery and visual perception: An fMRI study. Cogn. Brain Res. 20, 226–241 (2004).

82. Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R. & Fried, I. Internally generated reactivation of single neurons in human hippocampus during free recall. Science 322, 96–101 (2008).

83. O.'Craven & Kanwisher, N. G. Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions. J Cogn Neurosci 12, 1013–1023 (2000).

84. Cao, Q., Shen, L., Xie, W., Parkhi, O. M. & Zisserman, A. VGGFace2: A dataset for recognising faces across pose and age. in Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018 67–74 (2017).

85. Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. IEEE Signal Process Lett 23, 1499–1503 (2016).

86. Parkhi, O. M., Vedaldi, A. & Zisserman, A. D. F. R. 41 1-41 12 (2015).

87. Huang, G. B., Ramesh, M., Berg, T. & Learned-Miller, E. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. vol. ICCV (2007).

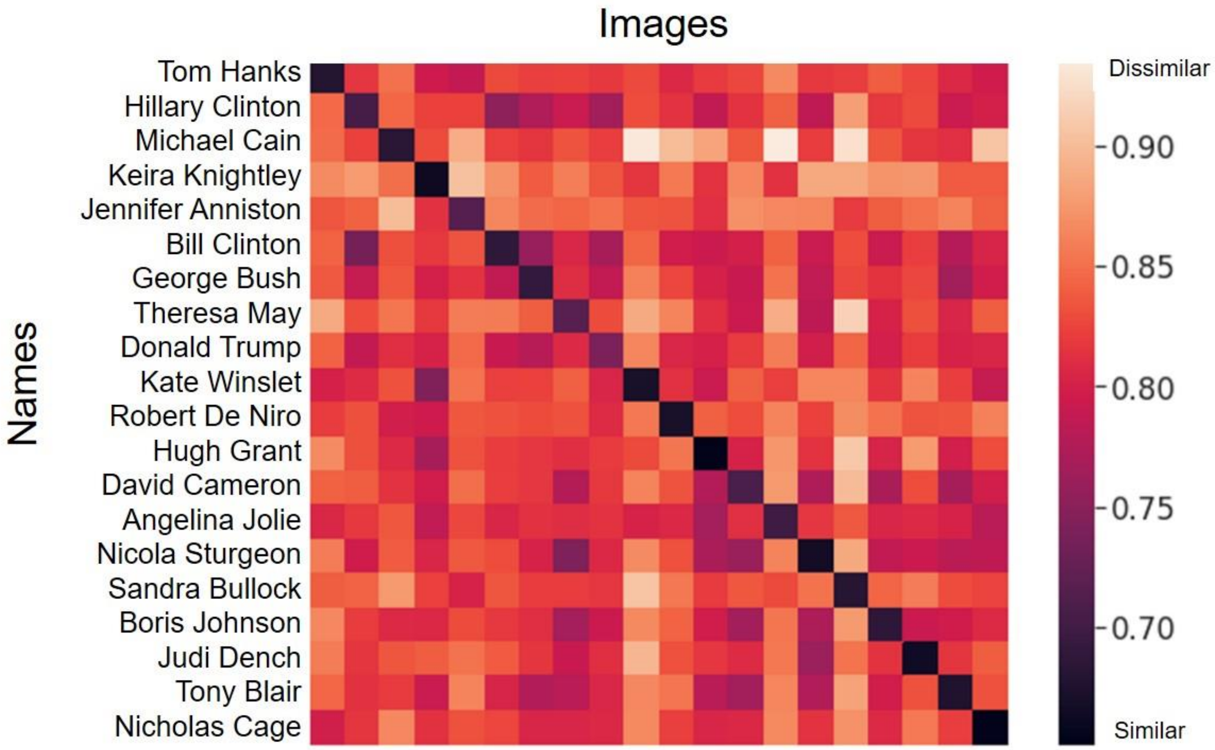
88. Paszke, A. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv Neural Inf Process Syst 32, (2019).

89. Deng, J. ImageNet: A large-scale hierarchical image database. in 2009 IEEE Conference on Computer Vision and Pattern Recognition 248–255 (IEEE, 2009).

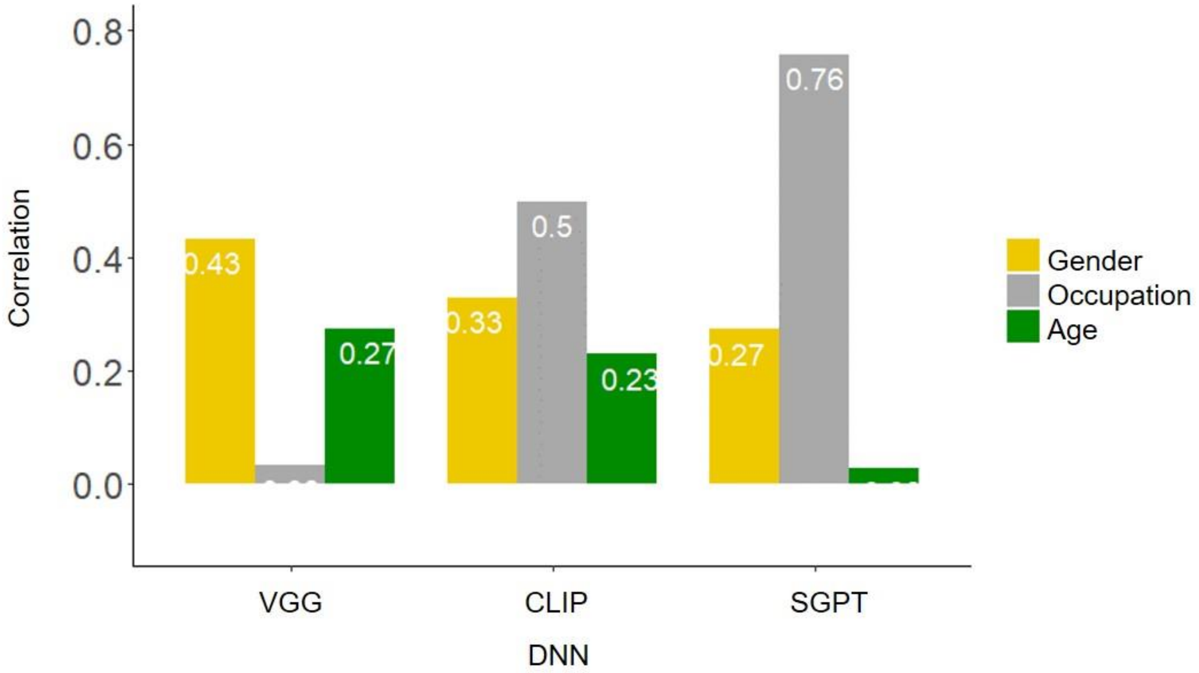
90. Ma, N., Baetens, K., Vandekerckhove, M., Van der Cruyssen, L. & Van Overwalle, F. Dissociation of a trait and a valence representation in the mPFC. Soc. Cogn. Affect. Neurosci. 9, 1506–1514 (2013).

91. R core team, R. core T. R: A language and environment for statistical computing. 3, (2019).

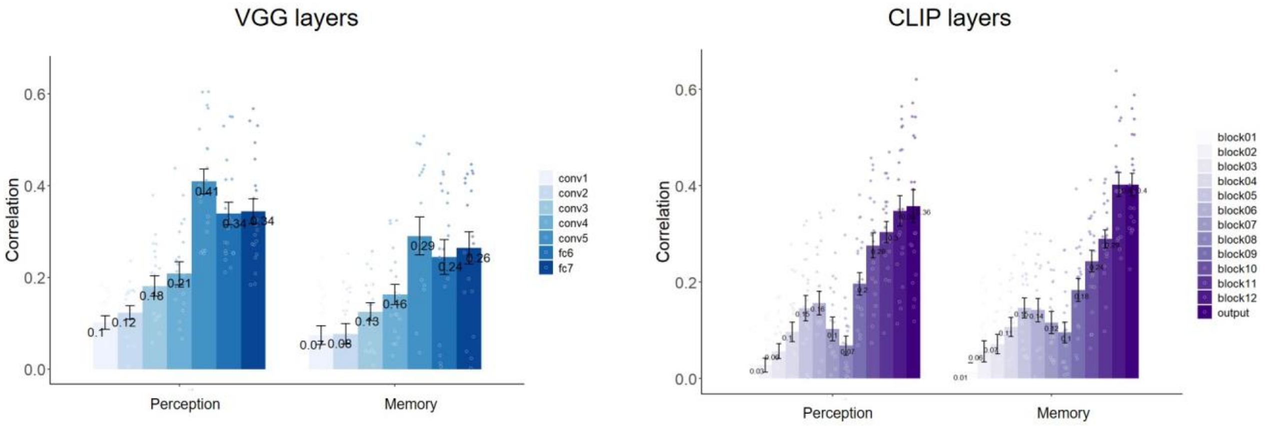
Extended data:



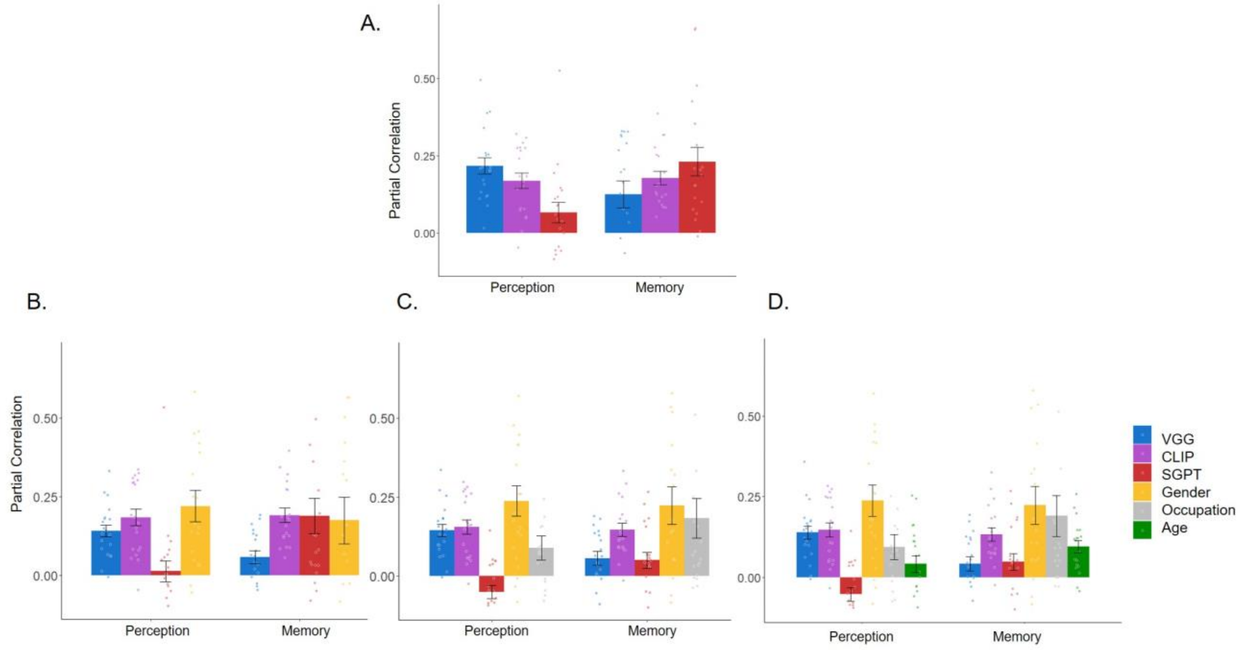
Extended Data Fig. 1 A distance matrix of CLIP embeddings of face images and their names. Only identities that were familiar to CLIP were selected to the study. An identity is considered familiar if the distance between the embedding of its name (rows) and the embedding of its corresponding face image (columns) is closest relative to all other identities. This is indicated by the dark diagonal of the matrix.



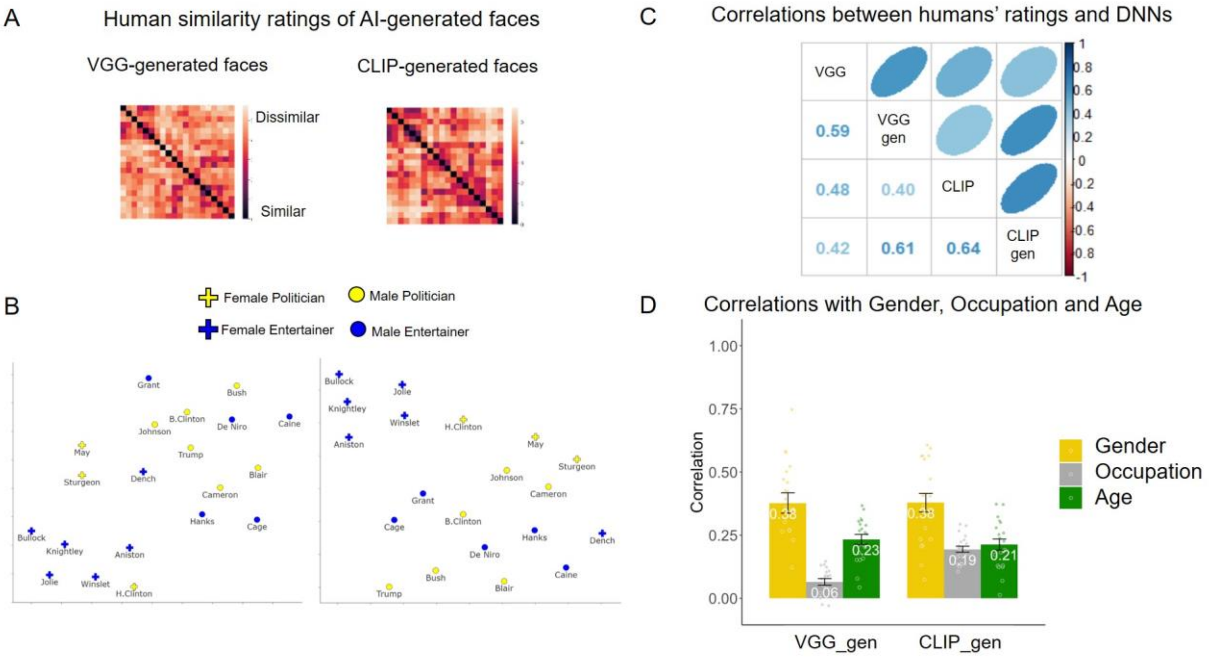
Extended Data Fig. 2 A. The mean partial correlations of the RDMs of visual (VGG-ft20), visual-semantic (CLIP) and semantic (SGPT) DNNs with RDMs of human perception (N = 20) and memory (N = 19) of the same identities, when B. gender, C. gender and occupation and D. gender, occupation and age are held out. Error bars indicate the standard error of the mean. See statistical analysis in Extended data Table 1.



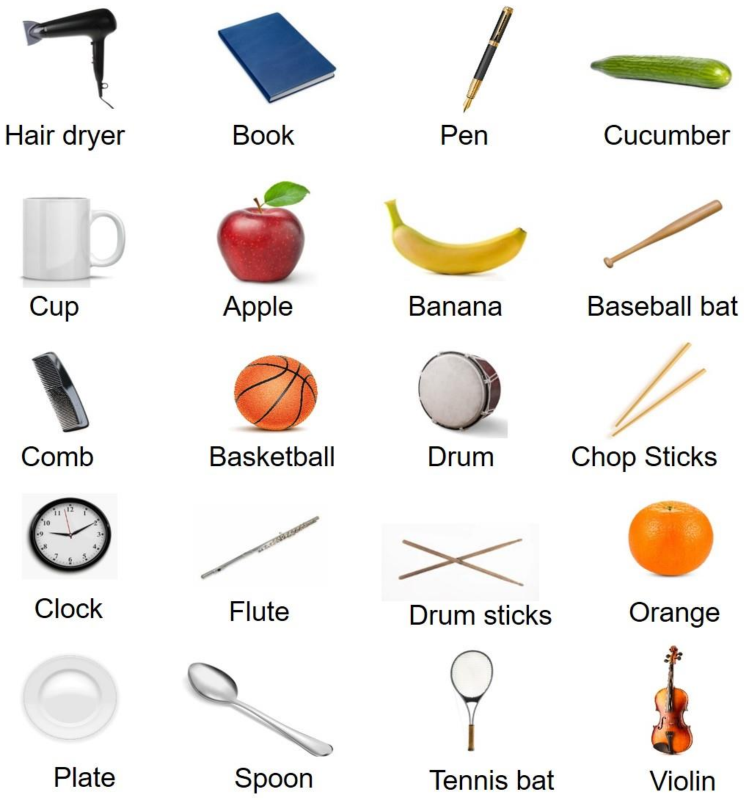
Extended Data Fig. 3 The mean correlations between RDMs of faces based on the embeddings of each layer of VGG-ft20 (left) and each layer of CLIP (right) with human visual similarity ratings in perception (left; N = 20) and memory (right; N = 19). Error bars indicate the standard error of the mean. Each dot indicates a participant.



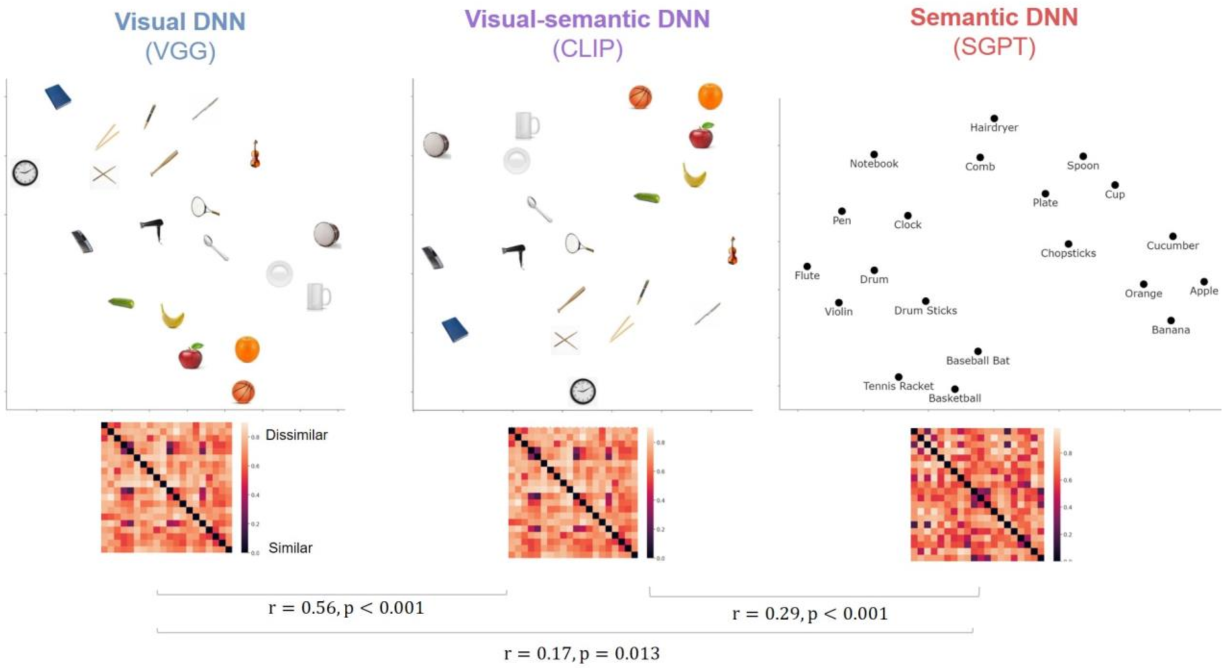
Extended Data Fig. 4 A. The mean partial correlations of the RDMs of visual (VGG-ft20), visual-semantic (CLIP) and semantic (SGPT) DNNs with RDMs of human perception (N = 20) and memory (N = 19) of the same identities, when B. gender, C. gender and occupation and D. gender, occupation and age are held out. Error bars indicate the standard error of the mean. See statistical analysis in Extended data Table 1.



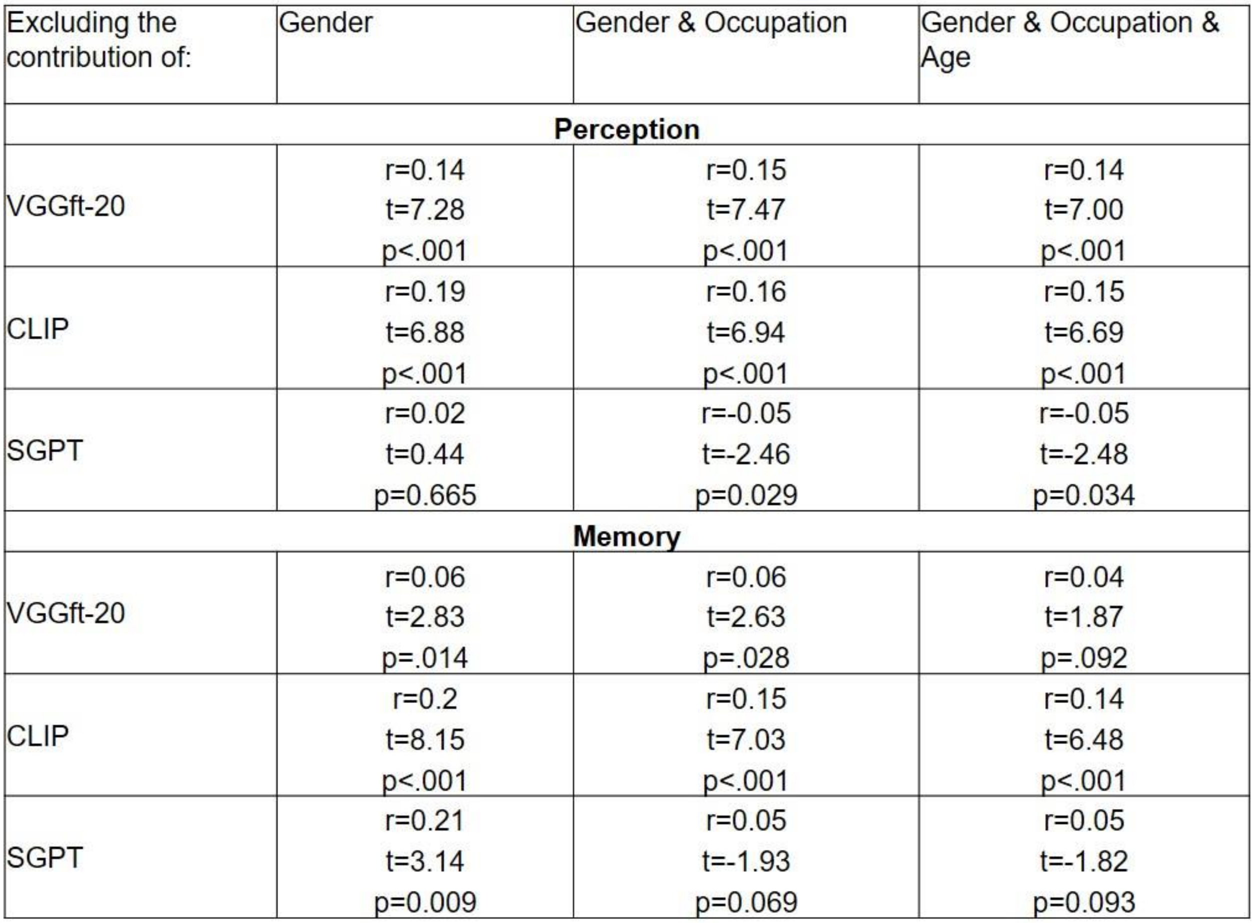
Extended Data Fig. 5 A. RDMs of human similarity rating of VGG-generated (left) and CLIP-generated (right) images of the celebrity faces. B. A t-SNE visualization of human RDMs of VGG-generated (left) and CLIP-generated (right) faces C. High correlations between human similarity ratings of VGG and CLIP-generated faces and RDMs of VGG and CLIP's embeddings of the original faces: D. The mean correlations between the RDM of human similarity ratings of VGG-generated (N = 20) and CLIP-generated faces (N = 20) with Gender, Occupation and Age. Error bars represent the standard error of mean correlations. VGG: VGG embeddings of original faces; VGG-gen: human similarity ratings of the VGG-generated faces. CLIP: CLIP embeddings of original faces; CLIP-gen: human similarity ratings of the CLIP-generated faces. The AI-generated faces cannot be copyrighted and are not shown in the figure. The images can be obtained by contacting the authors.



Extended Data Fig. 6 The 20 objects that were selected for Experiment 4. For display purpose, all images were replaced by licensed images with similar appearance from Freepik.com. A license certificate was obtained from Freepik for each of the images shown in the Figure.



Extended Data Fig. 7 The RDMs, the correlations between them and t-SNE visualization for objects based on embeddings of the images by VGG trained on ImageNet and CLIP and SGPT embeddings of their dictionary definitions (see supplementary Table 3).



Excluding the contribution of:	Gender	Gender & Occupation	Gender & Occupation & Age
Perception			

VGGft-20	r=0.14
	t=7.28
	p<.001
		r=0.15
		t=7.47
		p<.001
			r=0.14
			t=7.00
			p<.001

CLIP	r=0.19
	t=6.88
	p<.001
		r=0.16
		t=6.94
		p<.001
			r=0.15
			t=6.69
			p<.001

SGPT	r=0.02
	t=0.44
	p=0.665
		r=-0.05
		t=-2.46
		p=0.029
			r=-0.05
			t=-2.48
			p=0.034

Memory			

VGGft-20	r=0.06
	t=2.83
	p=.014
		r=0.06
		t=2.63
		p=.028
			r=0.04
			t=1.87
			p=.092

CLIP	r=0.2
	t=8.15
	p<.001
		r=0.15
		t=7.03
		p<.001
			r=0.14
			t=6.48
			p<.001

SGPT	r=0.21
	t=3.14
	p=0.009
		r=0.05
		t=-1.93
		p=0.069
			r=0.05
			t=-1.82
			p=0.093

Extended Data Table. 1 The table reports the mean partial correlations of the RDMs of visual (VGG-ft20), visual-semantic (CLIP) and semantic (SGPT) DNNs with the RDMs of human perception (N = 20) and memory (N = 19) of the same identities, when gender (left), gender and occupation (middle) and gender, occupation, and age (right) are held out. Statistical significance of partial correlations was tested with one-sample, two-sided, t-tests (FDR corrected). See Extended Data Figure 4.

Supplementary Methods and Results

Supplementary results of Experiment 1A:

The correlations between the representations of VGG, CLIP and SGPT with gender, occupation, and age

To further examine the contents of the representations generated by the three algorithms for faces, we generated RDMs for the age (difference in years), gender and occupation of the different identities and examined their correlations with each of the DNN's representations. Results show that different DNNs capture different aspects of information from the faces (Extended data Figure 2). The visual DNN (VGG) was moderately correlated with age (r = 0.27, p < .001, two-sided, CI: 0.13-0.40) and gender (r = 0.43, p < .001, CI: 0.30-0.54), but not occupation (r = 0.03, p = .650, two-sided, CI: -0.01-0.17). Age (r = 0.23, p = .001, CI: 0.09-0.36), gender (r = 0.33, p < .001, two-sided, CI: 0.2-0.45) and occupation (r = 0.50, p < .001, two-sided, CI: 0.13-0.4), were all similarly correlated with the visual-semantic DNN (CLIP). The semantic DNN (SGPT) was not correlated with Age (0.03, p = .691, two-sided, CI: -0.11-0.17), moderately correlated with Gender (r=0.27, p < .001, two-sided, CI: 0.13-0.4) and highly correlated with Occupation (r = 0.76, p < .001, two-sided, CI: 0.69-0.81). These results show that the different algorithms capture different types of information about the familiar identities.

We also assessed the additional variance of gender, occupation and age add beyond the three algorithms by computing the partial correlations between the algorithms and human mental representations and adding each of these RDMs serially (Extended data Figure 4). We first included gender and found that the contributions of the three algorithms remained roughly the same. When adding occupation, the contribution of SGPT decreased, indicating that it primarily reflects information about the identities' occupation. Further adding age did not change this pattern of results. The statistical tests of the partial correlations of each of the DNNs are reported in Extended data Table 1. Thus, the information that the visual and visual-semantic algorithms account for in the representations in perception and memory goes beyond age and gender, whereas the representation of the semantic DNN (SGPT) is primarily accounted for by the identities' occupation.

Supplementary methods of Experiment 1B

Training a model to generate StyleGAN's embeddings:

To train this model, we used a subset of images from the CelebA-HQ dataset [1], all of which were identities familiar to CLIP (tested with zero-shot classification of name-face images described in the methods section), and none of them depict identities used in the main experiment. We mapped the embeddings of these images according to VGG and CLIP to StyleGAN's embeddings. To obtain this mapping we trained this model in the following steps:

- Each image embedding was paired with the StyleGAN embedding describing the same image, which was calculated in a process termed StyleGAN inversion [2] by using the e4e algorithm [3].
- These image pairs were used to train the model to transform a DNN (VGG or CLIP) embedding to a StyleGAN embedding. We trained 18 independent normalizing flows using the RealNVP architecture [4], mapping between the 512-dimensional latent variable of each input layer of StyleGAN [5] , conditioned on the image's representation according to each DNN, to the 512-dimensional multivariate normal distribution.
- All RealNVP models were optimized using the Adam optimization algorithm [6] with a learning rate of 1e-5 and default PyTorch [7] parameters for 400,000 iterations with batch size of 12 images. Learning rate reduced by a factor of 10 after 200,000 iterations. Image reconstruction from CLIP embeddings was done with representations from the ViT-B/32 architecture [8].

We ran a similar procedure with a model that was trained on different datasets. We used CelebAHQ full dataset, VGGFace2 dataset (familiar to CLIP) or CelebA full dataset. We trained the model between StyleGAN and CLIP using each of these datasets and the procedure described above. To test if the different trained models generated similar images we generated an RDM using VGG-16 for each set and calculated the correlations between these RDMs. The correlations were very high (r = 0.89-95), indicating that the models generated similar representations. Extended data Figure 5 shows the generated images based on CLIP and VGG models.

Supplementary results of Experiment 1B

Recognition of VGG and CLIP-generated faces:

To test if the generated images can be recognized by humans, we performed a two-stage face recognition task: Participants either were assigned to recognize VGG-generated faces or CLIP-generate faces. The task included two stages: a free recall recognition and a face-name matching from a list of names. All the participants conducted the free recall task prior to the face-name matching task. In the free-recall task, the participants were presented with a set of face images and for each image, they had to indicate to which familiar person the face is mostly similar. The participants could also write a description of this person if they could not remember their name or to indicate that they do not recognize the face. Following the free recall, the participants performed face name matching. In this task, they were presented with the same set of face images, and each image was presented with a list of names of highly famous identities. They were asked to indicate for each face, to which of the identities in the list, the face is most similar. All female faces were presented with the same 18 names (9 names of faces that were included in our task and 9 names of novel identities) and all male faces were presented with the same 22 names (11 names of faces that were included in our task and 11 names of novel identities). The full list of names in shown in Supplementary Table 2. After they completed both recognition tasks, the participants were asked to indicate for the original 20 identity images whether they were familiar with them before the experiment. The experiment lasted about 15 minutes.

We excluded from the analysis trials that included generated images that are based on identities that the participants were not familiar with. In addition, in a few cases participants indicated that they are not familiar with an identity based on its original image but did match the face to the correct name or recognize them correctly based on the DNN-generated image. In such cases we marked them as a correct response and included these trials in the analysis.

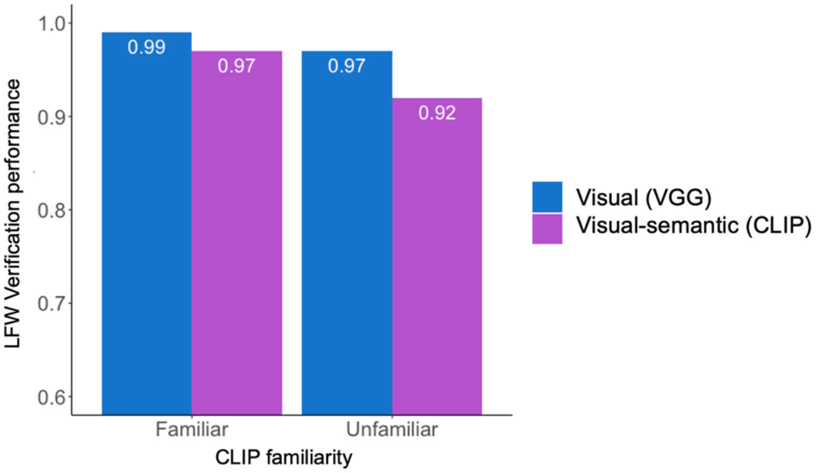
We first assessed whether the VGG- and CLIP-generated faces can be recognized as the original identities by humans. Participants were presented with each VGG- or CLIP-generated face and were asked to write their name or if they do not recall the name to write any unique semantic information that they know about them. We computed the proportion of participants who correctly recognized each face image. The VGG-generated faces were recognized by an average of 51% (median: 59%; range: 0-90%) of the participants across the 20 face images, and CLIP-generated faces were recognized by an average of 31% (median: 28%; range: 0-60%) of the participants across the 20 face images. To assess whether performance is improved in a face-name matching task, in a second phase of the experiment participants were presented with a list of different names of celebrities from the same gender (18 females and 22 males) and were asked to match the face image to the name of the person that is most similar to them. The correct name was selected by the majority of the participants for 18/20 images in VGG-generated set and for 19/20 in CLIP-generated set. The VGG-generated faces were correctly matched to their names by an average of 76% (median: 84%; range: 5-100%) of the participants across the 20 face images, and CLIP-generated faces were correctly matched to their names by an average of 57% (median: 57%; range: 0-95%) of the participants across the 20 face images. These results show that the VGG-generated faces were recognized better than CLIP-generated faces.

Supplementary results of Experiment 2:

CLIP classification performance for familiar and unfamiliar faces

One limitation of the visual-semantic DNN (CLIP) is that we have no knowledge about the face images that it was trained with. We therefore assessed whether this larger training set of CLIP indeed generates a better representation of face identity than the face-trained face DNN, that we trained on 8749 identities of faces from the VGGFace2 dataset. Face-trained DNN are often tested on benchmarks of faces that are outside of their training set, such as Labeled faces in the Wild[9]. These images are downloaded from the internet and we therefore assessed whether they were included in the training set of CLIP by measuring the distance between the embeddings of their images and the embeddings of their names. We selected a subset of faces that the embedding of their names was closest to (relative to 499 other names) as a set of 200 faces that are familiar to CLIP, and another set of 200 faces for which they were farther away, and there were other names that were closer to the images, as faces that were unfamiliar to CLIP. All the faces were not included in the training set of VGG. We then created a verification test using all the images of each of the selected identities to create same identity pairs and the same number of different identity pairs (a total of 1400 pairs). We then measured the verification performance of CLIP and VGG on each of the tests by measuring the cosine distance between each pair of images and calculate the accuracy of the best threshold for the verification test. Supplementary Figure 1 shows that VGG performance was better than CLIP for faces that they were both unfamiliar with CLIP reached VGG performance for unfamiliar faces only for faces it was familiar with. Thus, CLIP extensive training does not generate a representation that is better for face classification than VGG. These findings suggest that VGG generates a representation of face identity that better generalizes to unfamiliar identities. The visual-semantic representation that CLIP generates is better for the classes it is trained with than for classes that it is not trained with.

We conclude that CLIP, like humans, generates a visual-semantic representation that may not generalize well to unfamiliar faces. Given that the goal of human face recognition is to recognize socially relevant familiar faces, this representation suffices for the purpose of human's social interaction with people.



Supplementary Figure 1: Verification performance for Labeled faces in the wild (LFW) faces that were familiar or unfamiliar to CLIP, based on their face-name embedding distance. Both sets of images were unfamiliar to VGG.

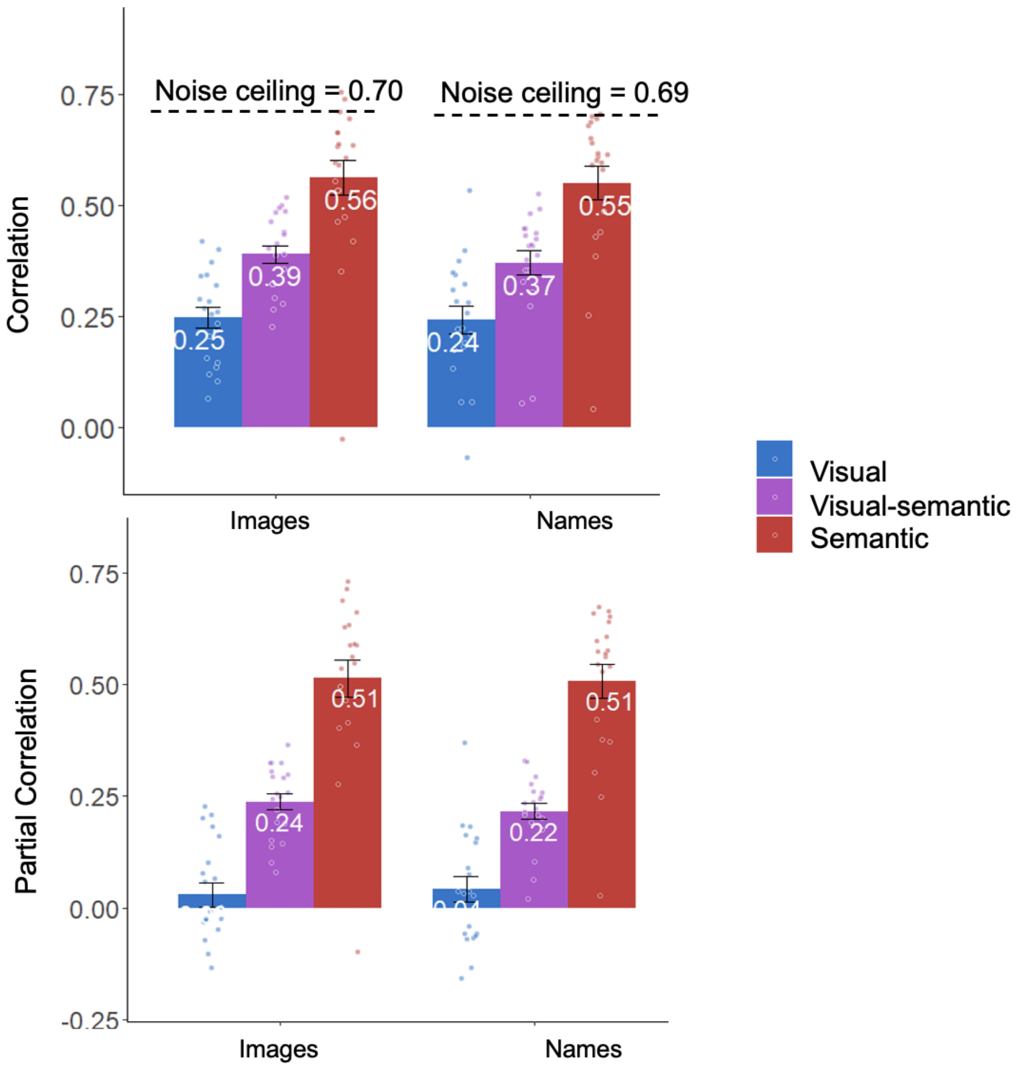
Supplementary results of Experiment 4:

Semantic similarity judgements of objects

For each participant we computed the correlations and the partial correlations of the three DNNs with semantic similarity ratings of objects images or names. The correlations were Fisher's z transformed. Results show significant correlations between the three DNNs and semantic similarity judgements based on images (VGG: r = 0.25, t(19) = 10.82, p < .001. two-sided, CI= 0.2,0.29, Cohen's d = 2.42; CLIP: r = 0.39, t(19) = 19.84, p < .001. two-sided, CI= 0.35, 0.39, Cohen's d = 4.44: SGPT r = 0.56, t(19) = 14.4, p < .001. two-sided, CI= 0.48,0.64, Cohen's d = 3.22) and semantic judgement similarity based on names (VGG: r = 0.24, t(19) = 7.95 p < .001. two-sided, CI= 0.18,0.31, Cohen's d = 1.78; CLIP: r = 0.37, t(19) =0.37, p < .001. two-sided, CI= 0.13,0.43, Cohen's d = 3.01, Cohen's d = 3.01: SGPT r = 0.55, t(19) = 14.55, p < .001. two-sided, CI=0.47,0.63, Cohen's d =

3.25. A 2-way ANOVA of DNN and Task (Perception, Memory) reveal a main effect of DNN F(1.12,42,43) = 94.75, p < .001, two-sided, $\eta_p^2 = 0.71$. Post hoc comparisons reveal a significant larger correlation with the semantic DNN than the visual semantic DNN t(117) = 6.72, p < .001, two-sided, and the visual DNN t(117) = 10.6, p < .001 two-sided, and between the visual-semantic and visual DNN t(117) = 4.16, p < .001, two-sided. Supplementary Figure 2 shows the raw correlations.

To assess the unique contribution of each DNN, we computed the partial correlations. The partial correlations with semantic judgements based on names were significant with CLIP: r = 0.24, t(19) = 12.9, p < .001. two-sided, CI: 0.2,0.28, Cohen's d = 2.89: and SGPT r = 0.51, t(19) = 12.2, p < .0001. two-sided, CI: 0.43,0.6, Cohen's d = 2.75, but not with VGG: r = 0.03, t(19) = 1.07, p = 0.296, two-sided, CI = -.03,0.09, Cohen's d = 0.24. The same pattern was found for semantic judgement based on images: CLIP: r = 0.22, t(19) = 12.23, p < .001. two-sided, CI: 0.18,0.25, Cohen's d = 2.74: SGPT r = 0.51, t(19) = 13.53, p < .001. two-sided, CI: 0.43,0.59, Cohen's d = 3.03 but not with VGG r = 0.04, t(19) = 1.46, p = 0.193, two-sided, CI: -.02,0.1, Cohen's d = 0.33). Thus, both semantic and visual-semantic DNNs uniquely contribute to semantic ratings of familiar objects based on their images or names. Supplementary Figure 2 shows the raw partial correlations.



Supplementary Figure 2: The mean correlations and partial correlations between RDMs of human semantic similarity ratings for objects based on their images (N = 19) or their names (N = 20) and the visual, visual-semantic and semantic representations of the DNNs, show very high correlations between human semantic similarity and SGPT and to a lesser extent CLIP similarity ratings. Error bars indicate the standard error of the mean.

Supplementary Table 1: The first paragraph in Wikipedia of each identity was used to extract embeddings of the semantic representations of the 20 identities from SGPT:

Identity	Text
Angelina Jolie	Angelina Jolie[3] DCMG (born Angelina Jolie Voight,[4] June 4, 1975; later Angelina Jolie Pitt[5]) is an American actress, filmmaker, and humanitarian. The recipient of numerous accolades, including an Academy Award and three Golden Globe Awards, she has been named Hollywood's highest-paid actress multiple times.
Jennifer Aniston	Jennifer Joanna Aniston (born February 11, 1969) is an American actress and producer. The daughter of actors John Aniston and Nancy Dow, she began working as an actress at an early age with an uncredited role in the 1988 film Mac and Me; her first major film role came in the 1993 horror comedy Leprechaun. Since her career progressed in the 1990s, she has become one of the worlds highest-paid actresses.
Judi Dench	Dame Judith Olivia Dench CH DBE FRSA (born 9 December 1934) is an English actress. Regarded as one of Britain's best actresses,[1][2][3] she is noted for her versatile work in various films and television programmes encompassing several genres, as well as for her numerous roles on the stage.[4] Dench has garnered various accolades throughout her career spanning over six decades, including an Academy Award, a Tony Award, two Golden Globe Awards, four British Academy Television Awards, six British Academy Film Awards and seven Olivier Awards.
Kate Winslet	Kate Elizabeth Winslet CBE (born 5 October 1975) is an English actress.[3] Known for her work in independent films, particularly period dramas, and for her portrayals of headstrong and complicated women, she has received numerous accolades, including an Academy Award, a Grammy Award, two Primetime Emmy Awards, three British Academy Film Awards, and five Golden Globe Awards. Time magazine named Winslet one of the 100 most influential people in the world in 2009 and 2021, and in 2012, she was appointed Commander of the Order of the British Empire (CBE).
Keira Knightley	Keira Christina Righton[1] OBE Knightley, born 26 March 1985) is an English actress.[2] She has starred in both independent films and big-budget blockbusters, and is particularly noted for her roles in period dramas. Her accolades include two Empire Awards and nominations for two Academy Awards, three British Academy Film Awards, three Golden Globe Awards, one Screen Actors Guild Award and one Laurence Olivier Award. Knightley was appointed an OBE in the 2018 Birthday Honours for services to drama and charity.[3]
Sandra Bullock	Sandra Annette Bullock (born July 26, 1964) is an American actress and producer. The recipient of various accolades, including an Academy Award and a Golden Globe Award, she was the world's

	highest-paid actress in both 2010 and 2014.[1][2][3] In 2010, she was named one of Time's 100 most influential people in the world.
Hugh Grant	Hugh John Mungo Grant[2] (born 9 September 1960) is an English actor. His awards include a Golden Globe Award, a BAFTA Award, Volpi Cup and an Honorary C?sar. As of 2018, his films have grossed a total of nearly US$3 billion worldwide from 29 theatrical releases.[3]
Michael Caine	Sir Michael Caine CBE (born Maurice Joseph Micklewhite; 14 March 1933) is an English actor. Known for his distinctive South London accent, he has appeared in more than 160 films in a career spanning seven decades, and is considered a British film icon.[2][3] He has received various awards including two Academy Awards, a BAFTA Award, three Golden Globe Awards, and a Screen Actors Guild Award. As of February 2017, the films in which Caine has appeared have grossed over $7.8 billion worldwide.[4] Caine is one of only five male actors to be nominated for an Academy Award for acting in five different decades.[nb 1] He has appeared in seven films that featured in the British Film Institute's 100 greatest British films of the 20th century. In 2000, he received a BAFTA Fellowship and was knighted by Queen Elizabeth II for his contribution to cinema.
Nicolas Cage	Nicolas Kim Coppola (born January 7, 1964),[2][3] known professionally as Nicolas Cage, is an American actor and filmmaker. Born into the Coppola family, Cage is the recipient of various accolades, including an Academy Award, a Screen Actors Guild Award, and a Golden Globe Award.
Robert De Niro	Robert Anthony De Niro Jr. (/d? ?n??ro?/ d? NEER-oh, Italian: [de ?ni?ro]; born August 17, 1943) is an American actor, producer, and director. He is particularly known for his nine collaborations with filmmaker Martin Scorsese, and is the recipient of various accolades, including two Academy Awards, a Golden Globe Award, the Cecil B. DeMille Award, and a Screen Actors Guild Life Achievement Award. In 2009, De Niro received the Kennedy Center Honor, and received a Presidential Medal of Freedom from U.S. President Barack Obama in 2016.
Tom Hankes	Thomas Jeffrey Hanks (born July 9, 1956) is an American actor and filmmaker. Known for both his comedic and dramatic roles, he is one of the most popular and recognizable film stars worldwide, and is regarded as an American cultural icon.[2] Hanks's films have grossed more than $4.9 billion in North America and more than $9.96 billion worldwide,[3] making him the fourth-highest-grossing actor in North America.[4]

Hillary Clinton	Hillary Diane Rodham Clinton (born October 26, 1947) is an American politician, diplomat, lawyer, writer, and public speaker who served as the 67th United States secretary of state from 2009 to 2013, as a United States senator representing New York from 2001 to 2009, and as first lady of the United States from 1993 to 2001 as the wife of President Bill Clinton. A member of the Democratic Party, she was the party's nominee for president in the 2016 presidential election, which she lost to Donald Trump.
Nicola Sturgeon	Nicola Ferguson Sturgeon (born 19 July 1970) is a Scottish lawyer and politician serving as First Minister of Scotland and Leader of the Scottish National Party (SNP) since 2014. She is the first woman to hold either position. She has been a member of the Scottish Parliament (MSP) since 1999, first as an additional member for the Glasgow electoral region, and as the member for Glasgow Southside (formerly Glasgow Govan) from 2007.
Theresa May	Theresa Mary, Lady May[1] (Brasier; born 1 October 1956) is a British politician who served as Prime Minister of the United Kingdom and Leader of the Conservative Party from 2016 to 2019. She served as Home Secretary from 2010 to 2016 in the Cameron government and has been the Member of Parliament (MP) for Maidenhead in Berkshire since 1997. Ideologically, May identifies herself as a one-nation conservative.[3]
Bill Clinton	William Jefferson Clinton (Blythe III; born August 19, 1946) is an American politician who served as the 42nd president of the United States from 1993 to 2001. He previously served as governor of Arkansas from 1979 to 1981 and again from 1983 to 1992, and as attorney general of Arkansas from 1977 to 1979. A member of the Democratic Party, Clinton became known as a New Democrat, as many of his policies reflected a centrist "Third Way" political philosophy. He is the husband of Hillary Clinton, who was a senator from New York from 2001 to 2009, secretary of state from 2009 to 2013 and the Democratic nominee for president in the 2016 presidential election.
Boris Johnson	Alexander Boris de Pfeffel Johnson (born 19 June 1964) is a British politician serving as Prime Minister of the United Kingdom and Leader of the Conservative Party since 2019. He was Secretary of State for Foreign and Commonwealth Affairs from 2016 to 2018 and Mayor of London from 2008 to 2016. Johnson has been Member of Parliament (MP) for Uxbridge and South Ruislip since 2015 and was previously MP for Henley from 2001 to 2008.
David Cameron	David William Donald Cameron (born 9 October 1966) is a British politician, businessman, lobbyist, and author who served as Prime Minister of the United Kingdom from 2010 to 2016. He was Member of Parliament (MP) for Witney from 2001 to 2016 and leader of the Conservative Party from 2005 to 2016. He identifies as a one-nation

	conservative, and has been associated with both economically liberal and socially liberal policies.
Donald Trump	Donald John Trump (born June 14, 1946) is an American politician, media personality, and businessman who served as the 45th president of the United States from 2017 to 2021.
George Bush	George Walker Bush (born July 6, 1946) is an American politician who served as the 43rd president of the United States from 2001 to 2009. A member of the Bush family and son of former president George H. W. Bush, he previously served as the 46th governor of Texas from 1995 to 2000 as part of the Republican Party.
Tony Blair	Sir Anthony Charles Lynton Blair KG (born 6 May 1953) is a British politician who served as Prime Minister of the United Kingdom from 1997 to 2007 and Leader of the Labour Party from 1994 to 2007. On his resignation he was appointed Special Envoy of the Quartet on the Middle East, a diplomatic post which he held until 2015. He has been the executive chairman of the Tony Blair Institute for Global Change since 2016. As prime minister, many of his policies reflected a centrist "Third Way" political philosophy.[b] He is the only living former Labour leader to have led the party to a general election victory; and one of only two in history to form three majority governments, the other being Harold Wilson.

Supplementary Table 2: The face-name matching task was performed against this list of names.

Female names:	Male names:
Angela Merkel	Arnold Schwarzenegger
Angelina Jolie	Bill Clinton
Hillary Clinton	Boris Johnson
Jennifer Aniston	Clint Eastwood
Judi Dench	David Cameron
Julia Roberts	Donald Trump
Kate Middleton	George Clooney
Kate Winslet	George WBush
Keira Knightley	Hugh Grant
Meryl Streep	Joe Biden
Nicola Sturgeon	Leonardo Dicaprio
Nicole Kidman	Michael Caine
Penelope Cruise	Nicola Sarkozy
Queen Elizabeth	Nicolas Cage
Reese Witherspoon	Prince William
Sandra Bullock	Richard Gere
Sarah Jessica Parker	Robert De Niro
Theresa May	Sylvester Stallone
	Tom Cruise
	Tom Hanks
	Tony Blair

Supplementary Table 3: The semantic representations from SGPT are based on the following definitions of each object.

Object	Text
Apple	A round fruit with shiny red or green skin that is fairly hard and white inside
Ball	A round object used for throwing, hitting or kicking in games and sports
Banana	A long-curved fruit with a thick yellow skin and that is soft inside, which grows on trees in hot countries
Baseball Bat	A baseball bat is a smooth wooden or metal club used in the sport of baseball to hit the ball after it is thrown by the pitcher.
Book	A set of printed pages that are fastened inside a cover so that you can turn them and read them
Chopsticks	A pair of thin sticks that are used for eating with, especially in some Asian countries
Clock	An instrument for measuring and showing time, in a room, on the wall of a building or on a computer screen (not worn or carried like a watch)
Comb	A flat piece of plastic or metal with a row of thin teeth along one side, used for making your hair neat; a smaller version of this worn by women in their hair to hold it in place or as a decoration
Cucumber	A long vegetable with dark green skin that is light green inside, usually eaten raw
Cup	A small container that is like a bowl in shape, usually with a handle, used for drinking tea, coffee, etc.
Drum Sticks	A stick used for playing a drum
Drum	A musical instrument made of a hollow round frame with plastic or skin stretched tightly across one or both ends. You play it by hitting it with sticks or with your hands.
Flute	A musical instrument of the woodwind group, like a thin pipe in shape. The player holds it to the side of his or her face and blows across a hole at one end.
Hairdryer	A small machine used for drying your hair by blowing hot air over it
Orange	A round citrus fruit with thick skin of a colour between red and yellow and a lot of sweet juice
Pen	An instrument made of plastic or metal used for writing with ink
Plate	A flat, usually round, dish that you put food on
Spoon	A tool that has a handle with a shallow bowl at the end, used for mixing, serving and eating food

Tennis Racket	The racket that you use when you play tennis
Violin	a musical instrument with strings, which you hold under your chin and play with a bow

1. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings (2017) doi:10.48550/arxiv.1710.10196.
2. Xia, W. et al. Gan inversion: A survey. IEEE Trans Pattern Anal Mach Intell (2022).
3. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O. & Cohen-Or, D. Designing an encoder for StyleGAN image manipulation. ACM Transactions on Graphics (TOG) (2021) doi:10.1145/3450626.3459838.
4. Dinh, L., Sohl-Dickstein, J. & Bengio, S. Density estimation using Real NVP. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (2016) doi:10.48550/arxiv.1605.08803.
5. Abdal, R., Qin, Y. & Wonka, P. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? Proceedings of the IEEE International Conference on Computer Vision 2019-October, 4431–4440 (2019).
6. Kingma, D. P. & Ba, J. L. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2014) doi:10.48550/arxiv.1412.6980.
7. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv Neural Inf Process Syst 32, (2019).
8. Radford, A. et al. Learning Transferable Visual Models From Natural Language Supervision. (2021).
9. Huang, G. B. & Learned-miller, E. Labeled faces in the wild : Updates and new reporting procedures. University of Massachusetts Amherst Technical Report 13–14 (2014).