Appendix: A Dual-Stream Neural Network Explains the Functional Segregation of Dorsal and Ventral Visual Pathways in Human Brains

Minkyu Choi[1], Kuan Han[1],
Xiaokai Wang[2], Yizhen Zhang[1,3], and Zhongming Liu[1,2]

[1] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109
[2] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109
[3] Department of Neurological Surgery , University of California, San Francisco, San Francisco, CA 94143
{cminkyu, kuanhan, xiaokaiw, zhyz, zmliu}@umich.edu

A Regions of Interests

In our study, we delineated our regions of interest (ROIs) into two primary segments: 1) the ventral visual stream and object recognition-related regions and 2) the dorsal visual stream and overt attention-related regions. This approach followed the parcellations proposed by [1]. For the dorsal visual stream, the ROIs includes V3A, V3B, V6, V6A, and V7. Within the parietal cortex, visuo-spatial information and overt attention are processed by the intraparietal sulcus (IPS) and the superior parietal lobule (SPL) [2, 3, 4, 5, 6]. The IPS encompasses V7, IPS1, IP0, IP1, and IP2; whereas the SPL consists of lateral intraparietal cortex (LIPv, LIPd), ventral intraparietal complex (VIP), anterior intraparietal (AIP), medial intraparietal area (MIP), 7PC, 7AL, 7Am, 7PL, and 7Pm. We also included the frontal eye field (FEF), which is acknowledged for controlling eye movements [7, 8, 9, 10]. In contrast, the ROIs associated with object recognition and the ventral visual stream encompassed V8, the posterior inferotemporal (PIT) complex, the fusiform face complex (FFC), and ventromedial visual (VMV) areas 1, 2, 3, along with the lateral occipital area (LO). In addition, we included the superior temporal sulcus (STS), which is recognized for processing multimodal signals, including auditory and visual cues [11, 12, 13]. Fig. S1 displays the full set of region labels, corresponding to Fig.3(a) from the main text. Among the parcellations by [1], regions including significantly predicted voxels either by the WhereCNN or WhatCNN are presented in Fig. S1.

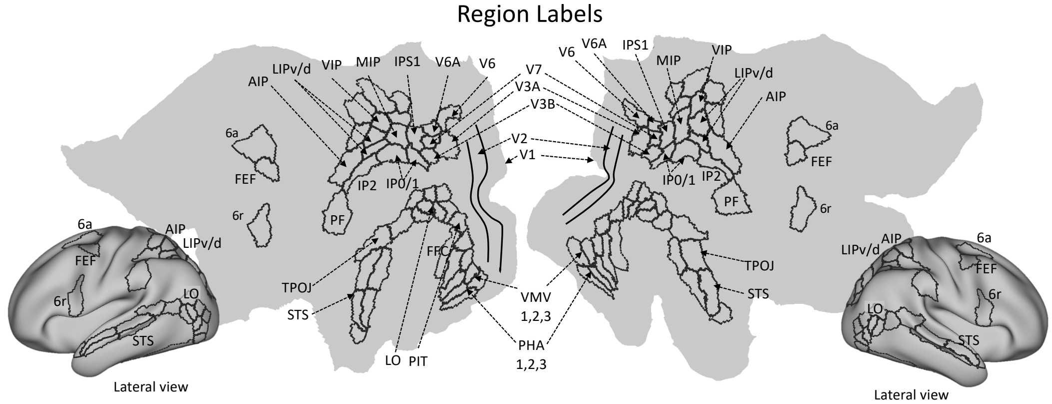


Figure S1: Region labels. Regions including significant voxels from Fig.3(a) in the main text are presented.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

B Training Details

The backbone Convolutional Neural Networks (CNNs) of both the WhereCNN and WhatCNN share the same architecture, consisting of four blocks of convolutional operations. Situated atop the backbone CNN, the WhereCNN and WhatCNN possess additional layers tailored to their specific objectives: the WhereCNN features two convolutional layers that produce 2D saliency maps, whereas the WhatCNN includes a Gated Recurrent Unit (GRU) layer followed by a fully connected layer for object classification.

During the pre-training of the backbone CNN, a global average pooling and a fully connected layer are integrated atop the backbone CNN, serving as a classifier. Upon completion of the pre-training process, the classifier is detached, allowing the pre-trained backbone CNN to be incorporated as a component of the WhereCNN or WhatCNN.

As detailed in Section 3.1 of the main text, our model underwent a three-stage training process. In this section, we will elaborate on the specifics of the pre-training phase.

Stage 1 - WhereCNN The backbone architecture of the WhereCNN was pre-trained on ILSVRC2012 [14] for an image classification task over 120 epochs. A batch size of $1,024$ was employed, along with the Adam optimizer [15] (lr=0.001, β_1=0.9, β_2=0.99). During pre-training, fixations for the retinal transformation were randomly generated across the image area. Once the backbone architecture had been pre-trained, we detached the classifier and initialized the WhereCNN using the model parameters obtained from the pre-training stage. We then performed SALICON training, as described in Section 3.1 of the main text.

Stage 2 - WhatCNN In a process mirroring Stage 1, the backbone of the WhatCNN was also pre-trained on ILSVRC2012 [14] for an image classification task over 120 epochs, utilizing random fixations and the Adam optimizer (lr=0.001, β_1=0.9, β_2=0.99). After pre-training the backbone CNN, we initialized the WhatCNN using the weights of the pre-trained backbone CNN.

Subsequently, the WhatCNN, initialized with the pre-trained weights as a whole, was trained on ILSVRC2012 [14] for object recognition using four fixations. Four randomly generated fixations were employed for training the WhatCNN for 55 epochs, again utilizing the Adam optimizer (lr=0.001, β_1=0.9, β_2=0.99). After this stage, we conducted a fine-tuning process using the learned fixations from the WhereCNN. In this stage, the WhereCNN, after the pre-training in Stage 1, was incorporated to guide the WhatCNN's fixations. However, only the WhatCNN was optimized, while the WhereCNN remained unchanged. This fine-tuning with learned fixations deployed four gazes, utilizing the Adam optimizer (lr=0.0001, β_1=0.9, β_2=0.99) over 25 epochs. Finally, the WhatCNN underwent further training on MSCOCO, as described in Section 3.1 of the main text.

Stage3 - WhereCNN & WhatCNN During this stage, both WhereCNN and WhatCNN, trained in the previous stages, were used to initialize model weights, followed by further end-to-end training, leveraging the stream-specific objectives (object recognition and saliency prediction, respectively). As the training requires labels for both tasks, the model was trained using images in the SALICON dataset, which contain labels for both saliency prediction and object recognition.

The model samples fixations from the predicted saliency maps from WhereCNN. As this sampling process is non-differentiable, the gradients from object recognition cannot optimize the weights of WhereCNN. To tackle this issue, we utilized REINFORCE [16] to approximate the gradient for WhereCNN. At the time t, a fixation $\boldsymbol{l}_t$ is generated by WhereCNN, based on which WhatCNN predicts a class prediction p_t. Then, in the context of REINFORCE, the reward r_t of choosing $\boldsymbol{l}_t$ as the fixation is calculated as the reduced classification loss relative to the previous time step $r_t = CE(p_{t-1}, \text{label}_c) - CE(p_t, \text{label}_c)$, where CE is the cross-entropy loss, label_c is class labels. The goal of REINFORCE is to maximize the discounted sum of rewards, $R = \sum_{t=1}^{T} \gamma^{t-1} r_t$, where $\gamma \in (0, 1)$ is the discount factor and set as 0.8.

In this stage, we strived to minimize the object recognition and saliency prediction losses while maximizing the discounted sum of rewards. As indicated in Section 3.1 of the main text, we utilized the Adam optimizer (lr=0.0002, β_1=0.9, β_2=0.99) for 25 epochs for this training stage.

For All Stages All training stages were conducted using four NVIDIA A40 GPUs. All codes are written in Pytorch 1.9.1.

C Saliency Maps and Inhibition of Returns

Once the saliency maps were generated by WhereCNN, inhibition of return (IOR) was used to prohibit future fixations to re-visit image areas that had been already explored. This process is illustrated in Fig. S2

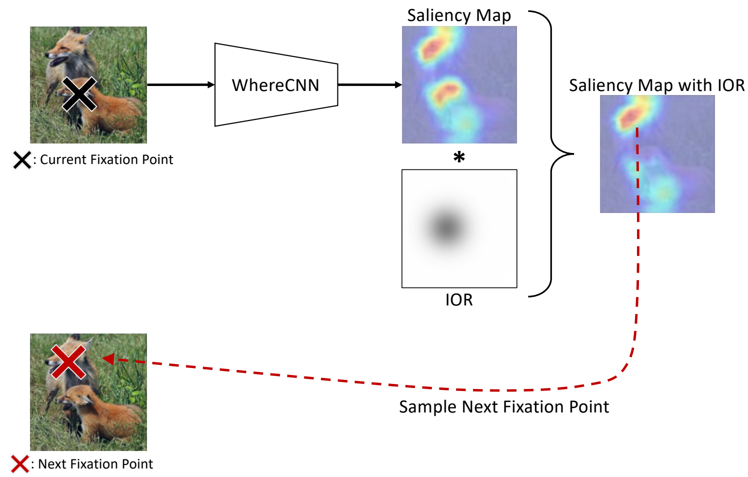


Figure S2: Process of determining the next fixation point given the current fixation. A saliency map, generated by WhereCNN, is multiplied element-wise (indicated by $*$) with the inhibition of return (IOR) to prevent future fixations from reverting to previous positions. In the IOR, white and black colors correspond to values of 1 and 0, respectively.

In the process of determining the next fixation, the WhereCNN generate a saliency map based on the current fixation. The location of this subsequent fixation is guided by the saliency map's probabilistic distribution. However, it's important to note that if the current fixation point possesses a high probability, subsequent fixations are likely to occur in proximity to the present fixation.

To ensure a more dynamic and comprehensive exploration of the visual field, we employed the principle of Inhibition of Return (IOR), detailed in Eq.2 of the main text, and presented again here in Eq.4.

$$\mathbf{IOR}(t) = \mathbf{ReLU}\Big(\mathbf{1} - \sum_{\tau=1}^{t} G(\boldsymbol{\mu} = \boldsymbol{l}_\tau, \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I})\Big) \tag{4}$$

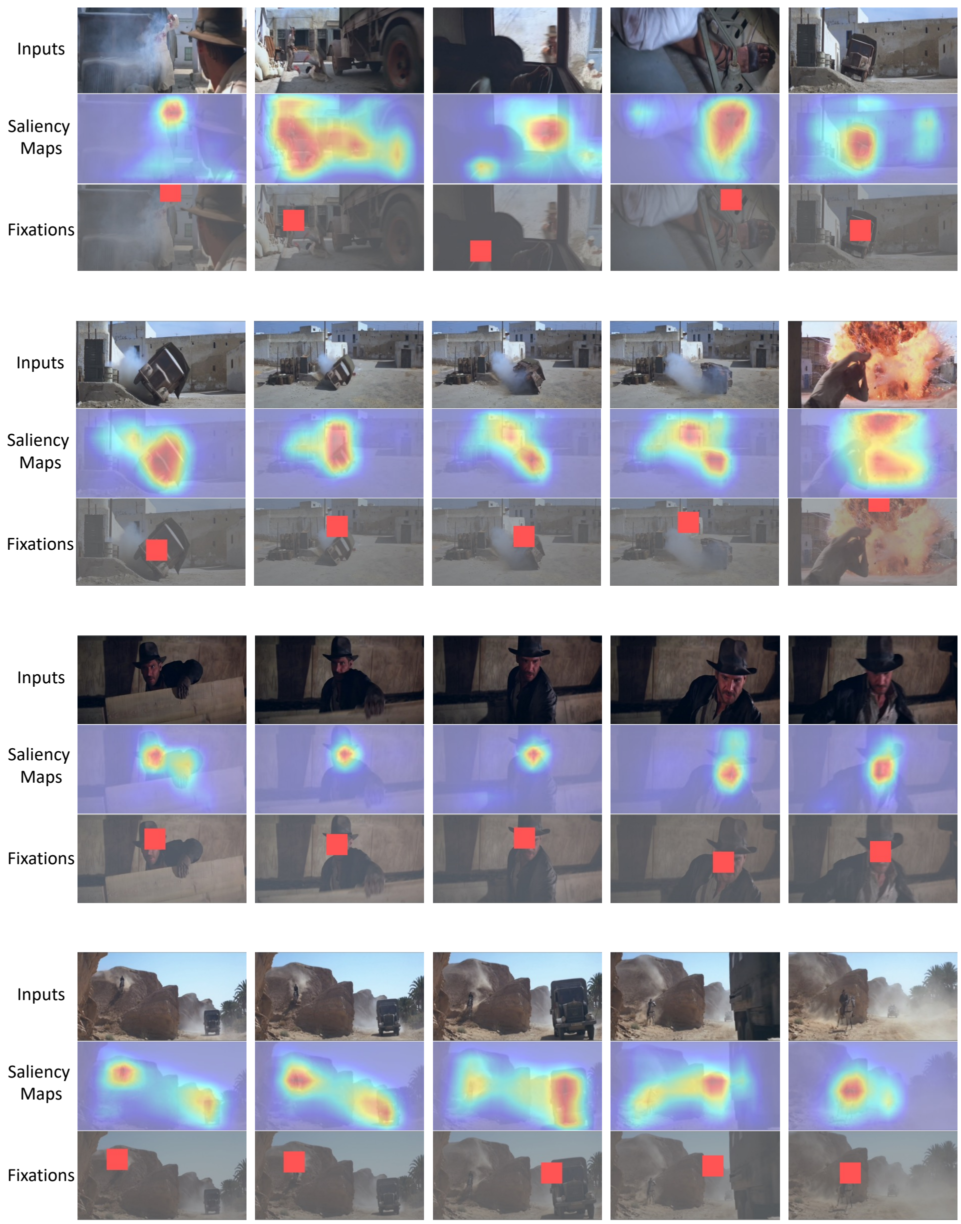
where $G(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a 2D Gaussian function centered at $\boldsymbol{l}_\tau$ (prior fixations) with a standard deviation σ at the τ-th step. The Inhibition of Return (IOR) is initially created at a resolution of 224×224 with $\sigma = 25$, and subsequently resized to align with the dimensions of the saliency map. IOR serves to decrease the saliency of previously attended areas, thereby preventing the model from repetitively focusing on these regions. This mechanism is informed by the model's all prior fixation history. The IOR map is designed such that it assigns lower values (approaching 0.0) in the vicinity of prior fixation points, and higher values (up to 1.0) in regions further away. Thus, when the IOR map is element-wise multiplied with the saliency map, it effectively reduces the saliency values in areas already explored.

Following the application of IOR, the subsequent fixation point is decided upon by considering the adjusted saliency map. It is then chosen based on the probabilistic distribution within this updated map. This strategy encourages more diverse fixations and facilitates a broader and more comprehensive understanding of the scene.

D WhereCNN's Saliency Maps and Fixation Points

The original images are presented in Cartesian coordinates. Once the retinal transformation is applied to these images, the resultant retinal images adopt retinal coordinates, as detailed in Eq.1 of the main text. Since the inputs to the WhereCNN operate in retinal coordinates, it naturally follows that the output saliency maps mirror this coordinate system. To visualize these within this paper, we utilize the inverse function of Eq.1, thereby transforming the saliency maps from retinal back to Cartesian coordinates.

In preparation for our model's processing of the movie *Raiders of the Lost Ark*, we reduce the frame rate to 6 frames per second (fps). This adjustment helps mitigate computational and memory costs associated with the handling of the extracted features. As the model engages with the movie, a solitary fixation point is established for each frame. Importantly, the Inhibition of Return (IOR) mechanism is not invoked during the model's interaction with the movie. Fig. S3 showcases saliency maps and fixation points derived from segments of the movie *Raiders of the Lost Ark*. Frames situated on the same horizontal axis are selected at a rate of 1 fps.



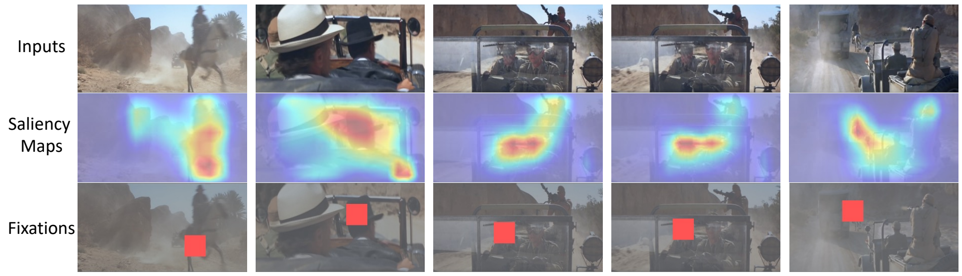


Figure S3: Given the movie frames (1st row), the WhereCNN generates saliency maps (2nd row) and fixations (3rd row). The red marker in the 3rd row presents the fixation point.

E Investigating Layer-wise Correspondence to Visual Cortex

In the main text, the whole features from the all layers of each stream are used for predicting voxel activities (noted as Stream-wise encoding). In an alternative way, the features from each layer can be used to predict voxel activities, instead of concatenating all the layers, (noted as Layer-wise encoding). In this way, the hierarchical correspondence between each layer in the model to the ROIs of the visual system can be observed.

With the layer-wise encoding scheme, we predicted fMRI responses using features from each layer in the WhereCNN and WhatCNN. Fig. S4 associates each voxel to one (color-coded) layer most predictive of that voxel for either (a) WhatCNN or (b) WhereCNN. Fig. S4 (a) shows that the lower layers of WhatCNN better predict earlier visual areas such as V1/V2, whereas the higher layers of WhatCNN better predict higher-order visual areas such as LO and PIT, consistent with prior studies [17, 18]. The results with the WhereCNN show different patterns, as shown in Fig. S4 (b). Within early visual areas, the lower layers of WhereCNN better predict foveal representations, whereas the higher layers better predict peripheral representations.

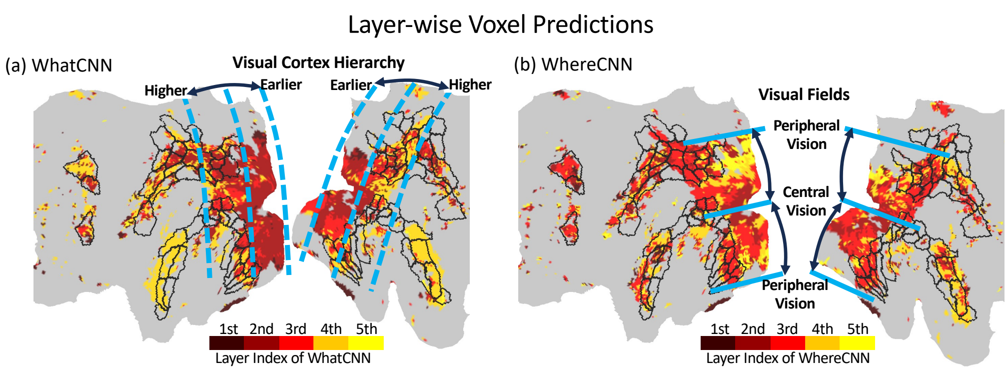


Figure S4: Each voxel is predicted by the features from a single layer from (a) WhatCNN and (b) WhereCNN. Layer indexes are color-coded so that the layer best predicting each voxel is presented.

F Implications to the Computer Visions

In the current study, we demonstrated that the biologically plausible components (two stream, retinal sampling and eye movements) can be used to build a better model for the human visual cortex in a naturalistic viewing condition. At the same time, those components we considered in this study may also bring benefits to the computer vision applications.

1) **Efficiency**. Unlike conventional CNNs that process entire images, our dual-stream model allows serial processing. It concentrates processing power on key image regions through attention directed fixations. This serial processing may significantly lower memory and computational overhead,

because resources are allocated only to the crucial image regions. It is plausible that such efficiency underpins the brain's adoption of dual stream processing due to biological constraints on energy use.

2) **Adaptability**. The dual streams of our model offer complementary lenses for visual exploration and perception in real-world environments. One stream provides a broad yet rough overview of the environment. The other gathers detailed observations with precision. Their synergistic interaction may facilitate adaptive behaviors for tasks like visual search, object detection in complex and cluttered scenes. Moreover, the distinct functions of each of the parallel streams present a combinatorial flexibility when leveraged together, potentially enhancing the model's overall capability to adapt to diverse visual challenges, including potential applications in robotics.

However, leveraging such potential benefits within the scope of current study face challenges. First, mainstream datasets like ImageNet and MS-COCO offer a narrow view and lack the high-resolution detail our model thrives on. Moreover, these datasets often focus on large, central objects, limiting our model's adaptability that benefits object recognition. A better benchmark to our model would be high-resolution panoramic images or synthetic virtual reality environments to accommodate unlimited fixation variances. In such settings, the efficiency and adaptability of our model should be more appealing.

References

[1] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[2] Jacqueline P Gottlieb, Makoto Kusunoki, and Michael E Goldberg. The representation of visual salience in monkey parietal cortex. *Nature*, 391(6666):481–484, 1998.

[3] Anna E Ipata, Angela L Gee, Jacqueline Gottlieb, James W Bisley, and Michael E Goldberg. Lip responses to a popout stimulus are reduced if it is overtly ignored. *Nature neuroscience*, 9(8):1071–1076, 2006.

[4] Makoto Kusunoki, Jacqueline Gottlieb, and Michael E Goldberg. The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus motion, and task relevance. *Vision research*, 40(10-12):1459–1468, 2000.

[5] James W Bisley, Koorosh Mirpour, Fabrice Arcizet, and Wei S Ong. The role of the lateral intraparietal area in orienting attention and its implications for visual search. *European Journal of Neuroscience*, 33(11):1982–1990, 2011.

[6] James W Bisley and Michael E Goldberg. Attention, intention, and priority in the parietal lobe. *Annual review of neuroscience*, 33:1–21, 2010.

[7] Hugo L Fernandes, Ian H Stevenson, Adam N Phillips, Mark A Segraves, and Konrad P Kording. Saliency and saccade encoding in the frontal eye field during natural scene search. *Cerebral Cortex*, 24(12):3232–3245, 2014.

[8] Kirk G Thompson and Narcisse P Bichot. A visual salience map in the primate frontal eye field. *Progress in brain research*, 147:249–262, 2005.

[9] Charles J Bruce, Michael E Goldberg, M Catherine Bushnell, and Gregory B Stanton. Primate frontal eye fields. ii. physiological and anatomical correlates of electrically evoked eye movements. *Journal of neurophysiology*, 54(3):714–734, 1985.

[10] David A Robinson and Albert F Fuchs. Eye movements evoked by stimulation of frontal eye fields. *Journal of neurophysiology*, 32(5):637–648, 1969.

[11] Jon Driver and Toemme Noesselt. Multisensory interplay reveals crossmodal influences on 'sensory-specific'brain regions, neural responses, and judgments. *Neuron*, 57(1):11–23, 2008.

[12] Gemma A Calvert. Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral cortex*, 11(12):1110–1123, 2001.

[13] Michael S Beauchamp. Statistical criteria in fmri studies of multisensory integration. *Neuroinformatics*, 3:93–113, 2005.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[17] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[18] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12):4136–4160, 2018.