ARTICLES
https://doi.org/10.1038/s42256-019-0111-7
nature machine intelligence

# Deep convolutional neural networks in the face of caricature

Matthew Q. Hill[1*], Connor J. Parde[1], Carlos D. Castillo[2], Y. Ivette Colón[1], Rajeev Ranjan[2], Jun-Cheng Chen[2], Volker Blanz[3] and Alice J. O'Toole[1]

**Real-world face recognition requires us to perceive the uniqueness of a face across variable images. Deep convolutional neural networks (DCNNs) accomplish this feat by generating robust face representations that can be analysed in a multidimensional 'face space'. We examined the organization of viewpoint, illumination, gender and identity in this space. We found that DCNNs create a highly organized face similarity structure in which identities and images coexist. Natural image variation is organized hierarchically, with face identity nested under gender, and illumination and viewpoint nested under identity. To examine identity, we caricatured faces and found that identification accuracy increased with the strength of identity information in a face, and caricature representations 'resembled' their veridical counterparts—mimicking human perception. DCNNs therefore offer a theoretical framework for reconciling decades of behavioural and neural results that emphasized either the image or the face in representations, without understanding how a neural code could seamlessly accommodate both.**

People recognize faces across changes in viewpoint, illumination, facial expression and appearance (for example, glasses, facial hair). The nature of the visual representation that supports this skill, however, is unknown, despite decades of research in psychology and neuroscience[1–5]. Historically, alternative hypotheses about face representations posited that the primate visual system reconstructs an object-centred facsimile of a face[1,6] or that it represents multiple image-based views of faces[3,4,7]. Computational models built on these hypotheses illustrate clearly the benefits and pitfalls of object-centred and image-based face representations.

In early image-based models, principal component analysis (PCA) was applied to sets of face images[8] to create a face space[9]. This model accounts for the human recognition costs incurred when imaging conditions change between learning and test[10]. It also provides insight into the gender[11], race[12], feature[13] and identity[11] information in face images. However, image-based PCA works only when the learned and test images are taken under similar conditions.

The limitations of image-based models led to the development of three-dimensional (3D) morphable models[14], which represent faces rather than images of faces. These models operate on densely sampled shape and pigmentation information from laser scans of faces. As with image-based models, a face space is created by applying PCA to sets of faces. In this space, individual identities are defined as trajectories that radiate out from the average face. As a face moves away from the average along its identity trajectory, it becomes increasingly distinctive, changing from anti-caricature to veridical, and then to caricature. The effect of this manipulation is to exaggerate differences between individual faces and an 'average' face. The paradox of caricatures is that they portray a good likeness of a person with a distorted image. Psychological studies indicate that caricature-based distortions do not hinder, and in some cases even enhance, human perception of face identity[15–19]. Morphable models provide a transparent account of human caricature perception[9,20]. However, they provide no mechanism to account for the human recognition costs incurred when images differ in viewpoint or illumination.

Over the years, object-centred and image-based models have made progress on the problem of generalized face/object recognition. However, neither provides a unified account of how the visual system simultaneously discriminates facial identities while managing (filtering out or encoding) image and appearance variation. Following these early approaches, practical progress on automatic face recognition was made primarily via the use of increasingly sophisticated techniques for extracting features from face images (for example, see refs. [21–23]), including features inspired by human vision[3,24]. Theoretical progress came from Bayesian approaches to face[25] and object recognition, which highlighted the importance of statistical priors in constraining solutions to generalized visual recognition. However, problems in obtaining adequate priors limited the practical application of these models.

Deep convolutional neural networks (DCNNs) are now the state of the art in machine-based face recognition, because they can generalize identity across variable images[26–30]. These networks are modelled after the primate visual system[31,32] and consist of multiple layers of simulated neurons that perform nonlinear convolution and pooling operations. DCNN representations expand in early layers of the network, but are compressed in the top layers through a bottleneck of neurons. The representation of facial identity that emerges at the final layer of a DCNN is compact and can operate robustly over changes in image parameters (for example, viewpoint) and appearance.

In DCNNs, decades of progress in face recognition models are brought to bear. These include critical aspects of prior probabilities, as well as elements of object-centred and image-based representations. DCNNs learn features from statistical mappings between images and label-based categories by training with large-scale datasets. This replaces preset features (for example, ref. [21]) with features learned from training data. The use of learned features incorporates prior probabilities to build a system with general knowledge of faces. Similar to object-centred models, DCNNs represent identity with a code that is robust to changes in view and illumination. Similar to image-based models, the representations retain information

[1]The University of Texas at Dallas, Richardson, TX, USA. [2]University of Maryland, College Park, MD, USA. [3]University of Siegen, Siegen, Germany.
*e-mail: matthew.hill@utdallas.edu

about the images they process[33,34]. Specifically, features from the top layer of DCNNs trained for face recognition support reliable linear readout of the viewpoint (yaw, pitch) of the input image[33]. This is consistent with a theoretical approach proposing that primate visual processing 'untangles' non-category information in an image, rather than removing it outright[35]. This theory has been supported by object recognition studies showing that high-performing object recognition DCNNs retain image information in their outputs[36,37].

Deep networks offer a proof of principle that a robust coding of high-level visual information can coexist with instance-based codes that retain characteristics of the imaging conditions. But how do DCNN codes accomplish the balancing act of accommodating facial identity and image information in a unitary representation? It has been difficult to directly address this question, because viewpoint, illumination and the number/quality of images for each identity are not controlled in the datasets used to train DCNNs. To overcome this challenge, we probed a network trained with 'in-the-wild' face images using an 'in-the-lab' dataset. Specifically, we used highly controlled laser scans of faces to examine how DCNNs represent faces in terms of subject parameters (identity, gender) and image characteristics (viewpoint, illumination).

To probe the representation of identity in a DCNN, we manipulated the strength of identity information in a face by caricaturing. Historically, artists have created caricatures by exaggerating the features in an individual's face that are distinctive relative to a population expectation. For example, caricatures of Angelina Jolie show exaggerated lip fullness and cheekbones. Computer-generated caricatures (for example, refs. [14,38]), which have existed since the mid-1980s[38], systematically distort face features relative to a computed average face. Decades of face recognition experiments indicate that people see caricatures as a 'good likeness' of the face[39]. Moreover, caricatures, despite their distortion, are recognized as well as, or more accurately than, veridical faces[39].

Manipulating the strength of identity information in a face through caricaturing and anti-caricaturing is a powerful tool for understanding how DCNNs code face identity. Given the psychological data, caricatures allow us to relate DCNN identity codes to those created by the human visual system. Specifically, caricatures are spatially and chromatically distorted in the image domain, but human perception creates an identity equivalence for these distortions. To model human perception, DCNNs should show identity constancy between faces and their caricatures. Moreover, this constancy should prevail over changes in viewpoint and illumination.

## Results

**Face space visualization.** We examined the organization of imaging characteristics and subject variables in the DCNN top-layer face representation using a face space framework[9,40]. In this framework, the distance between points in the space reflects the similarity of face images as 'perceived' by the top layer of the DCNN. We report data on a 101-layered face identification DCNN[30] trained with 5,714,444 in-the-wild images (Fig. 1a, for illustration) of 58,020 identities. The top-layer output of the network is a 512-element face representation. To demonstrate the robustness of the results across network architecture variation, we report a replication of these experiments on an architecturally distinct network (Supplementary Figs. 5–12 and Supplementary Tables 2, 3 and 5).

Images were created from laser scans of 70 male and 70 female heads registered to a parametric 3D face model[14]. Each face was rendered from five viewpoints (yaw: 0° (frontal), 20°, 30°, 45° and 60° (left profile)) under two illumination conditions (ambient and directional spotlight). This produced 1,400 images (Fig. 1b), which we processed through the DCNN to produce a top-layer representation for each image. To examine the structure and information content of the face space that emerges at the top layer of units in DCNNs, we visualized the face space representations of the 1,400 images using *t*-distributed stochastic neighbour embedding (*t*-SNE)[41].

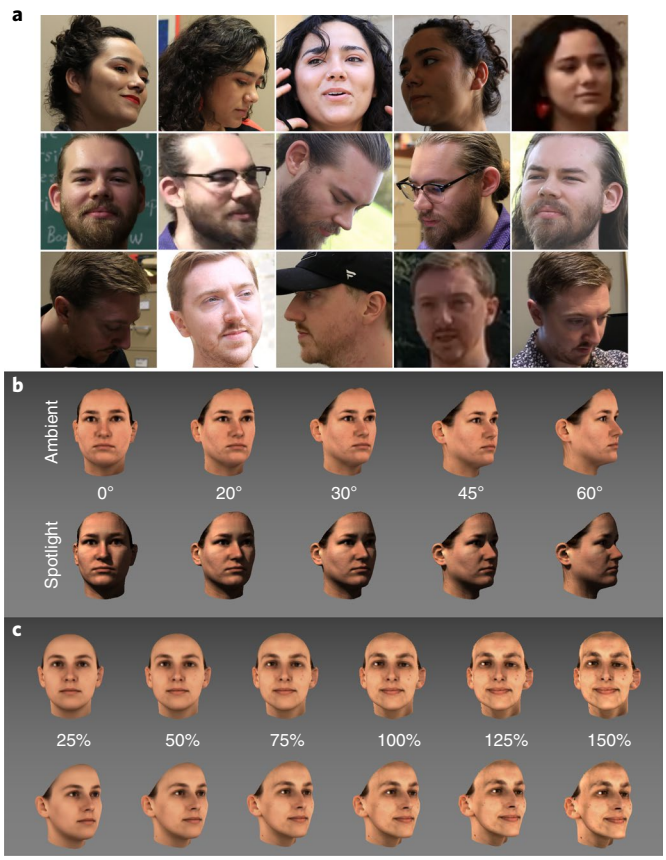


**Fig. 1 | Examples of face images. a**, Training was carried out on real-world unconstrained face images, similar to those pictured. **b**,**c**, Testing was performed on highly controlled laser-scan data varying by viewpoint and illumination (**b**) and identity strength (**c**).

Figure 2 shows the hierarchical organization of the face space with respect to gender, identity, illumination and viewpoint. Identities were separated with high accuracy (area under the curve (AUC) $\approx$ 1), indicating that the DCNN recognizes faces across substantial variability in viewpoint and illumination. The space is separated roughly into two clusters, by gender (Fig. 2a). Within each identity cluster, face images from the two illumination conditions separate into sub-clusters (Fig. 2b,e). Within each illumination subcluster, images are arranged systematically by viewpoint, like beads on a chain (Fig. 2c,f). This demonstrates a highly organized representation of image information in a robust identity code. For direct distance measures of this hierarchical structure in the full-dimensional space, see Supplementary Fig. 1 and Supplementary Table 1.

Next, we quantified the accessibility of gender, illumination, and viewpoint in the full high-dimensional space, using a linear classifier. All three variables were predicted accurately from the face representations ($P < 0.001$ in all cases). Viewpoint was detected with an average error of 6.34° (s.d. = 4.95°), illumination classification was 95.21% correct and gender classification was 98.21% correct (Supplementary Fig. 2). This demonstrates the accurate linear readout of image and subject information from the high-dimensional top-layer face representation. Notably, this indicates that the structure of the image-based information across identities resides in systematic directions in the high-dimensional space that do not explain sufficient variation to appear in a 2D projection.

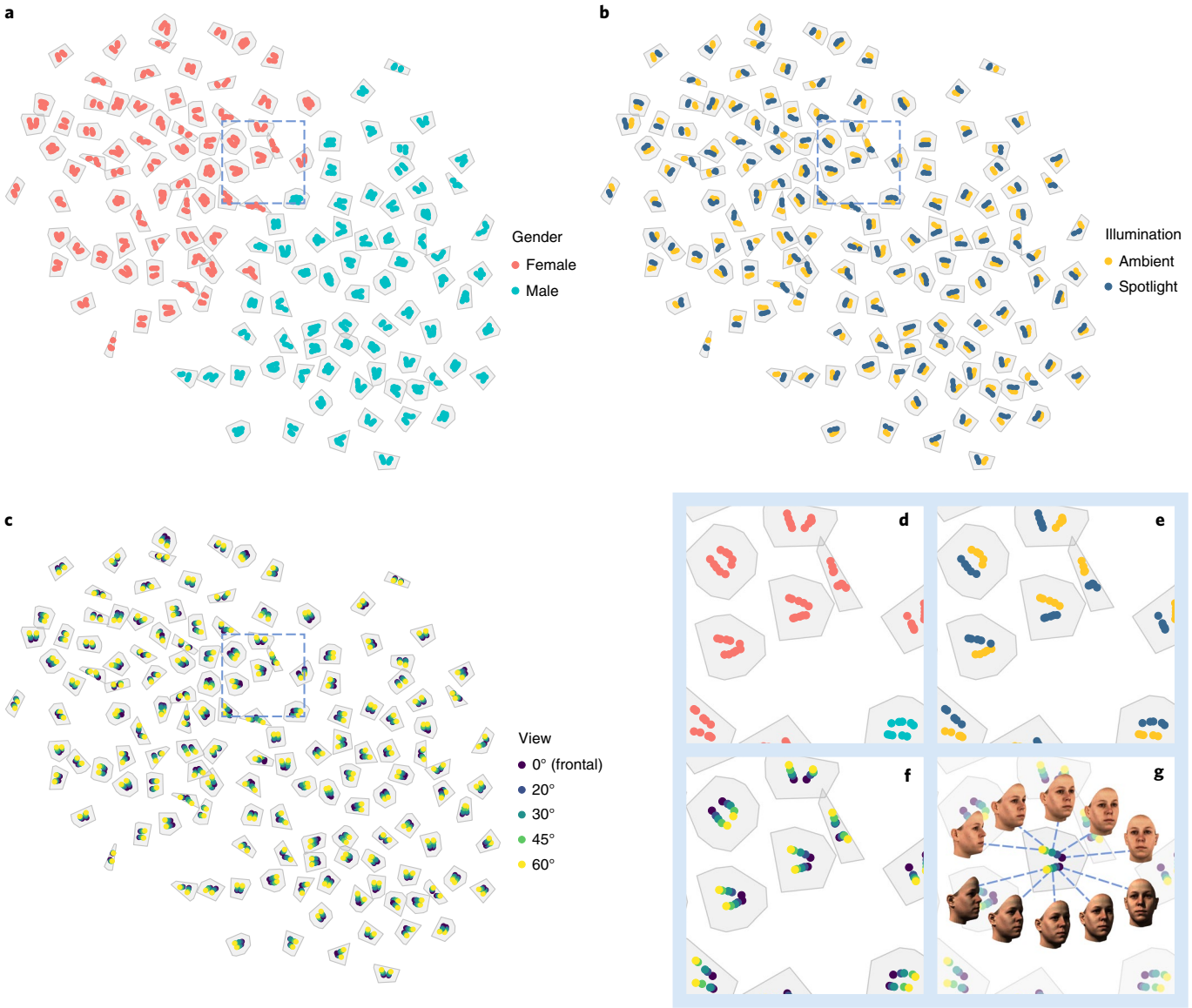


**Fig. 2 | Visualization of the top-level DCNN similarity space for all images. a–f**, Points are coloured according to different variables. Grey polygonal borders are for illustration purposes only and show the convex hull of all images of each identity. These convex hulls are expanded by a margin for visibility. The network separates identities accurately. In **a**,**d**, the space is divided into male and female sections. In **b**,**e**, illumination conditions subdivide within identity groupings. In **c**,**f**, the viewpoint varies sequentially within illumination clusters. Dotted-line boxes in **a**–**c** show areas enlarged in **d**–**g**.

**Identity strength in the face space.** To examine facial distinctiveness, we used the 3D head model to generate morphs that varied in the strength of the identity information (that is, caricature level) in the face[40]. Following the identity trajectory, each face was morphed from a caricature (high identity strength) to a near-average face (low identity strength) in five equal steps. This yielded six versions of each face (150%, 125% (caricature); 100% (veridical); 75%, 50%, 25% (anti-caricature)). The addition of identity strength increased the dataset to 8,400 images (Fig. 1c).

Figure 3 shows the *t*-SNE face space with the inclusion of identity strength variation. Faces with weak identity information are grouped according to other variables (gender, view, illumination). Specifically, Fig. 3a shows that mixed-identity clusters are scattered among correctly clustered identities and Fig. 3b shows that mixed-identity regions contain only faces with weak identity strength. Each identity-mixed cluster contains images of a single viewpoint (Fig. 3c), nested within a single illumination condition (Fig. 3d), and within a gender group (Fig. 3e). Enlarged view sections (Fig. 3f–h) show that within an identity cluster, images divide by illumination conditions (Fig. 3f). Viewpoints also divide, with caricature levels arranged in string-like groups (Fig. 3g). Caricatures are centred in identity clusters (Fig. 3h), showing that same-identity caricatures cluster more closely over image variation than veridicals and anti-caricatures (see Supplementary Information).

**Caricature and identity.** Face identification amounts to a decision as to whether two images depict the same or different identities. This decision is based on the cosine similarity between the top-layer representations of the two images (higher similarities suggest the same identity). Accuracy can be visualized using the similarity distributions for same- and different-identity image pairs (wider separation indicates higher accuracy).

Figure 4a shows that caricaturing improves the network's identification accuracy by increasing the 'perceptual' contrast between faces as caricature level increases. This is seen as a leftward drift of the different-identity distribution (Supplementary Fig. 1 and

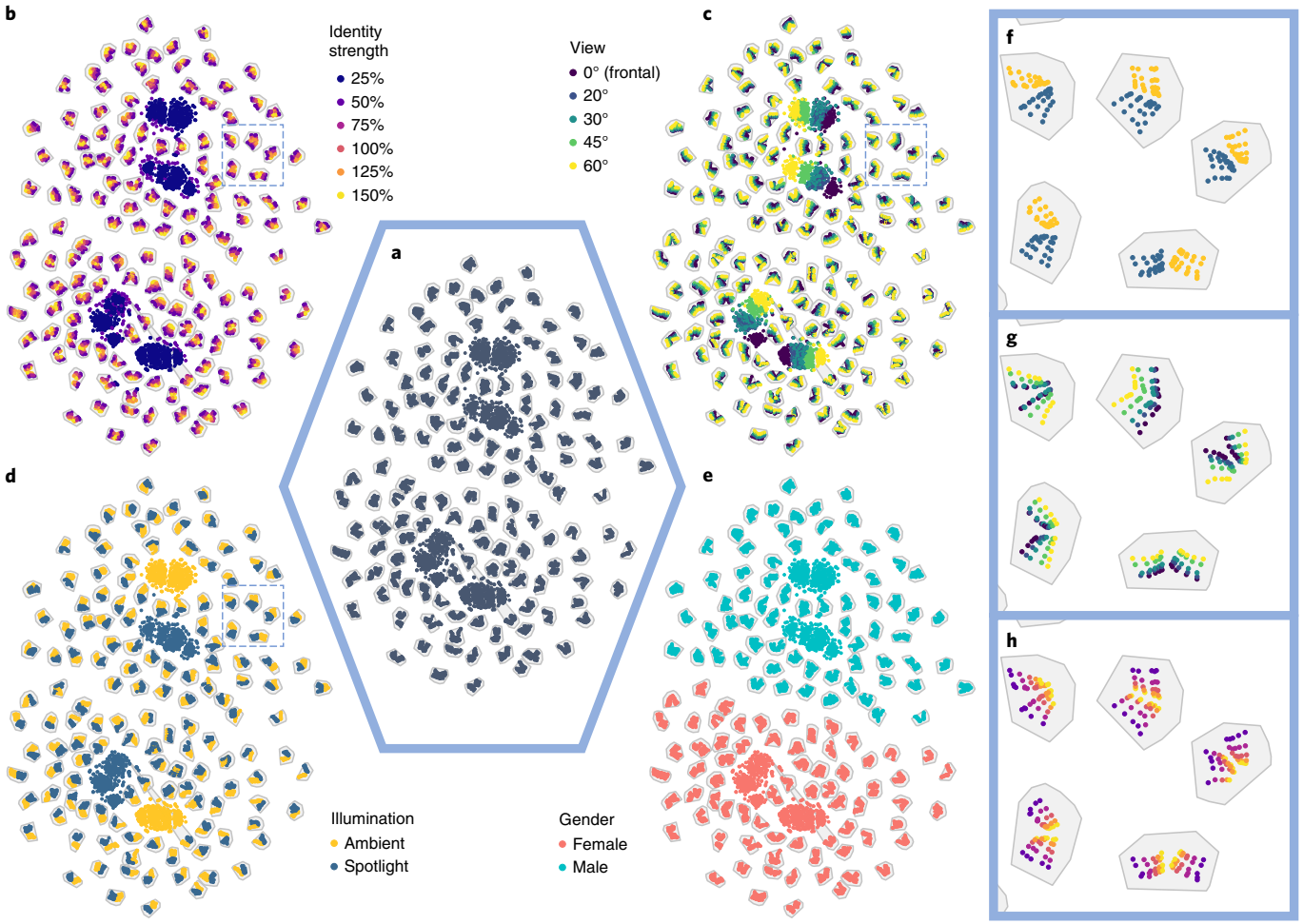


**Fig. 3 | Visualization of top-level similarity space with identity strength variation. a**, Grey polygons surround identity-constant clusters, which appear among mixed-identity clusters (not enclosed in polygons). Each polygon is calculated as the convex hull of all images of an identity, where identity strength is 75% or greater. The polygons are provided only for the purpose of visualization. **b**–**e**, The same plot as displayed in **a**, but colour-coded according to the different variables. Points colour-coded by identity strength show that mixed-identity regions contain weak identity strength images (**b**). Points colour-coded by viewpoint show that each identity-mixed cluster contains images of a single viewpoint (**c**). Points colour-coded by illumination show that viewpoints are nested within illumination conditions (**d**). Points colour-coded by gender show that the gender division of the space is retained with identity strength manipulation (**e**). **f**–**h**, Enlarged view sections show that within an identity-constant cluster, images divide by illumination conditions (**f**), viewpoints divide with caricature levels arranged in string shapes (**g**) and caricatures fall in the centre of the identity cluster (**h**).

Supplementary Table 1). Caricaturing does not appreciably move the same-identity distribution. However, consistent with its effects on minimizing the impact of imaging parameters (Fig. 3h), the range of similarity values in this distribution compresses toward the upper bound as caricature level increases (Supplementary Fig. 3 and Supplementary Table 4). Although the accuracy benefit here is modest due to a ceiling effect, larger performance gains would be expected if the data were more challenging (for example, more similar identities, more diverse imaging conditions).

Next, we asked whether the DCNN 'sees' the caricature as the same identity as its corresponding veridical face. Figure 4b indicates that it does. We looked at the similarity between veridicals and their corresponding images across caricature levels. The network perceives 75% anti-caricatures and caricatures as nearly equivalent to veridicals (Fig. 4b). The 25% and 50% anti-caricatures are less similar to their veridical faces. The caricatures (125%, 150%) are clustered with the correct identities. These data indicate that the difference in identity grouping for the caricatures (125% and 150%) is not due to distortion level, per se, but rather to the type of distortion. This is supported by the fact that identity-strength distortions of equal magnitude (50% to 100% and 100% to 150%) result in different similarity score outcomes (Fig. 4b).

Caricaturing, therefore, affects DCNN perception by exaggerating a face's unique identity information relative to other faces in the population without impairing identity perception.

**Caricature and image conditions, viewpoint and illumination.** How does image-based information interact with identity constancy? Figure 5 shows that imaging conditions affect the DCNN's perception of face similarity. Changes in viewpoint and/or illumination can be seen as peaks in the similarity score distributions for same-identity pairs (top row), at all levels of caricature. For higher identity strengths (≥75%), different-identity distributions (bottom row) separate visibly from same-identity distributions, and the salience of image-based similarity is attenuated. This shows that identity—not imaging condition—is the primary determiner of dissimilarity for different-identity pairs. Imaging condition effects reappear with weak identity strengths (≤50%). These near-average faces approach a single (average) identity that varies only by imaging condition. Therefore, similarity in the DCNN encompasses both

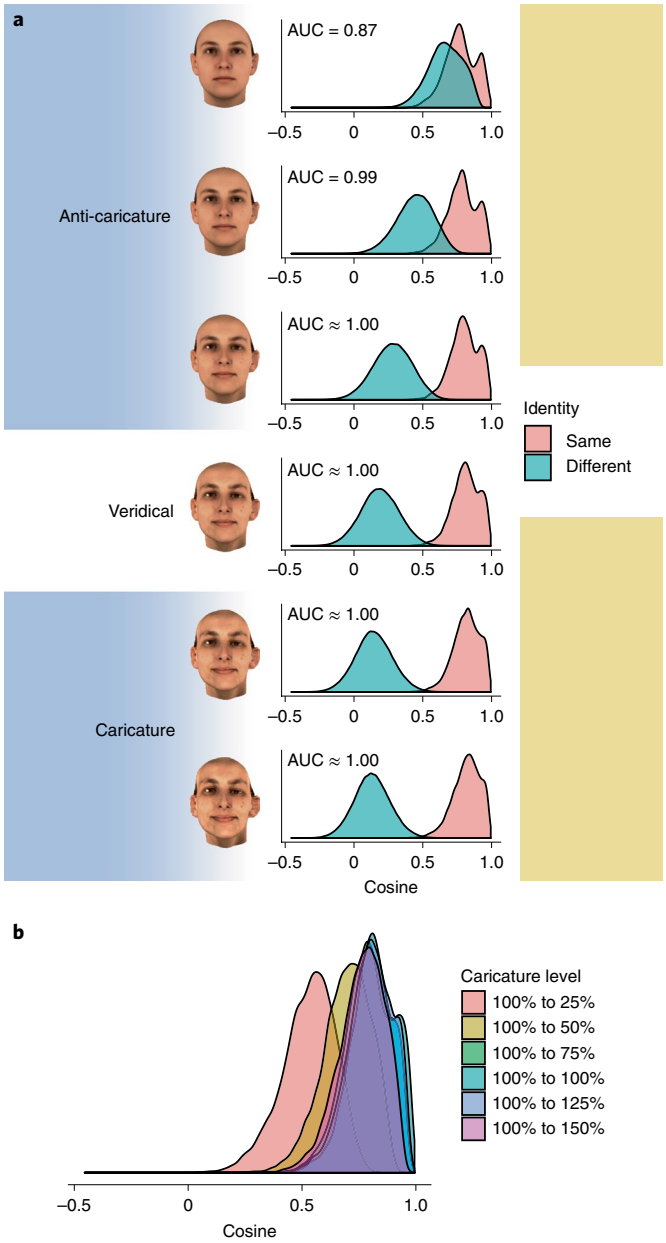


**Fig. 4 | Caricature effects. a**, Image-pair similarity score distributions show that accuracy increases with caricature level. This is due to the greater dissimilarity of caricatures to other identities (leftward drift of the different-identity distribution). **b**, Similarity distributions of same-identity pairs at differing levels of caricature, each compared to the veridical face. Low-identity-strength images (25% and 50%) are less similar to the veridical (pink and yellow distributions), whereas high-identity-strength images (75–150%) are largely equivalent to the veridical (green through purple).

identity and viewing conditions, but on a different scale. Identity contributes far more than image conditions.

## Discussion

Deep networks accomplish the balancing act of accommodating facial identity and image information in a unitary representation by generating an elegantly organized face similarity space. To understand this organization, we distinguish between person properties (for example, identity, gender and race) and specific images (for example, viewpoint and illumination). The former, immutable characteristics divide the face space into sub-space partitions that are homogeneous with respect to their defining person characteristics. The latter, variable properties, are accommodated within the homogeneous subspaces, yet they apply to all identities.

Being able to access information about object/person properties at multiple levels of abstraction is a computational goal of a visual categorization system[42]. In psychological terms, the topology of the DCNN space organizes faces to allow easy access to person properties at different levels of abstraction. The space itself defines a basic-level category of faces. The position of an image in the face space indicates a subordinate gender category, and the position in this gender category specifies an exemplar category of identity[42,43].

To access the fundamental person property of identity, the DCNN must code the uniqueness of a face across variable image conditions. The robust nature of this identity information in a DCNN is inherited from the topology of its similarity space, which is highly nonlinear with respect to image properties. Two widely different images (for example, frontal versus profile) are coded as similar, because the network represents identity categorically.

The use of caricaturing to probe the organization of the face space provides a unique vantage point for seeing how identity and image information interact in a DCNN. Caricaturing affects DCNN performance because it operates both within individual identity clusters and at the level of face populations. Within identity clusters, caricatured faces minimize the influence of imaging parameters. At the population level, caricaturing increases the separation between identities in the space, making them less confusable. The increased distinctiveness of caricatures is primarily the result of the leftward drift of the different-identity distribution. To a lesser extent, this is augmented by the compression of the same-identity distribution toward the maximum cosine value of one (that is, same identity).

From a psychological perspective, the DCNN's combined representation of identity and actual images provides a unified account of behavioural effects seen historically as evidence for exclusively image-based or object-centred theories of face processing. DCNN representations are compatible with a face recognition cost for changes in image parameters between learning and testing. They are also compatible with effects of face distinctiveness relative to a population. The accord between behavioural results and deep network representations, combined with the network's ability to produce a robust representation of identity, makes a DCNN a plausible model of human face processing.

In the present work, we address recognition of identities 'unfamiliar' to the network. Specifically, the DCNN uses its general face knowledge to process novel faces. Future work could address how 'familiarity' with a particular face, via exposure to in-the-wild images, alters the face representation to generalize recognition further. Multiple stages of DCNN training can be targeted in this endeavour[34]. Moreover, we have focused on the top-layer representation in assessing general accord to human face processing. This does not preclude equal or better accord at lower network layers, as has been found for different tasks in object recognition (for example, see ref. [44]). For face processing, additional research is needed to address questions about whether particular network layers model particular behavioural results.

From a neuroscience perspective, DCNN representations reconcile the seemingly paradoxical nature of ventral temporal cortex organization as both object-categorical and reflective of low-level image properties such as viewpoint[45,46], illumination[47], size[48] and position[49]. For the former, structure exists in the organization of person properties in subspaces. For the latter, structure within identity subspaces is duplicated across identities to index image properties. Although there are compelling similarities between DCNNs and the human visual system, there are also key differences. For example, DCNNs are supervised, whereas learning in neural systems is probably unsupervised. Notwithstanding this, the human–network

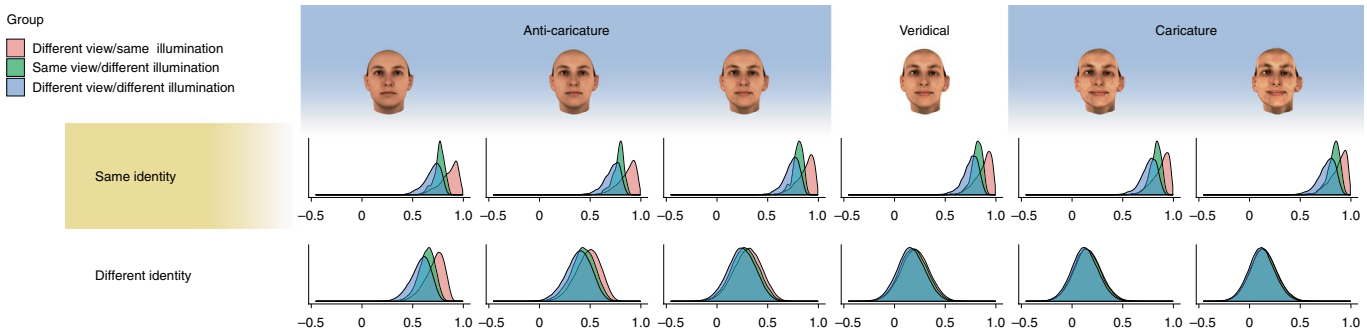


**Fig. 5 | Density curves of face image-pair cosine similarity scores.** Overlap between same-identity (top row) and different-identity (bottom row) distributions decreases as identity strength increases. Within same-identity distributions, viewpoint and illumination differences are visible at all caricature levels as peaks in the distributions. These peaks are visible in the different-identity distributions only for weak identity strengths.

accord we see here may be largely data-driven. The basic elements of face identity, captured in a non-retinotopic (categorical) domain, would probably group together without supervised learning. The key unsolved question concerns how it is possible to convert an image-based representation to a categorical identity representation without supervised learning.

Another pertinent question about the accord with neural systems involves the interpretability of the information captured by individual output units. Due to the nonlinear nature of the network, features deeper in the network are difficult to interpret semantically. Instead, semantic features are most likely found in directions through the multidimensional space[50]. For example, individual units in a face DCNN do not show consistent tuning properties for viewpoint[33,34]. However, the network retains information about both viewpoint and identity. Consequently, unit interpretability remains an active focus of research.

From a computational perspective, converting a representation in the image domain to one that operates in a categorical domain does not necessarily entail information loss. Instead, it can be achieved by reorganizing the space. Although much of this organization is sufficiently salient to be visualized in two dimensions, the full representation in the high-dimensional space drives these effects (and our computations). If the goal of a visual system is to reorganize the representational codes to 'untangle' information that is nonlinear in the image domain[35], then the data configurations we arrive at here may offer a first look at how cascades of neural-like computations can represent face identity robustly with limited loss of image context.

## Methods

**Networks.** To test the stability of the face space across network architectures and training data, we performed these simulations on two face identification DCNNs: network A[51,52] (main text) and network B[53]. Network A is a ResNet-based DCNN trained with the Universe dataset[51,52], which is a mixture of three datasets (UMDFaces[54], UMDVideos[51] and MS1M[55]). It includes images and video frames acquired in extremely challenging, in-the-wild conditions (pose, illumination and so on). We used the ResNet-101[56] architecture with the Crystal Loss (L2 Softmax) loss function for training[52]. ResNet-101 consists of 101 layers organized with skip connections that retain error signal strength to leverage very deep CNN architectures. The scale factor $\alpha$ was set to 50. The final layer of the fully trained network was removed and the penultimate layer (512 features) was used as the identity descriptor. Once the training was complete, this penultimate layer was considered the 'top layer'. Network B has 15 convolution and pooling layers, a dropout layer and a fully connected top layer that outputs a 320-dimensional identity descriptor. Network B was trained using a softmax loss function on the CASIA-WebFace dataset (494,414 images of 10,575 identities that vary widely in illumination, viewpoint and overall quality (blur, facial occlusion and so on)).

**Morphing stimuli.** Stimuli were made from 3D laser scans with densely sampled shape and reflectance data from faces. These scans were put into point-by-point correspondence with an average face[14]. In this format, a face is described as a deformation field from the average face, in both shape $[\delta x, \delta y, \delta z]$ and reflectance $[\delta r, \delta g, \delta b]$. Identity strength was manipulated by multiplying the face representation by a scalar value, $s$, such that $s > 1$ produces a caricature and $0 < s < 1$ produces an anti-caricature.

**Visualization.** Face space visualizations were performed with $t$-SNE, a nonlinear dimensionality reduction technique that uses gradient descent to preserve the distance between each point in a high-dimensional space, while reducing the number of dimensions[57]. $t$-SNE was used to reduce the DCNN's 512-dimensional feature space to a 2D space. DCNNs use the angular distance between representations to compare images. To preserve this relationship in the space, face representation vectors were normalized to unit length before computing the $t$-SNE. We used the Barnes–Hut implementation of $t$-SNE[41] with $\theta$ of 0.5, and perplexity coefficients of 30 (Fig. 2) and 100 (Fig. 3). We chose perplexity values following ref. [58]. Specifically, perplexity can be interpreted as the number of 'neighbours' of each point[57], and therefore this value should increase as the number of points per cluster increases[58]. This increase in points per cluster applies between Figs. 2 (10 images per identity) and 3 (60 images per identity). Again following ref. [58], we tested a range of values for each graph. We saw no change in the qualitative structure of groups based on perplexity value. Therefore, we selected the perplexity that yielded the most accessible visualization of the global structure.

Note that all quantitative analyses were conducted in the full 512-dimensional space.

**Classification.** Linear discriminant analysis (LDA) was applied to the full-dimensional face descriptors to classify gender and illumination. Linear regression with the Moore–Penrose pseudo-inverse was used to predict viewpoint. All predictions were conducted with identity-level cross-validation, as follows. Training data for each classifier in the cross-validation sequence consisted of the top-layer feature vectors for all images of all-but-one identity. The test data consisted of all images of the left-out identity. Output for gender and illumination was the predicted category. Output for viewpoint was the predicted viewpoint in degrees. This procedure was implemented 140 times, leaving out a different identity in each iteration. Performance for gender and illumination was measured as percent correct categorization. Performance for viewpoint was measured as the average error of the prediction in degrees.

Classifications generated from network B produced results similar to those generated from network A (see Supplementary Information). The statistical significances of both networks' predictions were evaluated with permutation tests. A null distribution was generated from the original data matrix (columns, deep features; rows, images). We permuted the column contents to break the relationship between deep features, thus creating a null distribution that preserves the statistical structure of the real data. Permutations ($n = 1,000$) were generated for each variable (gender, illumination, viewpoint). The resulting distributions were compared to the true value from each classification test. All permutation tests proved significant at $P < 0.001$, with no overlap between test value and null distribution. Network B produced the same results.

## Data availability

All data used for analysis are available via the Open Science Framework at https://osf.io/ebvys/.

## Code availability

All of the code used for plotting and analysis is available via the Open Science Framework at https://osf.io/ebvys/.

Received: 25 January 2019; Accepted: 8 October 2019;
Published online: 12 November 2019

## References


1. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, 1982).
2. Brunelli, R. & Poggio, T. Face recognition: features versus templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**, 1042–1052 (1993).
3. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
4. Bülthoff, H. H. & Edelman, S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl Acad. Sci. USA* **89**, 60–64 (1992).
5. Yuille, A. L. Deformable templates for face recognition. *J. Cogn. Neurosci.* **3**, 59–70 (1991).
6. Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* **94**, 115–147 (1987).
7. Poggio, T. & Edelman, S. A network that learns to recognize three-dimensional objects. *Nature* **343**, 263–266 (1990).
8. Turk, M. & Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**, 71–86 (1991).
9. Valentine, T. A unified account of the effects of distinctiveness, inversion and race in face recognition. *Q. J. Exp. Psychol. A* **43**, 161–204 (1991).
10. Troje, N. F. & Bülthoff, H. H. Face recognition under varying poses: the role of texture and shape. *Vision Res.* **36**, 1761–1772 (1996).
11. O'Toole, A. J., Abdi, H., Deffenbacher, K. A. & Valentin, D. Low-dimensional representation of faces in higher dimensions of the face space. *J. Opt. Soc. Am. A* **10**, 405–411 (1993).
12. O'Toole, A. J., Deffenbacher, K. A., Valentin, D. & Abdi, H. Structural aspects of face recognition and the other-race effect. *Mem. Cognit.* **22**, 208–224 (1994).
13. Nestor, A., Plaut, D. C. & Behrmann, M. Feature-based face representations and image reconstruction from behavioral and neural data. *Proc. Natl Acad. Sci. USA* **113**, 416–421 (2016).
14. Blanz, V. & Vetter, T. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* 187–194 (ACM Press/Addison-Wesley, 1999).
15. Benson, P. J. & Perrett, D. I. Perception and recognition of photographic quality facial caricatures: implications for the recognition of natural images. *Eur. J. Cogn. Psychol.* **3**, 105–135 (1991).
16. Benson, P. J. & Perrett, D. I. Visual processing of facial distinctiveness. *Perception* **23**, 75–93 (1994).
17. Byatt, G. & Rhodes, G. Recognition of own-race and other-race caricatures: implications for models of face recognition. *Vision Res.* **38**, 2455–2468 (1998).
18. Lee, K., Byatt, G. & Rhodes, G. Caricature effects, distinctiveness and identification: testing the face-space framework. *Psychol. Sci.* **11**, 379–385 (2000).
19. Rhodes, G., Byatt, G., Tremewan, T. & Kennedy, A. Facial distinctiveness and the power of caricatures. *Perception* **26**, 207–223 (1997).
20. Rhodes, G., Brennan, S. & Carey, S. Identification and ratings of caricatures: implications for mental representations of faces. *Cogn. Psychol.* **19**, 473–497 (1987).
21. Lowe, D.G. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision* Vol. 2, 1150–1157 (IEEE, 1999).
22. Dalal, N. & Triggs, B. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on* Computer Vision and Pattern Recognition, 2005 Vol. 1, 886–893 (IEEE, 2005).
23. Ojala, T., Pietikainen, M. & Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 971–987 (2002).
24. Riesenhuber, M. & Poggio, T. Models of object recognition. *Nat. Neurosci.* **3**, 1199–1204 (2000).
25. Moghaddam, B., Jebara, T. & Pentland, A. Bayesian face recognition. *Pattern Recognition* **33**, 1771–1782 (2000).
26. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. Deepface: closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1701–1708 (IEEE, 2014).
27. Sankaranarayanan, S., Alavi, A., Castillo, C. & Chellappa, R. Triplet probabilistic embedding for face verification and clustering. In *Proceedings of the IEEE International Conference on Biometrics Theory, Applications and Systems* 1–8 (IEEE, 2016).
28. Schroff, F., Kalenichenko, D. & Philbin, J. Facenet: a unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 815–823 (IEEE, 2015).
29. Chen, J.-C. et al. An end-to-end system for unconstrained face verification with deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* 118–126 (IEEE, 2015).
30. Ranjan, R., Sankaranarayanan, S., Castillo, C. D. & Chellappa, R. An all-in-one convolutional neural network for face analysis. In *12th IEEE International Conference on* Automatic Face and Gesture Recognition (*FG 2017*) 17–24 (IEEE, 2017).
31. Fukushima, K. Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Netw.* **1**, 119–130 (1988).
32. KrizhevskyA., SutskeverI. & HintonG. E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Proc. Syst.* **25**, 1097–1105 (2012).
33. Parde, C. J. et al. Face and image representation in deep CNN features. In *12th IEEE International Conference on Automatic Face and Gesture Recognition* (*FG 2017*) 673–680 (IEEE, 2017).
34. O'TooleA. J., CastilloC. D., PardeC. J., HillM. Q. & ChellappaR. Face space representations in deep convolutional neural networks. *Trends Cogn. Sci.* **22**, 794–809 (2018).
35. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
36. Hong, H., Yamins, D. L., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
37. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
38. Brennan, S. E. Caricature generator: the dynamic exaggeration of faces by computer. *Leonardo* **18**, 170–178 (1985).
39. Rhodes, G. *Superportraits: Caricatures and Recognition* (Psychology Press, 1997).
40. Leopold, D. A., O'Toole, A. J., Vetter, T. & Blanz, V. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat. Neurosci.* **4**, 89–94 (2001).
41. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
42. Grill-Spector, K. & Weiner, K. S. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* **15**, 536–548 (2014).
43. Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. Basic objects in natural categories. *Cogn. Psychol.* **8**, 382–439 (1976).
44. EberhardtS., CaderJ. G. & SerreT. How deep is the feature analysis underlying rapid visual categorization? *Adv. Neural Inf. Proc. Syst.* **29**, 1100–1108 (2016).
45. Kietzmann, T. C. et al. The occipital face area is causally involved in facial viewpoint perception. *J. Neurosci.* **35**, 16398–16403 (2015).
46. Natu, V. S. et al. Dissociable neural patterns of facial identity across changes in viewpoint. *J. Cogn. Neurosci.* **22**, 1570–1582 (2010).
47. Grill-Spector, K. et al. Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* **24**, 187–203 (1999).
48. Yue, X., Cassidy, B. S., Devaney, K. J., Holt, D. J. & Tootell, R. B. Lower-level stimulus features strongly influence responses in the fusiform face area. *Cerebral Cortex* **21**, 35–47 (2010).
49. Kay, K. N., Weiner, K. S. & Grill-Spector, K. Attention reduces spatial uncertainty in human ventral temporal cortex. *Curr. Biol.* **25**, 595–600 (2015).
50. Szegedy, C. et al. Intriguing properties of neural networks. Preprint at https://arxiv.org/abs/1312.6199 (2013).
51. Bansal, A., Castillo, C. D., Ranjan, R. & Chellappa, R. The do's and don'ts for CNN-based face verification. In *ICCV Workshops* 2545–2554 (IEEE, 2017).
52. Ranjan, R. et al. A Fast and Accurate System for Face Detection, Identification, and Verification. In *Proceedings of the IEEE Transactions on Biometrics, Behavior, and Identity Science* 82–96 (IEEE, 2019)
53. Chen, J.-C., Patel, V. M. & Chellappa, R. Unconstrained face verification using deep CNN features. In *IEEE Winter Conference on* Applications of Computer Vision (*WACV*) 1–9 (IEEE, 2016).
54. Bansal, A., Nanduri, A., Castillo, C. D., Ranjan, R. & Chellappa, R. UMDFaces: an annotated face dataset for training deep networks. In *IEEE International Joint Conference on* Biometrics (*IJCB*) 464–473 (IEEE, 2017).
55. Guo, Y., Zhang, L., Hu, Y., He, X. & Gao, J. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision* 87–102 (Springer, 2016).
56. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
57. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
58. Wattenberg, M., Viégas, F. & Johnson, I. How to use t-SNE effectively. *Distill* **1**, e2 (2016).


## Acknowledgements

This work had funding support from the Intelligence Advanced Research Projects Activity (IARPA). This research is based on work supported by the Office of the Director of National Intelligence (ODNI) and IARPA (via R&D contract no. 2014-14071600012). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA or the US Government.

## Author contributions

All authors were involved in the conceptualization and design of the methodology of the study. M.Q.H., C.D.C., R.R. and J.-C.C. handled software. The original draft of the

manuscript was prepared by M.Q.H. and A.J.O. Review and editing were carried out by M.Q.H., C.J.P., Y.I.C., C.D.C., V.B. and A.J.O. Formal analysis, investigation and visualization were done by M.Q.H. and C.J.P., with validation by M.Q.H., Y.I.C., C.J.P. and C.D.C. Supervision and funding acquisition were handled by C.D.C. and A.J.O., with project administration by A.J.O.

## Competing interests

University of Maryland has filed a US patent application that covers portions of network A. R.R. and C.D.C. are co-inventors on this patent.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s42256-019-0111-7.

**Correspondence and requests for materials** should be addressed to M.Q.H.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.