Human Visual Pathways for Action Recognition versus Deep Convolutional Neural Networks: Representation Correspondence in Late but Not Early Layers

Yujia Peng[1,2,3,4*], Xizi Gong[1*], Hongjing Lu[4,5], and Fang Fang[1,6,7,8]

Abstract

■ Deep convolutional neural networks (DCNNs) have attained human-level performance for object categorization and exhibited representation alignment between network layers and brain regions. Does such representation alignment naturally extend to other visual tasks beyond recognizing objects in static images? In this study, we expanded the exploration to the recognition of human actions from videos and assessed the representation capabilities and alignment of two-stream DCNNs in comparison with brain regions situated along ventral and dorsal pathways. Using decoding analysis and representational similarity analysis, we show that DCNN models do not show hierarchical representation alignment to human brain across visual regions when processing action videos. Instead, later layers of DCNN models demonstrate greater representation similarities to the human visual cortex. These findings were revealed for two display formats: photorealistic avatars with full-body information and simplified stimuli in the point-light display. The discrepancies in representation alignment suggest fundamental differences in how DCNNs and the human brain represent dynamic visual information related to actions. ■

INTRODUCTION

Humans possess an exquisite ability to recognize actions, even from stimuli that greatly deviate from everyday experiences, such as point-light displays consisting of only a few discrete dots representing joint movements (Johansson, 1973). This exceptional perceptual ability likely reflects the crucial role of actions in human learning. Actions, exemplified by body movements, stand as a prime example of biological motion, providing a form of "body language" for perception and cognition. When we observe the body movements of an individual, we not only perceive actions with well-controlled arm movements, but also gain a clear sense of whether this person is performing locomotory actions (e.g., a stretching exercise), or is interacting with an object (e.g., shooting a basketball, or making a golf swing) or with the other people (e.g., waving hands in greeting, or showing directions to someone). Hence, recognition of biological motion goes beyond just the assignment of action labels, also encompassing action semantic classification (Dittrich, 1993) and attribute identification of other people (Peng, Thurman, & Lu, 2017; van Boxtel & Lu, 2012; Pollick, Lestou, Ryu, & Cho, 2002).

Over several decades, psychophysical and neuroimaging studies have produced converging evidence that recognition of biological motion is supported by dual processes, with separate analysis pathways specialized for kinematics of body movements and the spatial structure of body forms (van Boxtel & Lu, 2015; Theusner, de Lussanet, & Lappe, 2011; Lange, Georg, & Lappe, 2006; Lu & Liu, 2006; Beintema & Lappe, 2002; Pinto & Shiffrar, 1999; Cutting, Moore, & Morrison, 1988). In particular, fMRI experiments have shown that point-light videos activate not only motion-selective regions such as middle temporal area (MT)/middle superior temporal area (MST), but also regions responsible for processing appearance information, which is located in the projection from primary visual cortex (V1) to inferotemporal cortex (Grossman & Blake, 2002). For example, the extrastriate body area (EBA), which is sensitive to human body form information, has been reported to be activated in recognizing biological motions (Lingnau & Downing, 2015; Downing, Jiang, Shuman, & Kanwisher, 2001). In conjunction with this, numerous studies have established that the posterior superior temporal sulcus (pSTS) plays a crucial role in biological motion perception, highlighting its function in integrating motion

[1]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, People's Republic of China, [2]Institute for Artificial Intelligence, Peking University, Beijing, People's Republic of China, [3]National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence, Beijing, China, [4]Department of Psychology, University of California, Los Angeles, [5]Department of Statistics, University of California, Los Angeles, [6]IDG/McGovern Institute for Brain Research, Peking University, Beijing, People's Republic of China, [7]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, People's Republic of China, [8]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, People's Republic of China
*Equal contribution.

Journal of Cognitive Neuroscience 36:11, pp. 2458–2480
https://doi.org/10.1162/jocn_a_02233

processing and appearance processing (Thurman, van Boxtel, Monti, Chiang, & Lu, 2016; Grossman, Jardine, & Pyles, 2010; Grossman, Battelli, & Pascual-Leone, 2005; Grossman & Blake, 2001, 2002; Vaina, Solomon, Chowdhury, Sinha, & Belliveau, 2001) or identifying social-related information in biological motion (McMahon, Bonner, & Isik, 2023).

Inspired by the findings from behavioral and neuroscience studies, Giese and Poggio (2003) developed a computational model for action recognition based on two parallel information processing streams: a "what" pathway and a "where" pathway. The "what" pathway is specialized for analyzing body forms in static image frames, whereas the "where" pathway is specialized for processing motion information. Each pathway comprises a hierarchy of feature detectors, with receptive fields increasing in size and complexity, to encode either configural forms or motion patterns. This model can account for a range of behavioral and neural effects, illustrating the effectiveness of adopting computational architectures mirroring the two parallel processes in the brain. However, the hierarchical features in the model were predefined, emulating neural receptive fields measured in neuroscience studies. Can these features used in action recognition be spontaneously learned from visual experiences of viewing human actions?

The recent advances in deep learning models provide a plausible way to learn visual representations from a massive amount of training data, given the architecture of networks and the objective function of learning (Krizhevsky, Sutskever, & Hinton, 2012; Lecun, Bottou, Bengio, & Haffner, 1998). Simonyan and Zisserman (2014) developed a two-stream deep convolutional neural network (DCNN) for the recognition of actions in videos. The two-stream DCNN consists of two parallel pathways: a spatial pathway that processes appearance information, taking pixel-level intensity of images as the input, and a temporal pathway that processes motion information using optical flow as the input. The two-stream DCNN performed well on the action classification task and reached human-level recognition performance for realistic action videos in two challenging data sets: UCF-101 (Soomro, Zamir, & Shah, 2012) and HMDB-51 (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011). Although the two-stream DCNN model achieved human-level recognition performance at the behavioral level, it is unclear whether this DCNN fully captures the neural representation of actions in the human brain.

Recent studies comparing DCNNs and the brain for object recognition in static images addressed the above question, as neuroimaging studies have reported brain–DCNN correspondences, suggesting that deep neural networks can learn similar representations as the human brain to show representation alignment between biological and artificial systems. The representation of DCNN layers exhibits characteristic properties of neural representations, and can predict image-driven neural responses along the ventral visual stream (Cadena et al., 2019; Yamins et al., 2014). Specifically, the DCNNs generate visual representations that capture increasingly abstract attributes of object categories (Seeliger et al., 2018; Cichy, Khosla, Pantazis, & Oliva, 2017; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Khaligh-Razavi, Henriksson, Kay, & Kriegeskorte, 2017; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Hong, Yamins, Majaj, & DiCarlo, 2016; Güçlü & van Gerven, 2015; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins, Hong, Cadieu, & DiCarlo, 2013; but see Xu & Vaziri-Pashkam, 2021). Furthermore, the evidence above also showed representation alignment between DCNN layers to brains, where the representational structure of lower and higher human visual areas aligns with early and later DCNN layers, respectively. While accumulating evidence increasingly supports the alignment of visual representations between DCNNs and the human brain for object recognition, further research is required to extend these findings to dynamic stimuli. This extension is particularly important because action recognition encompasses both the ventral and dorsal pathways, as well as the integration of visual processes from both pathways.

In the current study, we compared fMRI responses to human actions with network responses of three different DCNNs (i.e., a form-only spatial pathway DCNN, a motion-only temporal pathway DCNN, and a two-stream DCNN). To acquire robust representations of actions, we utilized action videos in two presentation formats: photorealistic avatar videos and point-light videos. These two formats share the same kinematic movements of human actors but differ in the level of detail regarding actor appearance in image frames. The photorealistic avatar videos, rendered with 3-D human avatars, offer detailed appearances and high ecological validity. In contrast, the point-light videos eliminated body shape and contextual information, retaining only the motion trajectories of major joints involved in actions. We selected five ROIs in the brain situated along the two visual pathways: V1 for low-level visual information processing, middle temporal/medial superior temporal (MT+) for motion processing, lateral occipital complex (LOC) for object perception, EBA for human body processing, and pSTS, which has been linked to biological motion perception and theory of mind. Using decoding analysis and representational similarity analysis (RSA), we examined how well DCNNs capture the representational structures in the human brain for action recognition.

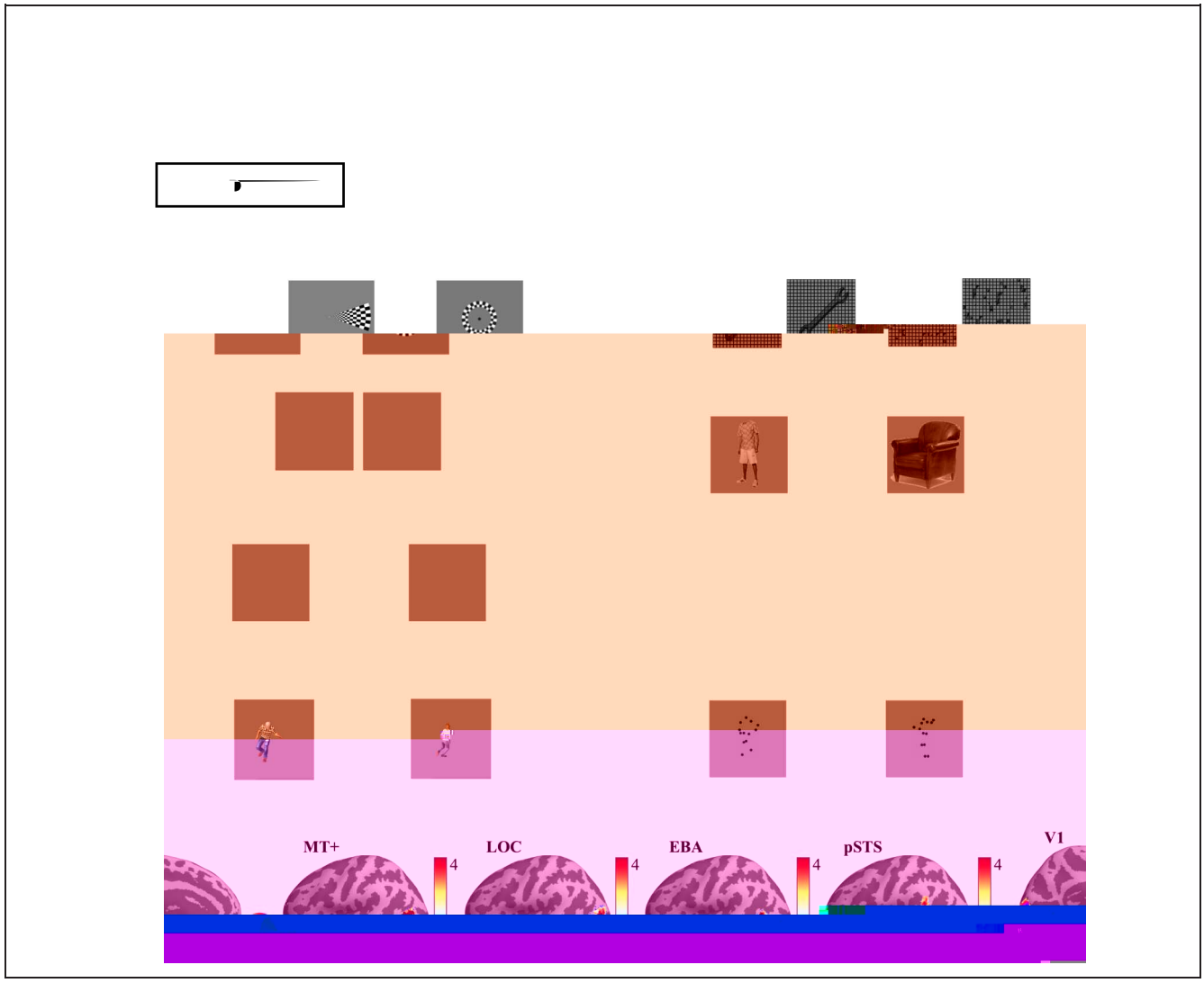
METHODS

Model Structure and Training for Action Recognition

The two-stream DCNN model (Simonyan & Zisserman, 2014) was adopted because its architecture resembles the dual pathways of brain in processing biological motion, and the model achieves human-level performance of action recognition. As shown in Figure 1, this two-stream DCNN takes two types of information as inputs to classify a video into action categories. One source of

information is the pixel-level appearance of moving bodies in a sequence of static images, which serves as inputs to the spatial pathway. The other source of information is motion information represented by optical flow fields (Horn & Schunck, 1981), which serves as inputs to the temporal pathway. The two-stream DCNN model then integrates outputs from both pathways and culminates into a final decision via fully connected (FC) layers. Note that we also examined two control DCNN models, in which these two types of information were independently processed by single-stream DCNNs, namely, a form-only spatial pathway DCNN (i.e., spatial DCNN) and a motion-only temporal pathway DCNN (i.e., temporal DCNN), as shown in the top two images in Figure 1.

The DCNN models were trained to perform an action classification task of 15 action categories using naturalistic videos in the Human 3.6 M data set (Ionescu, Papava, Olaru, & Sminchisescu, 2014). These 15 categories are



of fixation, followed by 36 trials of action presentations. Each action was displayed for 3 sec, interleaved with a period for response and jitter of 3, 5, or 7 sec. Each run ended with another 10 sec of fixation. Participants were asked to indicate the semantic class of the action by pressing one of the three buttons on the response box.

Data Analysis

RSA of ROI and DCNN Action Representations

Representational similarity analysis (RSA) (Kriegeskorte et al., 2008) was used to compare neural representations with DCNN representations. Using RSA, we calculated the similarity between response vectors of pairs of action videos, yielding a representational dissimilarity matrix (RDM) with the size of 36 × 36 for each DCNN layer and each brain ROI. Therefore, layer-specific DCNN RDMs can be compared with fMRI ROI RDMs, yielding a measure of brain-DCNN representational similarity.

Specifically, for each layer in the DCNNs, we extracted the DCNN layer output for each video clip (i.e., every 10 frames in each action instance). For convolutional layers, we used a max-pooling approach to take the maximum response value from each 2-D response field (e.g., for a convolutional layer of 7 × 7 × 512, we took the maximum value from each of 7 × 7 matrices, yielding a feature vector with the size of 512 dimensions). Max-pooling was conducted because of two reasons. First, this max-pooling method extracts location-invariant representations and reduces the dimensions of feature vectors. Hence, this method resembles the process of feature selection in visual areas. Second, the max-pooling operation reduces noise and increases the robustness of action representations in convolutional layers. In addition, we examined both the average and median values in the pooling

operation and found similar structures in RDM results. These extracted features for video clips in the same action instance were concatenated into a vector. The sizes of feature vectors were reported in the Appendix (Section 2, Table A1). Then, for one pair of actions, we used 1 minus Spearman correlation coefficients to represent dissimilarity between the model activation vectors, yielding a 36 × 36 RDM (i.e., DCNN RDM), summarizing the representational dissimilarities for each layer of a network. To increase the temporal flexibility of the model representation, we also computed dissimilarity scores by shifting the temporal window between feature vectors of two video clips and used the minimum dissimilarity score in RDM. The results of temporal shifting showed a similar trend to those obtained without temporal shifting.

For neural activities, the same correlation-based dissimilarity calculation as for model measurements was used to compute RDMs for brain ROIs. For each ROI (V1, MT+, LOC, EBA, pSTS), the z-normalized beta activation patterns of voxels were concatenated into vectors (see Appendix Sections 2 and 3 for details). We then calculated the correlation-based dissimilarity (1 minus Spearman's R) between beta patterns for every pair of action instances within the ROI, leading to a 36 × 36 ROI RDM indexed in rows and columns by the compared actions.

With DCNN RDMs and ROI RDMs, we compared layer-specific model representations to region-specific brain representations by calculating Spearman's correlation between the dissimilarity scores in the DCNN and ROI RDMs for each individual participant. To establish the upper bounds of model-brain comparisons, we first assessed the reliability of the ROI RDMs across the group of human participants by calculating the noise ceiling of the fMRI data. The noise ceiling defines the upper boundary of a model's capacity to account for dissimilarity variance in brain representations, constrained by the inherent and measurement noise in the brain activity. In other words, it quantifies the maximum variance in the brain representations that a model could explain. The noise ceiling was defined as $\frac{1}{n}\sum_{i=1}^{n} r(v_i, \bar{v})$, where n denotes the number of participants, v_i denotes each participant's RDM, $\bar{v}$ denotes the averaged RDM across participants, and r denotes Spearman's correlation coefficient (Khaligh-Razavi, Cichy, Pantazis, & Oliva, 2018; Nili et al., 2014). We then computed the proportion of brain variance explained by DCNN layers: dividing the brain–DCNN Spearman's R correlation coefficients by the corresponding noise ceiling of ROI representations.

Incorporating both semantic classes and action categories, we generated a full-knowledge design matrix of RDMs, as shown in Figure 2C. This design matrix assumes higher similarities between videos within one semantic class, in addition to greater similarities of videos within one action category (e.g., jumping, walking, and running are all locomotory actions). Hence, this design matrix captures the hierarchical structures of similarity, for example, actions in different action categories are similar to a certain degree if they are within the same semantic class. We also calculated an action-only design matrix as a comparison in the Appendix (Sections 9 and 10), which assumes only higher similarity between videos within one action category (e.g., jumping) but ignores similarities among actions within one semantic class. DCNN and brain ROI RDMs were compared with design matrices to investigate whether DCNN layers and ROIs demonstrate semantic-level representations beyond visual similarities of actions within the same action category.

Multivariate Pattern Analysis of ROI and DCNN Representations

To decode action categories, we applied multivariate pattern analysis (MVPA) based on linear support vector machines. Neural representations of ROIs to each action instance were concatenated and normalized before entering as inputs to the support vector machine training algorithm. To control for feature dimensions across ROIs in the decoding process, PCA was conducted on data of the selected voxels of each ROI to reach a limit of 20 feature dimensions. A nine-way decoder was then trained to classify the brain activity patterns into one of the nine action categories using a fourfold cross-validation procedure. Peak decoding accuracy across layers was examined through bootstrapping analyses (see Appendix Section 6 for details). The decoding analyses complement the design matrix correlation analyses by emphasizing action classifications, whereas the RSAs with the design matrix emphasize the hierarchical representations of semantic classes and action categories.

Statistical Analysis

Statistical analysis was performed using IBM Statistical Package for the Social Sciences (SPSS V20.0) and MATLAB. Repeated-measures ANOVAs were applied to examine the differences between DCNN pathways and between action presentation types. Permutation tests were conducted to assess the statistical significance of RDM correlations. Specifically, we permuted the correlations in the original RDM and calculated Spearman's correlation between the original and the shuffled RDMs. This process was repeated 10,000 times to generate a null distribution. The p values of the observed correlation were calculated as the percentile of the observed data in the permuted null distribution. In addition, the 95th percentile of each null distribution was provided as dashed lines in the corresponding figures.

RESULTS

Action Representations in DCNN Layers

To investigate how semantic classes and action categories were represented in DCNN layers, we conducted the representation similarity analysis using the activations of each

DCNN layer. In particular, deeper DCNN layers demonstrated increasingly clear diagonal block patterns, indicating enhanced discrimination of action categories for both photorealistic avatar and point-light videos (Figure A1). Interestingly, the diagonal block patterns in RDMs of the spatial pathway, which processed appearance information, were less clear for photorealistic avatar videos than those for point-light videos. This result may be due to more varied visual cues (i.e., actor appearances) among the photorealistic avatar videos. In contrast, for both types of videos, the RDMs of the temporal pathway, which processes optical-flow information, all exhibited clear diagonal block patterns.

To quantitatively assess how well action categories were represented in DCNN layers, we trained classifiers to classify activations in DCNN layers to one of the nine action categories. As shown in Figure 4A and 4B, the later layers (Conv5, FC1, and FC2) in all three DCNNs showed significantly greater than chance-level decoding performance (ps $< .001$), suggesting that the DCNN activations in the later layers capture features crucial for recognizing action categories. Notably, even early layers of DCNN (e.g., for photorealistic avatar videos, starting from Conv3 for the spatial DCNN, and Conv1 for the temporal DCNN; for point-light videos, starting from Conv1 layers for both the spatial and temporal pathways) demonstrated significant action classification, although yielding worse performance than later layers. To compare the decoding accuracies in the spatial and temporal pathways, a repeated-measures ANOVA was performed with DCNN Models (spatial and temporal) as a between-subject factor and DCNN Layers as a within-subject factor. Results showed a significant main effect of DCNN Layers, photorealistic avatar: $F(5, 30) = 19.57$, $p = .049$, $\eta_p^2 = .980$; point-light: $F(5, 30) = 40.69$, $p < .001$, $\eta_p^2 = .871$, indicating that as layers go deeper, the decoding accuracies for action categories increase. Moreover, the main effect of DCNN Models was significant, photorealistic avatar: $F(1, 6) = 25.01$, $p = .002$, $\eta_p^2 = .806$; point-light: $F(1, 6) = 43.56$, $p = .001$, $\eta_p^2 = .879$, indicating that the temporal pathway DCNN yielded significantly greater decoding accuracy than those of the spatial pathway DCNN. No significant interaction was found. We also examined when the DCNN layers reached the peak decoding performance but did not find a significant difference between the spatial and temporal pathways (Appendix Section 6).

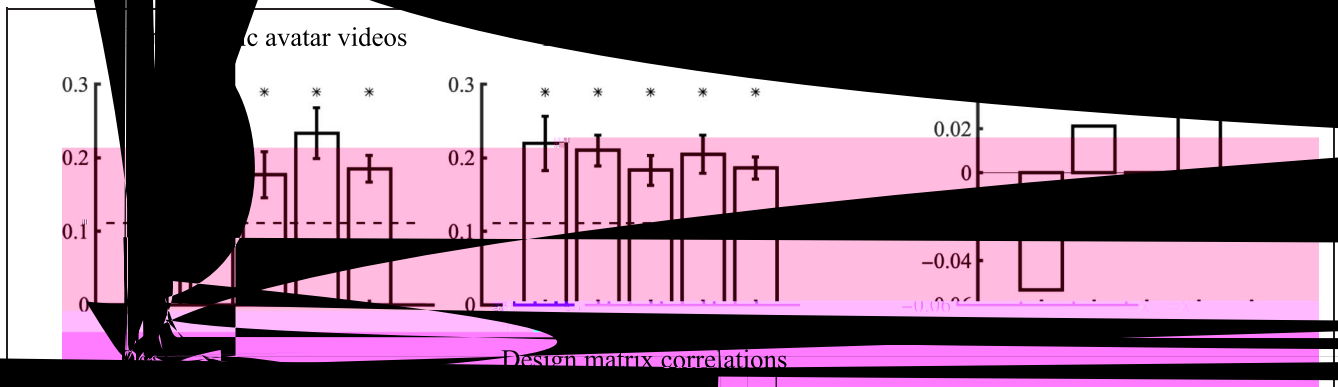
To examine the impact of action presentation types on action category decoding, we conducted a repeated-measures ANOVA with Action Presentation Types (photorealistic avatar vs. point-light videos) as a between-subject factor and DCNN Pathway and Layer as two within-subject factors. Results showed a significant two-way interaction between Action Types and DCNN Pathways, $F(1, 6) = 141.25$, $p < .001$, $\eta_p^2 = .959$, indicating that the decoding accuracy was influenced more strongly for the spatial pathway than for the temporal pathway when the presentation of actions changed from photorealistic avatar videos to point-light videos. The main effect of Actions Types, DCNN Pathways, and Layers were all significant (ps $< .001$). No other two-way or three-way interaction was found. Differences between decoding accuracies for photorealistic avatar and point-light videos were shown in Figure 4C.

Correlations between DCNN RDMs and the full-knowledge design matrix were calculated for each DCNN layer (Figure 4D and 4E). In the spatial pathway DCNN, RDMs of deeper layers increasingly resemble the full-knowledge design matrix, revealed by a significant linear relationship between layer depth and the resemblance to the full-knowledge design matrix (photorealistic avatar videos [$b = 0.051, p = .001$], point-light videos [$b = 0.068$, $p = .002$]). In the temporal pathway DCNN, we observed a significant linear relationship for photorealistic avatar videos ($b = 0.049, p = .006$), and a trending relationship (but not significant) for point-light displays ($b = .048, p = .083$). These findings were consistent with the findings of decoding performance.

Differences in DCNN-design correlations between photorealistic avatar and point-light videos were shown in Figure 4F. To examine the impact of action presentation types on DCNN-RDM correlations, we conducted ANCOVA analyses using the Design Matrix as the dependent variable, Action Presentation Type as the independent variable, and DCNN RDM as the covariates. Significant interactions were found across all spatial pathway layers, as well as Conv5 layers of both the temporal and two-stream pathway (ps $< .05$, false discovery rate [FDR]-corrected), suggesting significant correlation differences between the two action presentation types.

To quantitatively examine whether the representations in DCNN layers capture the action category and semantic class information, we conducted a representation-boundary effect analysis similar to that used by Kriegeskorte and colleagues (2008). For each DCNN layer, we first calculated mean dissimilarities for within-action-category video pairs, within-semantic-class pairs, and between-semantic-class pairs (Figure 4G). We then defined two representation-boundary effects: the action-category boundary effect and the semantic-class boundary effect. The action-category boundary effect referred to the difference between the mean dissimilarities for within-action-category pairs and within-semantic-class pairs, and the semantic-class boundary effect referred to the difference between the mean dissimilarities for within-semantic-class pairs and between-semantic-class pairs. To assess the statistical significance, we performed permutation analyses by shuffling the DCNN RDMs 10,000 times and recalculating both boundary effects for each permutation. As shown in Figure 4G, the two boundary effects were significant across all layers in the two-stream DCNN for both photorealistic avatar and point-light videos, suggesting that the two-stream DCNN can effectively capture both action category and semantic class information. Moreover, significant action category boundary effects

emerged as
DCNN for ph
tial and tem
trast, signifi
in later laye
layers for bo



Design matrix correlations

pr
ted a
g accura
listic avat
ject factor and
howed a significa
n Presentation Type

point-light videos, V1 yielded significantly lower correlations to the full-knowledge design matrix than MT+, pSTS, and EBA (paired t tests, ps $< .05$). Differences in ROI-design correlations between photorealistic avatar and point-light videos were shown in Figure 5F. To examine the impact of action presentation types on ROI-design correlations, we performed independent t tests for each ROI. Results showed that only MT+ yielded a significant difference between action presentation types ($p = .042$), where the MT+ correlation to the design matrix was significantly higher for photorealistic avatar videos ($M = 0.61$) than the point-light videos ($M = 0.43$). This result suggests that the dense optical flow captured in natural videos conveys more informative cues for decoding action categories than sparse movements of only joints included in point-light displays.

Moreover, we constructed two RDMs based on the average speed (i.e., how fast the dots move) and the average scatterness (i.e., how scattered the dots are from the center of the body) in the point-light videos to examine the representation of low-level visual features in different ROIs. We calculated the correlations between ROI RDMs and the constructed RDMs of speed and scatterness. As shown in Figure A5, V1 yielded stronger correlations to the speed and scatterness in actions, whereas pSTS showed the lowest correlations to these low-level visual features.

To quantitatively examine whether the representations in these ROIs capture both action category and semantic class information, we conducted a representation-boundary effect analysis (Kriegeskorte et al., 2008) by calculating both the action-category and the semantic-class boundary effects. As shown in Figure 5G and 5H, for the action-category boundary effect, all ROIs exhibited significantly greater within-semantic-class dissimilarity than within-action-category dissimilarity (paired-samples t tests, ps $< .05$, FDR-corrected). For the semantic-class boundary effect, except V1 in both experiments and pSTS in Experiment 1 of photorealistic avatar videos, all other ROIs demonstrate significant effects, with greater between-semantic-class dissimilarity than within-semantic-class dissimilarity (ps $< .05$, FDR-corrected). Significant differences between photorealistic avatar and point-light videos were observed in V1, where photorealistic avatar videos yielded much greater within-action dissimilarity (Figure 5I).

Action Representation Alignment between Brain ROIs and DCNN

To investigate representation alignment between DCNNs and brain regions, we computed the correlations of RDMs between DCNNs and brain ROIs (Figure 6). Specifically, RDMs of human fMRI responses to photorealistic avatar and point-light action videos were compared with RDMs of network responses of the three different DCNNs (i.e., a spatial pathway DCNN, a temporal pathway DCNN, and a two-stream DCNN) across model layers. The proportion of brain variance explained by DCNN layers was computed by dividing the brain–DCNN Spearman's R correlation coefficients by the corresponding noise ceiling. A greater proportion of brain variance explained would indicate greater representation similarities between DCNNs and the brain.

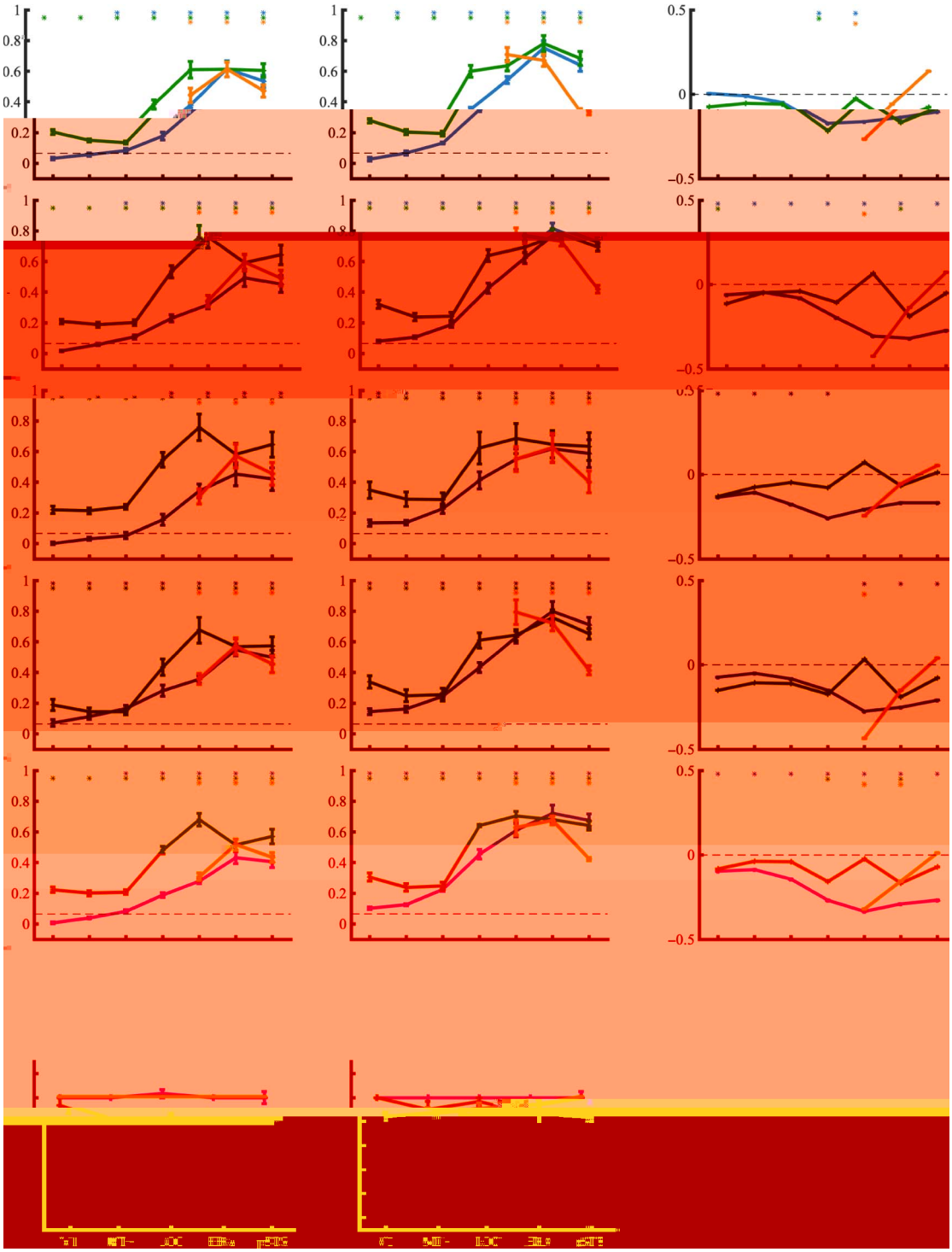
To assess the degree of alignment between brain and DCNN representations for actions, we identified, for each DCNN, the layer number that explained the largest proportion of variance for each of the five ROIs in each participant. If brain–DCNN representation alignment exists (i.e., a correspondence in action representation between earlier and later DCNN layers to lower and higher visual processing regions, respectively), we would expect a greater proportion of V1 variance explained by early DCNN layers (e.g., Conv1 and 2 in both the spatial and temporal DCNNs), MT+ by early/middle layers in the temporal pathway, EBA and LOC by later layers in the spatial pathway (e.g., spatial Conv5, FC1, and FC2), and pSTS by later layers in the temporal pathway or layers in the two-stream DCNN (McMahon et al., 2023). We then assessed whether the resulting layer numbers increased from lower to higher visual regions by examining the proportion of brain variance explained by each DCNN layer.

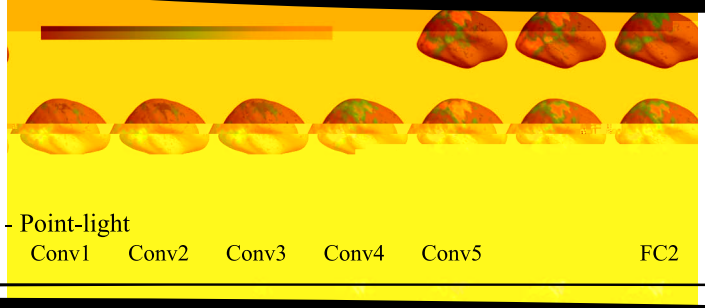
As depicted in Figure 6A and 6B, the layers showing the maximum explained variance of brain RDMs did not exhibit an increase in order from lower to higher visual regions (ps $> .05$). Thus, in contrast to the finding in object recognition that the proportion of explained variance of lower and higher ROIs reached the maximum by earlier and later DCNN layers, respectively (e.g., Cichy et al., 2016), for action recognition, the later layers of DCNNs, specifically Conv5, FC1, and FC2, consistently explained the most variance in all brain ROIs. Differences in ROI-DCNN correlations between photorealistic avatar and point-light videos were shown in Figure 6A (last column). Results showed that the differences were overall below 0, suggesting that DCNN models were more effective at explaining ROI representation variance for point-light videos than photorealistic avatar videos. Specifically, for all ROIs, several layers in the spatial DCNN yielded significant differences between experiments (ps $< .05$, FDR-corrected), whereas the differences of the temporal DCNN were predominantly insignificant, suggesting a greater impact of presentation type on the spatial pathway.

Brain–DCNN Mapping Revealed by Searchlight Analysis

To further identify brain areas with action representations similar to those of the DCNN layers and the full-knowledge design matrix, we used a spatially unbiased volume-based searchlight approach. For each participant, brain RDMs for every voxel (3-voxel radius) were constructed based on the local activity patterns4a1.6i

Point-





correlated with the layer-sp
Spearman's *R*, yielding a 3-D
each DCNN layer. The result
ticipants were normalized
maps.

First, as illustrated in Figu
extended upon the previ
showing significant correlat
corrected) between several a
ncluding the somatosensory
lcus, and superior pariet
, Conv5, FC1, and FC2 i
parietal regions also e
o the full-knowledge
-corrected). These
anterior parietal co
correlations w
nd FC2) were

represented in later layers, where representational dissimilarity matrices yielded clearer clustering patterns consistent with action categories. For human visual areas, all selected brain regions, even V1, were able to demonstrate representations that discriminate action categories. However, the comparisons between DCNN layers and ROIs across two experiments consistently revealed a lack of hierarchical representation alignment between DCNN layers and human brain ROIs, where later DCNN layers (Conv5, FC1, and FC2) yielded similarity to brain representations in both early visual areas and high-level ROIs. The finding that actions are best represented in later layers of DCNNs is consistent with the hypothesis that later layers in DCNNs increasingly capture the semantic knowledge about human actions (e.g., whether the action demonstrates social interaction). As proposed in Saxe, McClelland, and Ganguli (2019), the ability of DCNNs to capture abstract semantic knowledge consisting of useful categories may be inherent in the deep connection structure. Global optimality in DCNNs, achieved by acquiring features to enhance visual recognition, necessitates downplaying the contributions of unique local features for a small set of individual instances, and emphasizing the importance of global features shared by instances within one category, which are likely encoded in deeper layers of DCNNs.

The lack of representation alignment between two-stream DCNN and brain ROIs in action recognition was manifested by low similarities between V1 and early DCNN layer, which was contrary to findings in some studies in object recognition (Seeliger et al., 2018; Cichy et al., 2016, 2017; Eickenberg et al., 2017; Khaligh-Razavi et al., 2017; Hong et al., 2016; Güçlü & van Gerven, 2015; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2013). The discrepancy is likely because of differences between the nature of the stimuli and tasks. Previous research focused mostly on object recognition, and the current study targeted dynamic visual stimuli of human actions. Although the brain performs action recognition with little effort, the process is sophisticated and slow progress has been made in past decades largely due to the obstacle of the representation issue. Specifically, we do not have adequate knowledge regarding how to represent actions in a comprehensive and robust way that supports sophisticated inference, so that such representations can be used by high-level reasoning systems. Object recognition from static images can possibly be achieved in a single stream of feedforward processing, emulating the bottom–up visual processing in the human brain, where low-level visual features are extracted, and complex object patterns are gradually built upon combinations of features extracted by early layers. However, processing of dynamic visual information requires longer periods to accumulate information. Action recognition unfolds over time, during which communications between brain regions happen. Hence, iterations of bottom–up feature extraction and top–down regulations both play important roles in making decisions of observed actions. After a few seconds of visual processing, all ROIs may fine-tune representations beyond low-level visual features. Consistent with the speculations above, the present study found that all the selected brain regions were able to demonstrate semantic-level differentiation of action categories within seconds of stimuli presentation. However, the two-stream DCNNs operate in a purely bottom–up driven manner and lack top–down regulations (Peng et al., 2021). Hence, the lack of clear DCNN-brain alignment may result from the absence of top–down connections in DCNNs. In contrast, the human visual system operates through the seamlessly integrated interaction of feedforward processing and feedback regulations in a concerted manner.

Nevertheless, the current results are consistent with a recent evaluation (Xu & Vaziri-Pashkam, 2021) that found limited visual representational correspondence between DCNNs and the human brain for object recognition. The current results are not likely to be driven by low signal-to-noise ratio (SNR) issue, where we used an event-related fMRI design with 5 days of repetitions, aiming to establish relatively good SNR. This design enhanced the SNR by showing high brain–DCNN correlation values, with the highest correlation being 0.6 in the present study. Furthermore, the current study used functional localizers to define brain ROIs for individual participants, whereas previous studies defined human brain regions anatomically or through atlas.

There are other deep learning models developed for action recognition in videos. For example, Feichtenhofer, Fan, Malik, and He (2019) introduced SlowFast network, which also use two pathways for action recognition from videos. However, both pathways operate on a clip of video as a spatiotemporal volume with different frame rates. The architecture in SlowFast networks does not map to the "what" and "where" pathways in the brain. The other popular network is a two-stream inflated 3-D convnet developed by Carreira and Zisserman (2017). The inflated 3-D convnet model is built based on the inception-V1 network structure and includes nine inception layers. In contrast to many studies comparing human visual regions with the DCNN models (such as AlexNet), there exists little evidence on the representation alignment between visual areas and inception layers. Hence, this article focused on the two-stream DCNNs as an extension of standard DCNN models, which have rich literature on human and model comparisons. In addition, we examined AlexNet as a control model and, once again, failed to find representation alignment between AlexNet layers and visual areas in action recognition (Appendix Section 7).

We found a relatively consistent impact of action display format (e.g., photorealistic avatar vs. point-light video) on both neural activities in the brain and layer activities in DCNNs. The spatial DCNN, as well as the early visual ROI V1, were more impacted by a change from point-light to photorealistic avatar videos. This indicates that for

visual representations of more varied visual cues (i.e., actor appearances) in the photorealistic avatar videos, the temporal DCNN and visual regions higher on the hierarchy were able to rely on motion cues or more abstract cues and were less distracted by the appearances. It is worth noting that DCNNs showed similar results in processing motion flow information between two display formats, suggesting that DCNN models demonstrate a certain degree of generalization ability, as these models are trained using natural videos. Furthermore, the visual vividness of human actors may not be a vital factor in determining the hierarchical processing across brain regions. However, we observed a decrease in the discrimination ability of V1 from point-light videos to photorealistic avatar videos, indicating that the presentation format may have a greater impact on early visual regions. For photorealistic avatar videos, differences in visual appearance (e.g., color) may reduce V1 representation homogeneity of video instances in one semantic class. Interestingly, we did not observe superior performance of pSTS during MVPA decoding or superior correlation to the full-knowledge design matrix. Furthermore, the whole-brain searchlight analysis revealed a distributed network involved in human action recognition, including early motion processing regions, parietal cortex (somatosensory cortex, anterior intraparietal sulcus, and superior parietal lobule). These regions overlap with the mirror neuron system (MNS), consisting of the precentral gyrus, the inferior parietal lobule, the inferior frontal gyrus, and the STS (Cattaneo & Rizzolatti, 2009; Iacoboni & Dapretto, 2006).itrconsistin

and the motion pathway in the human visual system, respectively.

Simulation work (Feichtenhofer et al., 2016) suggests that the fusion of the activities in the last convolutional layers (i.e., "Conv5") of both streams consistently yields the best recognition accuracy across different data sets. Accordingly, the present article adopted this fusion architecture, where the two-stream DCNN model uses the fusion layer to first stack the outputs from the "conv5" layers of spatial pathway and temporal pathway. The stacked activities from $7 \times 7 \times 1024$ tensors provide inputs to a convolutional layer (also referred to as Conv5) consisting of 512 filters, followed by three FC layers, including a softmax layer.

1.2. Model training with natural action videos. The present article used the Human 3.6 M data set (Ionescu et al., 2014) to train the DCNN models with naturalistic red, green, blue videos. The Human 3.6 M data set (Ionescu et al., 2014; https://vision.imar.ro/human3.6m/description.php) includes 15 categories of actions: giving directions, discussing something with someone, eating, greeting someone, phoning, posing, purchasing (i.e., hauling up), sitting, sitting down, smoking, taking photos, waiting, walking, walking dog, and walking together. Each action was performed twice by each of the seven actors. For details of video generation, see Peng and colleagues (2021).

The DCNN models are trained to perform a 15-category action classification task. The action category with the highest score in the softmax layer is considered to be the model prediction for that instance. We follow a two-phase protocol to train the network as developed by Feichtenhofer and colleagues (2016). We first train the single-stream networks (spatial pathway and temporal pathway) independently with the task of 15-category action recognition. Then, activities from the conv5 layers of these two trained single-stream DCNNs are concatenated as inputs to train the fusion network in the two-stream DCNN. Simulation codes used in the current study are available at (Peng et al., 2021).

The training of DCNN models was assigned with a maximum of 100 epochs. Each epoch ran through a series of mini-batches of size 16. Gradient descent was calculated after each mini-batch through a Stochastic Gradient Descent optimizer (learning rate 10^{-4}) to update model weights. After each epoch, validation loss was calculated and model weights were saved if validation loss decreased compared with the previous epoc Training was terminated before reacng 100 epochs if vaidation loss remains without an increase for 10 consecutive epochs. Drop-out operations were implemented for training FC layers with a fraction of the input units to drop of 0.5 to prevent overfitting.

2. MRI Data Acquisition

MRI data were collected using a 3 T Siemens Prisma scanner with a 64-channel phase-array coil. In the scanner, the stimuli were back-projected via a projector (refresh rate: 60 Hz; spatial resolution: 1024×768) onto a translucent screen placed inside the scanner bore. Subjects viewed the stimuli through a mirror located above their eyes. The viewing distance was 90 cm. BOLD signals were measured using an EPI sequence with a multiband acceleration factor of (echo time = 30 msec; repetition time = 2000 msec; flip angle = 90°; acquisition matrix size = 112×112; filed of view = 224×224 mm^2; slice thickness = 2 mm; number of slices = 62; slice orientation = transversal). On each scan day, a T1-weighted, high-resolution, 3-D structural data set (3-D magnetization prepared rapid gradient echo, $0.5 \times 0.5 \times 1$-mm^3 resolution) was collected before functional runs.

3. MRI Data Analysis

3.1. Anatomical MRI analysis. We reconstructed the cortical surface of each participant using Freesurfer based on the T1 structural scan. This yielded a discrete triangular mesh representing the cortical surface used for the surface-based 2-D searchlight procedure outlined below.

3.2. Definition of fMRI ROIs. We defined five ROIs for each participant. Retinotopic visual areas were defined by a standard phase-encoded method (Engel et al., 1997; Sereno et al., 1995), in which participants viewed rotating wedge and expanding ring stimuli that created traveling waves of neural activity in visual cortex. Two runs of wedge rotation and two runs of expanding ring were conducted consecutively. MT+ was defined by the standard contrasts of moving dots with directions uni-

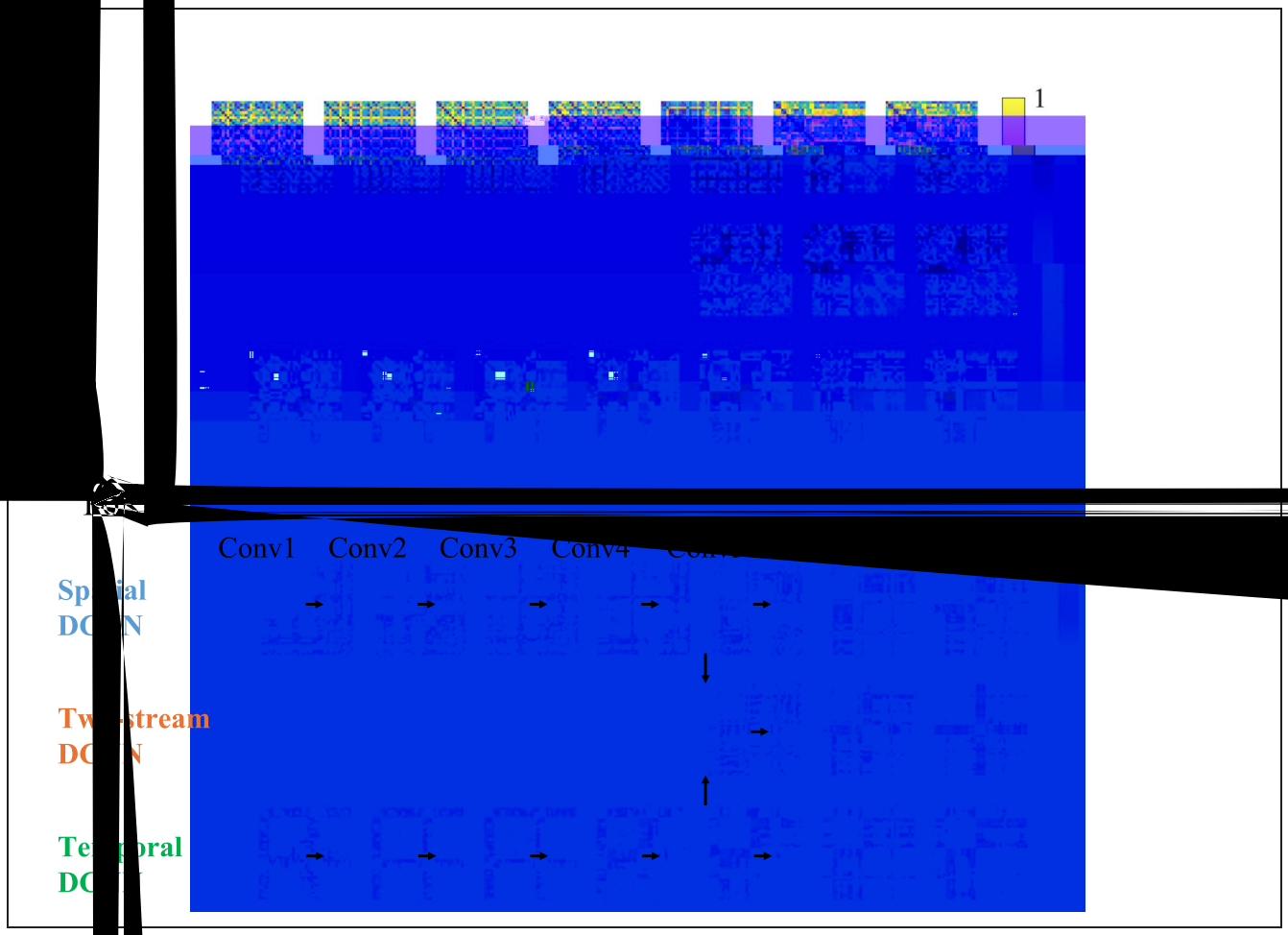
vo ns of localizers were employed for LOC, pS , each run containing 10 times of presen- each timulus condition, interleaved by 10 sec of A g eral linear model procedure was used for RO . Cortical reconstruction and volumetric atio were performed using the automated pro- bipe ne in FreeSurfer (Version 5.1.0; https:// mr. gh.harvard.edu/). ROIs were manually as a t of significantly responsive voxels to cor- ng c nditions (e.g., moving dots vs. static dots) in reeSurfer, guided by locations defined in lite ure, in a generally non-overlapping man- con cted analyses based on the top 50% of act ted voxels of the manually defined ROIs.

I an ysis. We preprocessed fMRI data using http /www.fil.ion.ucl.ac.uk/spm). For each of s, f I data were realigned and co-registered to ruct l scan collected on the same day. The T1 l sca from the last 4 days were further realigned egis red to the first day. Data were neither ed n smoothed because our analysis used the raw voxel-based data. We estimated the fMRI response patterns of the 36 videos in each of the 32 runs through a Least Squares Single voxelwise general linear model (Mumford, Davis, & Poldrack, 2014; Mumford, Turner, Ashby, & Poldrack, 2012), where each trial was modeled separately with one regressor and all other trials were modeled as a second regressor. Video onsets and duration were convolved with a hemodynamic response function. Movement parameters entered the general linear model as nuisance regressors. We then averaged the fMRI response patterns across all runs and applied *z*-normalizations to the beta values in each action instance to remove run-to-run variance.

Supplemental Results

4. Action Representation Dissimilar Matrices in DCNN and Brain Visual Regions

We inspected the representational geometry of each DCNN layer. Figure A1 showed that in all three DCNNs (a spatial pathway DCNN, a temporal pathway DCNN,



and a two-stream DCNN), each layer's RDM exhibited distinct structures in dissimilarity matrices, with a set of diagonal mini-blocks presented in some layers. Figure A2 showed similar representation structures of actions in all ROIs.

5. Examining ROI Hierarchy

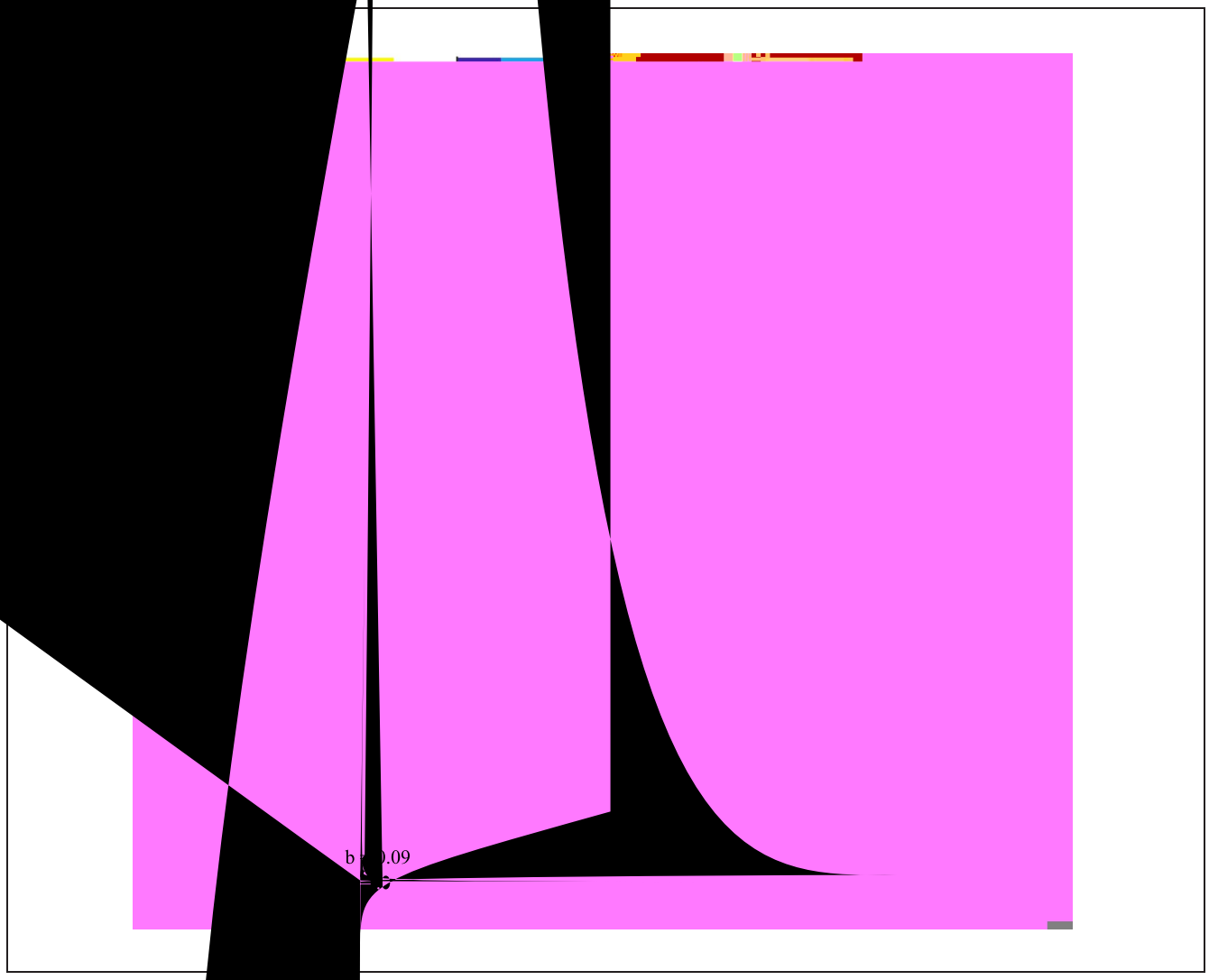
The hierarchical organization of brain ROIs serves as a simplified, yet very useful model for understanding information flow in the brain. This assumption is well supported by a wealth of anatomical and functional evidence, proving its efficacy in numerous studies, particularly those examining brain and CNN correlations.

To examine the hierarchical relationships between ROIs, we have computed the pairwise correlations of the RDMs from the brain ROIs, as shown in Figure A3A. Then, we grouped these pairwise correlations by their hierarchical "distance" (Figure A3B). Specifically, the distance between V1 and MT+, or MT+ and LOC, is considered one hierarchical step. Therefore, the distance between V1 and pSTS constitutes four hierarchical steps. Next, we computed the correlation between the group averaged pairwise correlation and the hierarchical step size. As illustrated in Figure A3C, these two variables demonstrated a significant linear relationship; as the hierarchical step size increases, the dissimilarity between ROI representations also increases (photorealistic avatar experiment: $b = 0.09$, $p = .043$; point-light experiment: $b = 0.11$, $p = .001$; averaged: $b = 0.09$, $p = .005$). This finding supports the assumption of hierarchical organization of selected ROIs in our study.

6. Peak Decoding Accuracy of DCNNs

For DCNNs, peak decoding accuracy was examined between spatial and temporal DCNN models. We also conducted bootstrapping analyses to examine where different DCNN pathways reached the maximum decoding accuracy at different layers. For the spatial and temporal DCNNs, layer numbers from 1 to 6 (5 convolutional layers and 1 FC layer) were sampled with replacement, weighted by the decoding accuracy (i.e., layers with greater decoding accuracy would be more likely to be selected). One hundred iterations were conducted, yielding 600 samples of layer indices. In each iteration, we calculated the frequ3o-1.1dd.8(ly)-3134[(of)-12[(th)-10.3(e)-404.9(t4.1(a)311.9rt4.1gt)-710

samples



RDMs of low-level based on the average speeds v and scat int-lights in biological motions (reflecting efined as:

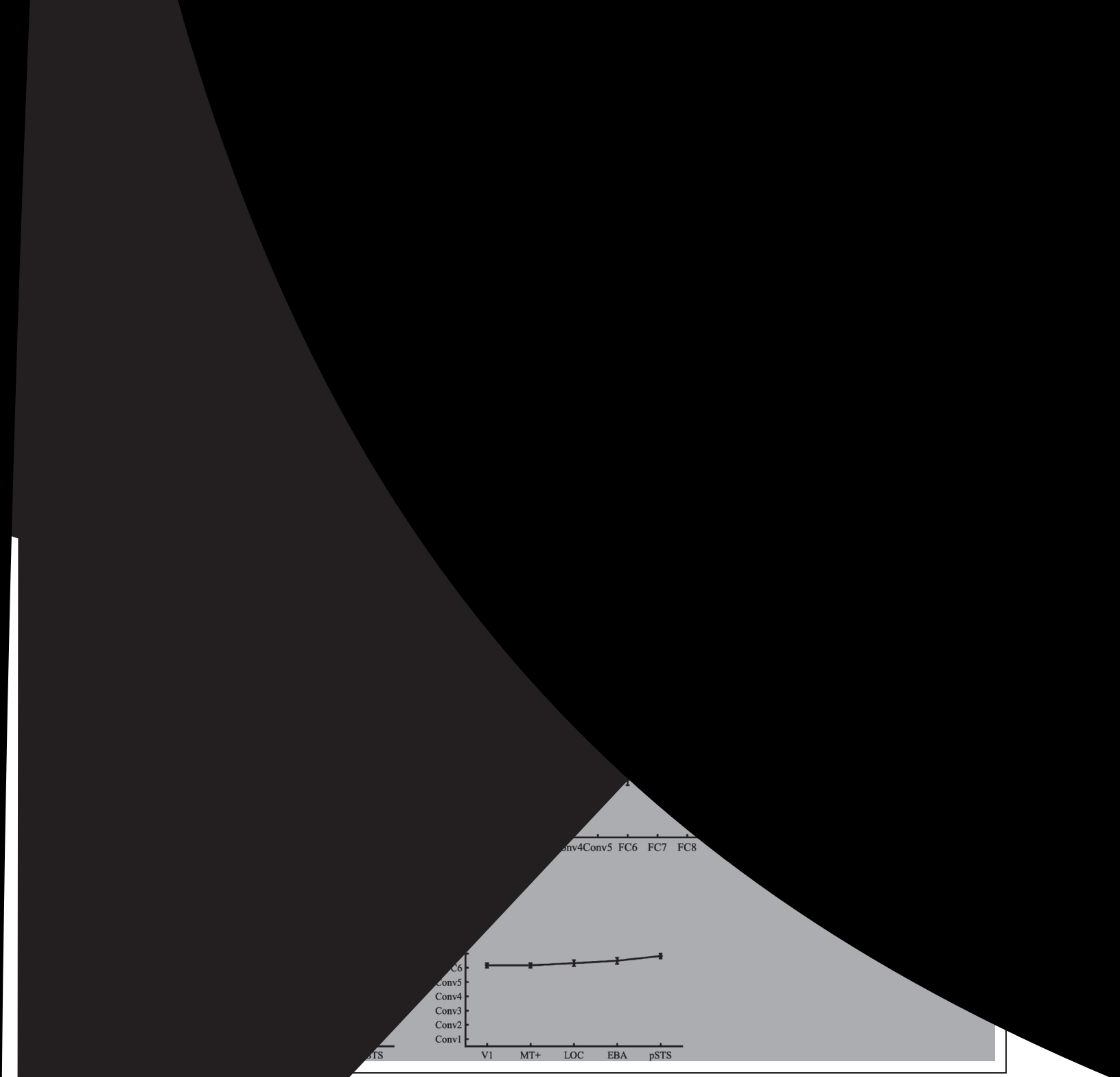
$$\langle v \rangle \equiv \sqrt{\left(\overline{\left(y_{t,i} - y_{t-1,i}\right)^2} \right.}$$

$$\langle r_c^2 \rangle \equiv \sqrt{\left(\overline{\left(y_{t,i} - y_{t,c}\right)^2} \right.}$$

where $x_{t,i}$ and $y_{t,i}$ ontal and vertical coordinates of dot i, refer to the averaged horizontal and ver among all dots at frame t of a point-light c presents how fast dots moved across tir he degree of motion information. Scat nts how far away dots are surrounding th body, indicating form information. RDMs ed and dot scatterness were calculated absolute differences between actions. correlations between ROI RDMs and RDMs of dot speed and scatterness were examined (Figure A5). Results showed that V1 yielded the greatest correlations to the speed and scatterness RDMs in Experiment 1, significantly greater than pSTS, LOC, and EBA on speed, and greater than pSTS and LOC on radius ($ps < .05$). This may suggest a dominant processing of low-level visual cues in V1. In contrast, pSTS yielded the lowest correlations to speed and scatterness RDM matrices in Experiment 1, and the lowest scatterness correlation in Experiment 2. These results confirm that pSTS is oriented toward higher-level visual processing of human actions compared with V1 and other selected early visual areas.

9. Action-only Design Matrix Correlations

In addition to the full-knowledge design matrix, we calculated an action-only design matrix to examine the



on of low-level visual information associated
action category. The action-only design matrix
s only higher similarity between videos within

Correlations between DCNN layer
only design matrix were sh
showed similar

design matrices than the full-knowledge design matrix for

Experiment 2 with point-light displays, whereas correlations to the action-only design matrix did not change much across the two experiments. We also found a significant two-way interaction between Design Matrices and ROIs, $F_{\text{Greenhouse–Geisser}}(1.94, 19.40) = 18.40$, $p < .01$, $\eta_p^2 = .684$. The three-way interaction between Experiments, Design Matrices, and ROIs was also significant, $F(4, 7) = 24.90$, $p < .001$, $\eta_p^2 = .934$.

Acknowledgments

We thank Qi Xie for helping generate the computer-rendered photorealistic avatar action stimuli, Tianmin Shu for assistance on DCNN model training, Junshi Lu for providing helpful advice on ROI selection and data analysis, and all the participants for their contribution to this research.

Corresponding authors: Yujia Peng or Fang Fang, School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Peking University, Haidian District, Beijing 100871, China, or via e-mail: yujia_peng@pku.edu.cn or ffang@pku.edu.cn.

Data Availability Statement

Data and scripts used in the current study can be found at: https://osf.io/jv8gp/.

Author Contributions

Yujia Peng: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Supervision; Visualization; Writing—Original draft; Writing—Review & editing. Xizi Gong: Data curation; Formal analysis; Methodology; Writing—Original draft; Writing—Review & editing. Hongjing Lu: Conceptualization; Formal analysis; Funding acquisition; Methodology; Supervision; Writing—Review & editing. Fang Fang: Conceptualization; Funding acquisition; Supervision.

Funding Information

This work was funded by National Science and Technology Innovation 2030 Major Program, grant number: 2022ZD0204802; the National Natural Science Foundation of China (https://dx.doi.org/10.13039/501100001809), grant numbers: 31930053 (to F. F.) and 32200854; and China Association for Science and Technology (https://dx.doi.org/10.13039/100010097), grant number: 2021QNRC00 to Y. P.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549,

https://doi.org/10.1371/journal.pcbi.1003963, PubMed: 25521294
Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299–6308). https://doi.org/10.1109/CVPR.2017.502
Cattaneo, L., & Rizzolatti, G. (2009). The mirror neuron system. *Archives of Neurology*, *66*, 557–560. https://doi.org/10.1001/archneurol.2009.41, PubMed: 19433654
Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *Neuroimage*, *153*, 346–358. https://doi.org/10.1016/j.neuroimage.2016.03.063, PubMed: 27039703
Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755. https://doi.org/10.1038/srep27755, PubMed: 27282108
Cutting, J. E., Moore, C., & Morrison, R. (1988). Masking the motions of human gait. *Perception & Psychophysics*, *44*, 339–347. https://doi.org/10.3758/BF03210415, PubMed: 3226881
Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception*, *22*, 15–22. https://doi.org/10.1068/p220015, PubMed: 8474831
Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*, 2470–2473. https://doi.org/10.1126/science.1063414, PubMed: 11577239
Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, *152*, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001, PubMed: 27777172
Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, *7*, 181–192. https://doi.org/10.1093/cercor/7.2.181, PubMed: 9087826
Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6201–6210). https://doi.org/10.1109/ICCV.2019.00630
Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1933–1941). https://doi.org/10.1109/CVPR.2016.213
Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, *4*, 179–192. https://doi.org/10.1038/nrn1057, PubMed: 12612631
Grossman, E. D., Battelli, L., & Pascual-Leone, A. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. *Vision Research*, *45*, 2847–2853. https://doi.org/10.1016/j.visres.2005.05.027, PubMed: 16039692
Grossman, E. D., & Blake, R. (2001). Brain activity evoked by inverted and imagined biological motion. *Vision Research*, *41*, 1475–1482. https://doi.org/10.1016/S0042-6989(00)00317-5, PubMed: 11322987
Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, *35*, 1167–1175. https://doi.org/10.1016/S0896-6273(02)00897-8, PubMed: 12354405
Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience*, *12*, 711–720. https://doi.org/10.1162/089892900562417, PubMed: 11054914
Grossman, E. D., Jardine, N. L., & Pyles, J. A. (2010). FMR-adaptation reveals invariant coding of biological motion on human STS. *Frontiers in Human Neuroscience*, *4*, 15. https://doi.org/10.3389/neuro.09.015.2010, PubMed: 20431723
Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*, 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015, PubMed: 26157000
Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*, 613–622. https://doi.org/10.1038/nn.4247, PubMed: 26900926
Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*, 185–203. https://doi.org/10.1016/0004-3702(81)90024-2
Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, *7*, 942–951. https://doi.org/10.1038/nrn2024, PubMed: 17115076
Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*, 1325–1339. https://doi.org/10.1109/TPAMI.2013.248, PubMed: 26353306
Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *Neuroimage*, *14*, S103–S109. https://doi.org/10.1006/nimg.2001.0832, PubMed: 11373140
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*, 201–211. https://doi.org/10.3758/BF03212378
Khaligh-Razavi, S.-M., Cichy, R. M., Pantazis, D., & Oliva, A. (2018). Tracking the spatiotemporal neural dynamics of real-world object size and animacy in the human brain. *Journal of Cognitive Neuroscience*, *30*, 1559–1576. https://doi.org/10.1162/jocn_a_01290, PubMed: 29877767
Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., & Kriegeskorte, N. (2017). Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, *76*, 184–197. https://doi.org/10.1016/j.jmp.2016.10.007, PubMed: 28298702
Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*, e1003915. https://doi.org/10.1371/journal.pcbi.1003915, PubMed: 25375136
Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*, 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043, PubMed: 19109916
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.
Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision* (pp. 2556–2563). https://doi.org/10.1109/ICCV.2011.6126543
Lange, J., Georg, K., & Lappe, M. (2006). Visual perception of biological motion by form: A template-matching analysis. *Journal of Vision*, *6*, 836–849. https://doi.org/10.1167/6.8.6, PubMed: 16895462
Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324. https://doi.org/10.1109/5.726791
Lingnau, A., & Downing, P. E. (2015). The lateral occipitotemporal cortex in action. *Trends in Cognitive Sciences*, *19*, 268–277. https://doi.org/10.1016/j.tics.2015.03.006, PubMed: 25843544

Lu, H., & Liu, Z. (2006). Computing dynamic classification images from correlation maps. *Journal of Vision*, *6*, 475–483. https://doi.org/10.1167/6.4.12, PubMed: 16889481

Mahowald, K., & Fedorenko, E. (2016). Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *Neuroimage*, *139*, 74–93. https://doi.org/10.1016/j.neuroimage.2016.05.073, PubMed: 27261158

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., et al. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *92*, 8135–8139. https://doi.org/10.1073/pnas.92.18.8135, PubMed: 7667258

McMahon, E., Bonner, M. F., & Isik, L. (2023). Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, *33*, 5035–5047. https://doi.org/10.1016/j.cub.2023.10.015, PubMed: 37918399

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *Neuroimage*, *103*, 130–138. https://doi.org/10.1016/j.neuroimage.2014.09.026, PubMed: 25241907

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*, 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076, PubMed: 21924359

Naselaris, T., Allen, E., & Kay, K. (2021). Extensive sampling for complete models of individual brains. *Current Opinion in Behavioral Sciences*, *40*, 45–51. https://doi.org/10.1016/j.cobeha.2020.12.008

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*, e1003553. https://doi.org/10.1371/journal.pcbi.1003553, PubMed: 24743308

Peng, Y., Lee, H., Shu, T., & Lu, H. (2021). Exploring biological motion perception in two-stream convolutional neural networks. *Vision Research*, *178*, 28–40. https://doi.org/10.1016/j.visres.2020.09.005, PubMed: 33091763

Peng, Y., Thurman, S., & Lu, H. (2017). Causal action: A fundamental constraint on perception and inference about body movements. *Psychological Science*, *28*, 798–807. https://doi.org/10.1177/0956797617697739, PubMed: 28481714

Pinto, J., & Shiffrar, M. (1999). Subconfigurations of the human form in the perception of biological motion displays. *Acta Psychologica*, *102*, 293–318. https://doi.org/10.1016/S0001-6918(99)00028-1, PubMed: 10504885

Pollick, F. E., Lestou, V., Ryu, J., & Cho, S.-B. (2002). Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Research*, *42*, 2345–2355. https://doi.org/10.1016/S0042-6989(02)00196-7, PubMed: 12350423

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230, PubMed: 15217330

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *Neuroimage*, *206*, 116189. https://doi.org/10.1016/j.neuroimage.2019.116189, PubMed: 31521825

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences, U.S.A.*, *116*, 11537–11546. https://doi.org/10.1073/pnas.1820226116, PubMed: 31101713

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., et al. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *Neuroimage*, *180*, 253–266. https://doi.org/10.1016/j.neuroimage.2017.07.018, PubMed: 28723578

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., et al. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, *268*, 889–893. https://doi.org/10.1126/science.7754376, PubMed: 7754376

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *arXiv*. https://doi.org/10.48550/arXiv.1406.2199

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv*. https://doi.org/10.48550/arXiv.1212.0402

Theusner, S., de Lussanet, M. H. E., & Lappe, M. (2011). Adaptation to biological motion leads to a motion and a form aftereffect. *Attention, Perception, & Psychophysics*, *73*, 1843–1855. https://doi.org/10.3758/s13414-011-0133-7, PubMed: 21598067

Thurman, S. M., van Boxtel, J. J., Monti, M. M., Chiang, J. N., & Lu, H. (2016). Neural adaptation in pSTS correlates with perceptual aftereffects to biological motion and with autistic traits. *Neuroimage*, *136*, 149–161. https://doi.org/10.1016/j.neuroimage.2016.05.015, PubMed: 27164327

Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., & Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences, U.S.A.*, *98*, 11656–11661. https://doi.org/10.1073/pnas.191374198, PubMed: 11553776

van Boxtel, J. J. A., & Lu, H. (2012). Signature movements lead to efficient search for threatening actions. *PLoS One*, *7*, e37085. https://doi.org/10.1371/journal.pone.0037085, PubMed: 22649510

van Boxtel, J. J. A., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. *Journal of Vision*, *13*, 7. https://doi.org/10.1167/13.12.7, PubMed: 24130256

van Boxtel, J. J. A., & Lu, H. (2015). Joints and their relations as critical features in action discrimination: Evidence from a classification image method. *Journal of Vision*, *15*, 20. https://doi.org/10.1167/15.1.20, PubMed: 25604612

Watson, J. D. G., Myers, R., Frackowiak, R. S. J., Hajnal, J. V., Woods, R. P., Mazziotta, J. C., et al. (1993). Area V5 of the human brain: Evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cerebral Cortex*, *3*, 79–94. https://doi.org/10.1093/cercor/3.2.79, PubMed: 8490322

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, *12*, 2065. https://doi.org/10.1038/s41467-021-22244-7, PubMed: 33824315

Yamins, D., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)* (pp. 3093–3101). Curran Associates Inc.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111, PubMed: 24812127

Zeki, S., Watson, J. D., Lueck, C. J., Friston, K. J., Kennard, C., & Frackowiak, R. S. (1991). A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience*, *11*, 641–649. https://doi.org/10.1523/JNEUROSCI.11-03-00641.1991, PubMed: 2002358