# 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], **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
{cminkyu, kuanhan, xiaokaiw, zhyz, zmliu}@umich.edu

## Abstract

The human visual system uses two parallel pathways for spatial processing and object recognition. In contrast, computer vision systems tend to use a single feedforward pathway, rendering them less robust, adaptive, or efficient than human vision. To bridge this gap, we developed a dual-stream vision model inspired by the human eyes and brain. At the input level, the model samples two complementary visual patterns to mimic how the human eyes use magnocellular and parvocellular retinal ganglion cells to separate retinal inputs to the brain. At the backend, the model processes the separate input patterns through two branches of convolutional neural networks (CNN) to mimic how the human brain uses the dorsal and ventral cortical pathways for parallel visual processing. The first branch (WhereCNN) samples a global view to learn spatial attention and control eye movements. The second branch (WhatCNN) samples a local view to represent the object around the fixation. Over time, the two branches interact recurrently to build a scene representation from moving fixations. When compared to brain responses in humans watching a movie, the WhereCNN and WhatCNN branches matched the dorsal and ventral pathways of the visual cortex, respectively. We also conducted experiments to disentangle the roles of retinal sampling, learning objective, and attention-guided eye movement in the functional alignment between the model and the brain. By evaluating the linear alignment between dual-stream models and brains during dynamic natural vision, we infer that the distinct responses and representations of the ventral and dorsal streams are more influenced by their distinct goals in visual attention and object recognition than by their specific bias or selectivity in retinal inputs. This dual-stream model takes a further step in brain-inspired computer vision, enabling parallel neural networks to actively explore and understand the visual surroundings.

## 1 Introduction

The human visual system comprises two parallel and segregated streams of neural networks: the "where" stream and the "what" stream [54]. The "where" stream originates from magnocellular retinal ganglion cells and extends along the dorsal visual cortex. The "what" stream originates from parvocellular retinal ganglion cells and extends along the ventral visual cortex [50]. The two streams exhibit selective responses to different aspects of visual stimuli [57]. The "where" stream is tuned to coarse but fast information from a wide view, while the "what" stream is selective to fine but slow information from a narrow view [50, 48]. The two streams are thought to serve different purposes.

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

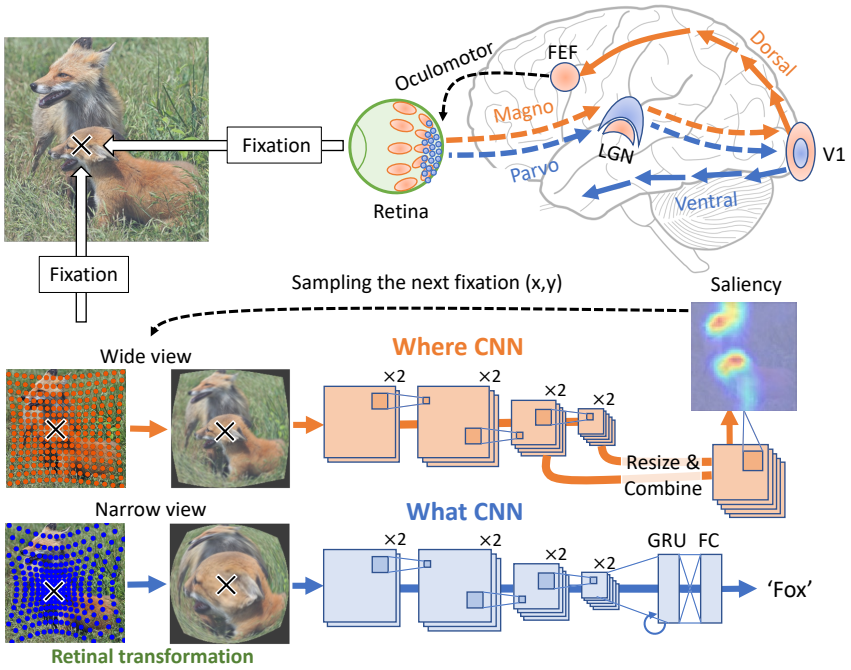


Figure 1: **Brain-inspired dual-stream vision model.** The top illustrates the subcortical (dashed arrows) and cortical (solid arrows) pathways for parallel visual processing in the brain. Given a scene (e.g., "two foxes on the lawn"), the retina samples incoming light relative to the fixation of the eyes (shown as the cross). Magnocellular (orange) and parvocellular (blue) retinal ganglion cells encode complementary visual information into two sets of retinal inputs relayed onto separate layers in the lateral geniculate nuclei (LGN) and further onto different neurons in the primary visual cortex (V1). Within V1, the relative ratio of magnocellular vs. parvocellular projections is higher for the periphery and lower for the fovea. Beyond V1, the magnocellular pathway continues along the dorsal visual cortex towards the intraparietal areas and further onto the frontal eye field (FEF) for oculomotor control, while the parvocellular pathway continues along the ventral visual cortex towards the inferior temporal cortex and further onto the superior temporal areas for semantic cognition. The bottom illustrates our model architecture including WhereCNN and WhatCNN. The model's frontend mimics the human retina and generates two separate input patterns relative to the fixation. One pattern is wider but coarser while the other is narrower but finer, providing the respective inputs to WhereCNN and WhatCNN. With the wide-view input, WhereCNN generates a probability map of saliency from which the next fixation is sampled. With a narrow-view input, WhatCNN generates an object representation per each fixation and constructs a scene representation recurrently from multiple fixations.

The "where" stream zooms out for spatial analysis [33], visual attention [59, 13, 17], and guiding actions [26] such as eye movements [11], while the "what" stream zooms in to recognize the object around the fixation [64]. While being largely parallel, the two streams interact with each other [51]. In one way of their interaction, the "where" stream decides where to look next and guides the "what" stream to focus on a salient location for visual perception. As the eyes move around the visual environment, the interaction between the "where" and "what" streams builds a scene representation by accumulating object representations over time and space. This dual-stream architecture allows the brain to efficiently process visual information and support dynamic visual behaviors [36].

In contrast, computer vision systems tend to use a single stream of feedforward processing, acting as passive observers that sample visual information all at once with fixed and uniform patterns [45, 35, 19]. Compared to human vision, this processing is less robust, especially given adversarial attacks [63, 27]; it is less efficient since it samples visual information equally regardless of salience or nuisance [8]; it is less adaptive, lacking spatial attention for active sensing [38, 55]. These distinctions define a major gap between human and computer vision. Many visual tasks that are straightforward for humans are still challenging for machines [46, 16]. Therefore, computer vision may benefit from

taking further inspiration from the brain by using a dual-stream architecture to learn adaptive and robust visual behaviors.

To gain insights into the computational mechanisms of human vision, researchers have developed image-computable models by utilizing goal-driven deep neural networks that simulate human perceptual behavior. In particular, convolutional neural networks (CNNs) are leading models of visual perception, capturing the hierarchical processing by the brain's ventral visual stream [74, 28, 20, 41, 73]. Previous models of this nature commonly utilize CNNs trained through supervised learning [74, 41, 28, 10, 20, 71], adversarial training [15, 5], unsupervised learning [31], or self-supervised learning [76, 67, 44]. However, models of the dorsal stream remain relatively under-explored, despite few studies [58, 53, 29, 3]. Existing testing of these models has primarily focused on static images presented briefly to the fovea, thus limiting their assessment to a narrow range of visual behaviors and processes [62]. A more comprehensive approach is needed to develop models that incorporate both dorsal and ventral stream processing and to assess those models against brain responses when humans engage both the dorsal and ventral streams to freely explore complex and dynamic visual environments, which may be simulated in experimental settings [42].

To meet this need, we have developed a dual-stream model to mimic the parallel ventral and dorsal streams in the human brain [54, 50, 57, 26]. The model includes two branches of convolutional neural networks: WhereCNN and WhatCNN, which share the same architecture but receive distinct visual inputs and generate different outputs. WhereCNN samples a wide view to learn spatial attention and where to direct the subsequent gaze, while WhatCNN samples a narrow view to learn object representations. By taking multiple gazes at a given scene, the model sequentially samples the salient locations and progressively constructs a scene representation over both space and time. To evaluate this dual-stream model as a model of the human visual system, we have tested its ability to reproduce human gaze behavior and predict functional brain scans from humans watching a movie with unconstrained eye movements. Our hypothesis is that the model's WhereCNN and WhatCNN branches can effectively predict the brain responses along the brain's dorsal and ventral visual pathways, respectively. In addition, we have also conducted experiments to evaluate the underlying factors contributing to the functional segregation of the brain's dorsal and ventral visual streams. Of particular interest were the relative contributions of retinal sampling, spatial attention, and attention-guided eye movement in shaping the function of the dorsal stream and its interplay with the ventral stream during dynamic natural vision.

## 2 Related Works

### 2.1 Dorsal-stream vision

Image-computable models of the brain's dorsal stream have been relatively limited compared to models of the ventral stream. Previous work has attempted to model the dorsal stream by training deep neural networks to detect motion [58] or classify actions [29] using video inputs. However, these models do not fully capture the neuroscientific understanding that the dorsal stream is involved in locating objects and guiding actions, leading to its designation as the "where" or "how" visual pathway. More recent work by Mineault et al. focused on training a dorsal-stream model to emulate human head movements during visual exploration [53]. Additionally, Bakhtiari et al. utilized predictive learning to train parallel pathways and observed the ventral-like and dorsal-like representations as an emergent consequence of structural segregation [3]. However, no prior work has explored neural network models that emulate how the dorsal stream learns spatial attention and guides eye movements for visual navigation.

### 2.2 Spatial attention and eye movement

Prior research in the field of computer vision has attempted to train models to attend to and selectively focus on salient objects within a scene [55, 61, 22], rather than processing the entire scene as a whole. This approach aligns with the brain's mechanism of spatial attention, where the dorsal stream acts as a global navigator, and the ventral stream functions as a local perceiver. In line with this mechanism, previous studies have employed dual-stream neural networks that process global and local features in parallel, aiming to achieve enhanced computational efficiency as a unified system [21, 61, 69, 30, 72]. However, these models do not fully replicate the way human eyes sample visual inputs during active exploration of the scene and thus still fall short in biological relevance.

### 2.3 Foveated vision and retinal transformation

The human retina functions as a sophisticated camera that intelligently samples and transmits visual information. It exhibits the highest visual acuity in the central region of the visual field, a phenomenon referred to as foveated vision [18, 12, 14]. In contrast, peripheral vision uses lower spatial acuity but higher temporal sensitivity, making it better suited for detecting motion. These properties of retinal sampling are potentially useful for training neural networks to enhance performance [52, 65, 40, 2] or robustness [32, 66, 8], augment data or synthesize images [68]. The retina also transmits information to the brain using distinct types of cells. The magnocellular and parvocellular retinal ganglion cells have different distributions, selectivity, and relay information through largely separate pathways. Taken together, the retina transforms visual information into segregated inputs for parallel visual processing. This biological mechanism has not been systematically investigated.

Unlike the above prior works, our work combines multiple biologically inspired mechanisms into an integral learnable model. It uses a frontend inspired by the human retina and applies complementary retinal sampling and transformation. It uses two parallel pathways inspired by the human dorsal and ventral streams. It uses spatial attention and object recognition as the distinct learning objectives for training the two pathways. It further uses the attention to move fixations and thus allows the two pathways to interact for active sensing. Although these mechanisms have been explored separately in prior studies, their combination is novel and motivates our work to build and test such a model against the human brain and behavior in a naturalistic condition of freely watching a movie.

## 3 Methods

In our model, WhereCNN and WhatCNN serve distinct functions in processing objects within a scene. WhereCNN identifies the spatial location of an object, determining "where" it is situated, while WhatCNN focuses on recognizing the identity of the object, determining "what" it is. When multiple objects are present in a scene, WhereCNN learns spatial attention and "how" to sequentially locate and fixate on each object. This allows the model to selectively attend to different objects in a sequence, mirroring the dynamic nature of human eye movements during visual exploration. Fig.1 illustrates and describes the human visual system that inspires us to design our model.

### 3.1 Model design and training

Akin to the human eyes [7, 23], our model uses retinal transformation [4] to generate separate inputs to WhereCNN and WhatCNN. For both, the retinal input consists of $64 \times 64$ samples non-uniformly distributed around the fixation. When describing a point in the retinal image and its corresponding point in the visual world in terms of the radial distance and the polar angle with respect to the fixation, their polar angles are the same while their radial distances are related by Eq. 1.

$$r = g(r') = \frac{b}{\sqrt{\pi}} \frac{1 - \exp(\ln(a) r'/2)}{1 - \exp(\ln(a)/2)} \tag{1}$$

where $r'$ and $r$ are the radial distances in the retinal and original images, respectively, $b$ is a constant that ensures $r_{\max}/g(r'_{\max}) = 1$, and $a$ controls the degree of center-concentration. Given a larger $a$, more retinal samples are closer to the fovea relatively to the periphery. We set $a = 15$ for WhatCNN and $a = 2.5$ for WhereCNN. In this setting, WhereCNN is more selective to global features, while WhatCNN is more selective to local features, mirroring the sampling bias of magnocellular and parvocellular retinal ganglion cells, as illustrated in Fig.1.

Both WhereCNN and WhatCNN use similar backbone architectures. The backbone consists of four blocks of convolutional layers. Each block includes two `Conv2D` layers (kernel size $3 \times 3$) followed by `ReLU` and `BatchNorm`. Applying $2 \times 2$ `MaxPool` between adjacent blocks progressively reduces the spatial dimension. The feature dimension following each block is $64$, $128$, $256$, or $512$. Atop this backbone CNN, both WhereCNN and WhatCNN use additional components to support different goals. For WhereCNN, the feature maps from the 3rd and 4th convolutional blocks are resized to $16 \times 16$ and concatenated, providing the input to an additional convolutional block. Its output feature map is subject to `SoftMax` to generate a probability map of visual saliency. By random sampling by the probability of saliency, WhereCNN decides a single location for the next fixation. To avoid

future fixations to revisit previously attended areas, inhibition of return (IOR) [37] is used. IOR keeps records of locations of prior visits as defined in Eq.2.

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

where $G(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a 2D Gaussian function centered at $\boldsymbol{l}_\tau$ (prior fixations) with a standard deviation $\sigma$ at the $\tau$-th step. Its values are normalized so that its maximum equals 1. By applying the IOR to the predicted saliency map using an element-wise multiplication, the future fixations would not revisit the areas already explored.

For WhatCNN, the output feature map from the 4th convolutional block, after global average pooling, is given as the input to an additional layer of Gated Recurrent Units (GRU) [9], which recurrently update the representation from a sequence of fixations to construct a cumulative representation following another fully-connected layer.

We first pre-train each stream separately and then fine-tune them together through three stages.

**Stage 1 - WhereCNN**. We first train WhereCNN for image recognition using ILSVRC2012 [60] for object recognition and then fine-tune it for generating the saliency map to match human attention using SALICON dataset [39]. In this stage, we use random fixations to generate the retinal inputs to WhereCNN. Adam optimizer [43] (lr=$0.002$, $\beta_1$=0.9, $\beta_2$=0.99) is used with $25$ epochs for SALICON training. At this stage, WhereCNN learns spatial attention from humans.

**Stage 2 - WhatCNN**. We first train WhatCNN for single-object recognition using ILSVRC2012 [60] and then fine-tune it for multi-object recognition using MSCOCO [47]. In this stage, we use the pre-trained WhereCNN to generate a sequence of eight fixations and accordingly apply the retinal transformation to generate a sequence of retinal inputs to WhatCNN for recurrent object recognition. Note that the training in Stage 2 is confined to WhatCNN, while leaving WhereCNN as pre-trained in Stage 1. Adam optimizer (lr=$0.002$, $\beta_1$=0.9, $\beta_2$=0.99) is used with $40$ epochs for MSCOCO training.

**Stage 3 - WhereCNN & WhatCNN** Lastly, we equally combine the two learning objectives to train both WhereCNN and WhatCNN altogether and end-to-end using eight fixations. Adam optimizer (lr=$0.0002$, $\beta_1$=0.9, $\beta_2$=0.99) is used with $25$ epochs for training. In this stage, SALICON, which contain labels for both saliency prediction and object recognition, is used for training. More details can be found at Appendix B.

### 3.2 Model evaluation with human gaze behavior and fMRI responses

We use two criteria to evaluate how well a model matches the brain given naturalistic and dynamic visual stimuli. First, the model should generate similar human visual behaviors, such as visual perception and gaze behavior. Second, the model's internal responses to the stimuli should predict the brain's responses to the same stimuli through linear projection implemented as linear encoding models [56]. For our dual-stream model, we hypothesize that WhereCNN better predicts dorsal-stream voxels and WhatCNN better predicts ventral-stream voxels.

For this purpose, we use a publicly available fMRI dataset from a prior study [34], in which a total of 11 human subjects ($4$ females) were instructed to watch the movie Raiders of the Lost Ark ($115$ minutes) with unconstrained eye movements. The movie was displayed on an LCD projector with a visual angle of $17^\circ \times 22.7^\circ$. Whole-brain fMRI data was acquired in a 3-T MRI system with a gradient-recalled echo planar imaging sequence (TR/TE = $2.5$s/$35$ms, flip angle = $90^\circ$, nominal resolution = $3\text{mm} \times 3\text{mm} \times 3\text{mm}$). We preprocess the data by using the minimal preprocessing pipeline released by the Human Connectome Project (HCP) [25]

We test how well the model can predict the voxel-wise fMRI response to the movie stimuli through a learnable linear projection of artificial units in the model. To evaluate whether and how the two branches in the model differentially predict the two streams in the brain, we define two encoding models for each voxel: one based on WhereCNN and the other based on WhatCNN. We train and test the encoding models with data during different segments of the movie. To avoid overfitting, we apply dimension reduction to the internal responses in either WhereCNN or WhatCNN by applying principal component analysis (PCA) first to each layer and then to all layers while retaining $99\%$ of the variance [71, 31]. We further convolve the resulting principal components with a canonical

hemodynamic response function (HRF) that peaks at 5 seconds, down-sample them to match the sampling rate of fMRI, generating the linear regressors used in the encoding model. Using the training data (81% of the total data), we estimate the encoding parameters using L2-regularized least squares estimation. Using the held-out testing data (19%), we test the encoding models for their ability predicting the fMRI responses observed at each voxel and measure the accuracy of prediction as the correlation between the predicted and measured fMRI responses, denoted as $r_{where}$ and $r_{what}$ for the encoding models based on WhereCNN and WhatCNN. We test the significance of the prediction using a block permutation test [1] with a block size of 20-seconds and $100,000$ permutations and apply the false discovery rate (FDR) ($p < 0.05$). We further differentiate the relative roles of the brain's WhereCNN vs. WhatCNN in predicting the brain's dorsal and ventral streams for single voxels as well as regions of interest. For this, we define a relative performance (Eq.3).

$$p_{where} = \frac{r^2_{where}}{r^2_{where} + r^2_{what}} \tag{3}$$

In the range from 0 to 1, $p_{where} > 0.5$ indicates better predictive performance by WhereCNN, while $p_{where} < 0.5$ indicates better predictive performance by WhatCNN.

### 3.3 Alternative models and control experiments

By design, the WhereCNN and WhatCNN branches within our model exhibit two key distinctions. WhereCNN is specifically trained to learn spatial attention by utilizing wider views, while WhatCNN focuses on object recognition through the use of local views. To explore the impact of input views and learning objectives on the model's capacity to predict brain responses, we introduce ControlCNN, which is a hybrid stream that learns to predict human attention (like WhereCNN) but uses retinal samples from local views (like WhatCNN). By replacing either WhereCNN or WhatCNN with ControlCNN, we create two alternative dual-stream models (illustrated in Fig.4) and examine their abilities to explain the functional segregation of the brain's dorsal and ventral streams.

## 4 Results

### 4.1 WhereCNN learns attention and WhatCNN learns perception

The WhereCNN and WhatCNN branches in our model are specifically designed to fulfill different objectives: predicting human visual saliency and recognizing visual objects, respectively. In Fig.2, we present examples comparing human attention with the model's attention based on the SALICON's validation set. WhereCNN can successfully identify salient locations where humans are more likely to direct gaze. WhereCNN can mimic human saccadic eye movements by generating a sequence of fixations that navigate the model's attention to those salient locations. In contrast, WhatCNN can recognize either single or multiple objects (macro F1 score on MSCOCO's validation set: 61.0).

### 4.2 WhereCNN and WhatCNN matches dorsal and ventral visual streams

By using linear encoding models, we use the WhereCNN and WhatCNN branches to predict fMRI responses during the processing of identical movie stimuli by both the model and the brain. Together, these two branches can predict responses across a wide range of cortical locations involved in visual processing. However, they exhibit distinct predictive power in relation to the dorsal and ventral streams. Generally, the WhereCNN branch exhibits superior predictive performance for the dorsal stream, while the WhatCNN branch performs better in predicting responses within the ventral stream (Fig.3). In early visual areas (V1, V2, V3), WhereCNN better predicts the peripheral representations, while WhatCNN better predicts the foveal representations.

### 4.3 Factors underlying the functional segregation of the dorsal and ventral streams

We further investigate the underlying factors contributing to the model's ability to explain the functional segregation of the brain's dorsal and ventral visual streams (as depicted in Fig.3). Specifically, we examine the input sampling pattern and output learning objective, both of which are distinct for the WhereCNN and WhatCNN branches in our dual-stream model.


Figure 2: **Saliency prediction**. Given an image (1st row), WhereCNN generates a saliency map (3rd row) similar to the map of human attention (2nd row). Sampling this saliency map generates a sequence of fixations (as red/orange/blue circles in the order of time in the 4th row) similar to human saccadic eye movements (not shown).

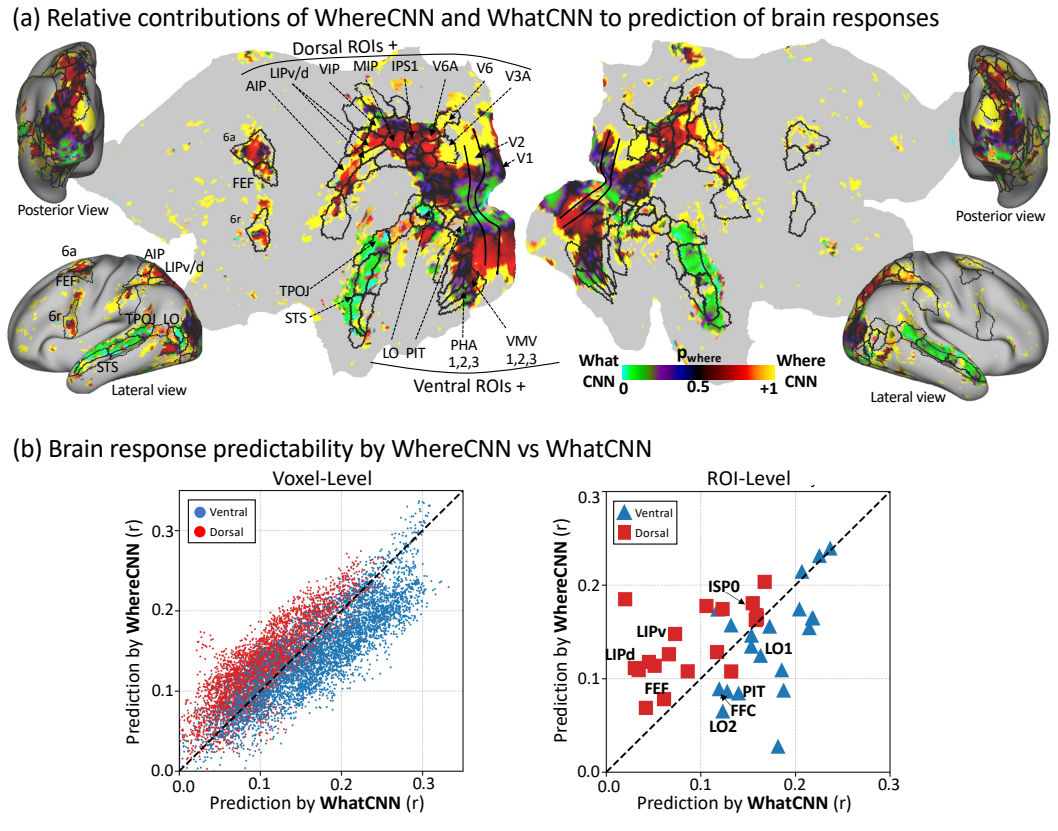


Figure 3: **Differential encoding of the dorsal and ventral streams**. (a) Relative contributions of WhereCNN and WhatCNN to the prediction of the fMRI response observed at each cortical location. Color highlights the locations significantly predictable by the model (FDR<0.05, block permutation test). The color itself indicates the degree by which WhereCNN is more predictive than WhatCNN (warm-tone) or the opposite (cool-tone). Visual areas are delineated and labeled based on brain altas [24]. Panel (b) plots the predictive performance by WhereCNN (y-axis) against that by WhatCNN (x-axis) and shows a clear separation of voxels (left panel) or ROIs (right panel) along the dorsal stream (red) vs. ventral stream (blue) relative to the dashed line of equal predictability. See Appendix A for the full ROI labels.

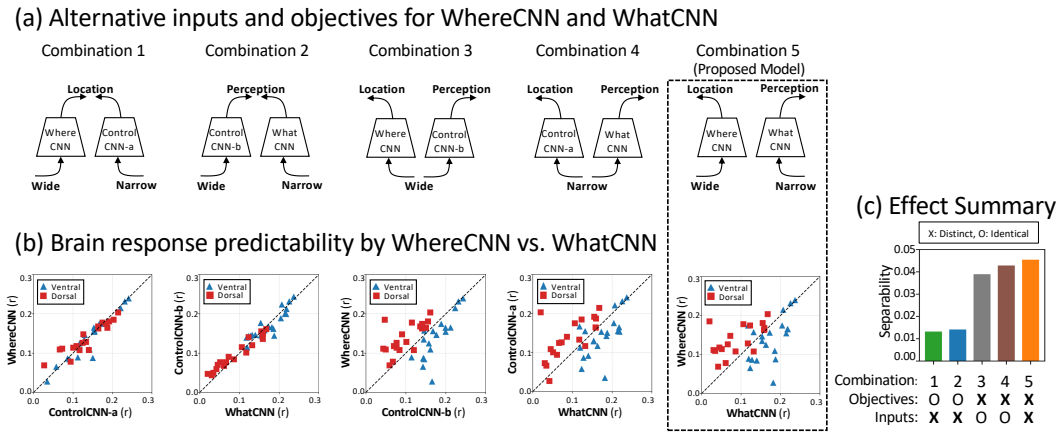


Figure 4: **Contributing factors of the dorsal-ventral functional segmentation.** (a) Alternative designs of the two-stream model for investigating the contributing factors of the functional segregation of two streams in the proposed model. In addition to WhereCNN and WhatCNN in the proposed model, we included ControlCNN-a (wide input field of view for predicting location) and ControlCNN-b (narrow input field of view for predicting perception) for an ablation study. (b) The predictive performances and functional segregations by the two streams are plotted for the dorsal (red squares) vs. ventral (blue triangles) ROIs for each of the alternative models in (a), correspondingly. The dashed line represents equal predictive abilities of ventral and dorsal ROIs. (c) Quantitative evaluation of the functional segregation of the dorsal and ventral ROIs relative to the dashed line of equal predictability. Separability measures how far the predictions are away from the dashed line. Assuming the coordinates of a certain ROI is $(x, y)$, separability is calculated as the average of $|x - y|$ for all ROIs.

To investigate the contributing factors of the functional segregation in predicting human ventral and dorsal streams, we introduce four variations of the proposed model, where the two branches either share their inputs or have the same learning objectives, and compare their capacity to account for the functional segregation of the dorsal and ventral visual streams (as shown in Fig. 4). When the two branches solely differ in their input sampling, they are unable to explain the dorsal-ventral segregation (the first two models from the left). However, when the two branches exclusively differ in their learning objectives, the functional segregation is better explained (third and fourth models from the left). Moreover, when the two branches differ in both input sampling and learning objectives (rightmost model), as utilized in our proposed model, the functional segregation is even more pronounced and elucidated. These ablation experiments suggest that the distinct learning objectives of the brain's dorsal and ventral streams is the primary factor underlying their functional segregation.

### 4.4 Dual-stream: a better brain model than single-stream

We also compare our dual-stream model with single-stream alternatives. One of these alternatives is a baseline CNN that shares the same backbone architecture as a single branch in our dual-stream model. However, this baseline CNN is trained with original (224x224) images to recognize objects in ImageNet [60] and MS-COCO [47]. Thus, it serves as a direct comparison with either the WhereCNN or WhatCNN branch in our model. In addition, we also include AlexNet [45], ResNet18, and ResNet34 [35] as additional alternatives, which have been previously evaluated in relation to brain responses [71, 70]. We compare these single-stream alternatives with either branch in our model in terms of their ability to predict brain responses within dorsal or ventral visual areas (Fig.5). Despite their use of the same backbone architecture, the baseline under-performs WhatCNN in predicting responses in ventral visual areas, and under-performs WhereCNN in predicting responses in dorsal visual areas. This result suggests that the interactive and parallel nature of the dual-stream model renders each stream more akin to the functioning of the human brain, surpassing the performance of isolating a single stream. Moreover, WhatCNN or WhereCNN also performs better than AlexNet and comparably with ResNet18 and ResNet34, which are deeper than the architecture of our model.

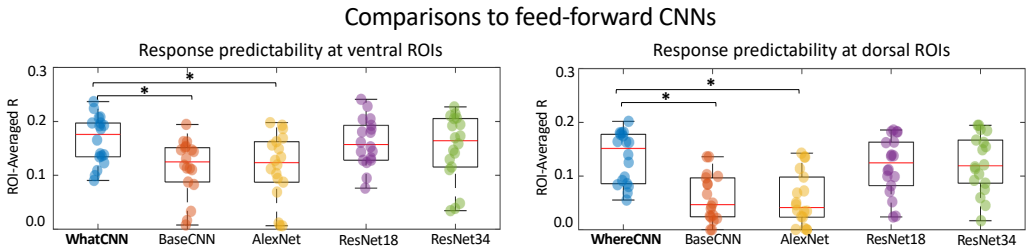


Figure 5: WhatCNN (left) or WhereCNN (right) vs. alternative single-stream CNNs. The boxplot shows the encoding performance of different models for ventral or dorsal visual areas. Each dot within the box plot signifies the average prediction accuracy r within a respective ROI in the ventral or dorsal region. Asterisk (*) represents a significant difference by the Wilcoxon signed-rank test ($\alpha = 0.05$).

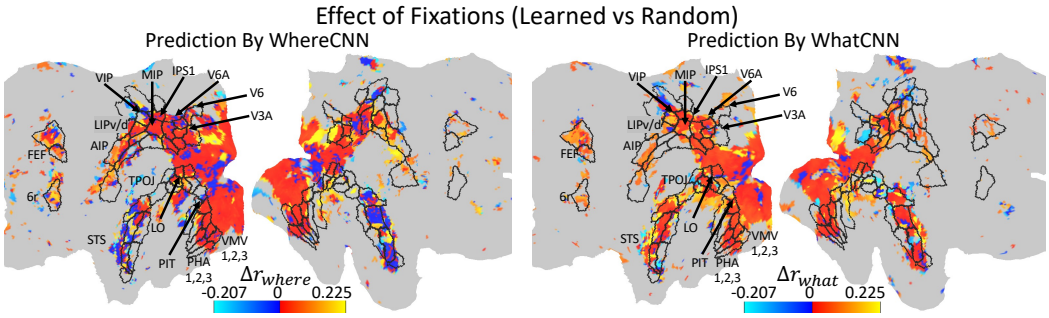


Figure 6: **Effects of attention-driven eye movements.** The use of attention to determine fixations vs. the use of random fixations is evaluated in terms of the resulting difference in the encoding performance by WhereCNN (left) and WhatCNN (right), denoted and color-coded as $\Delta r_{where}$ and $\Delta r_{what}$, respectively.

### 4.5 Attention-driven eye movements improve encoding

Similar to human gaze behavior towards salient objects, our model learns spatial attention to guide fixations for parallel visual processing. In this study, we investigate whether and how the model's ability to predict brain responses depends on its utilization of attention-driven fixations. To examine this, we conduct experiments where the model is allowed to use either attention-driven fixations or random fixations to collect retinal samples, and we evaluate how this choice impacts the model's capability to predict brain responses. As depicted in Fig.6, employing attention-driven fixations leads to higher encoding accuracy by both WhereCNN and WhatCNN compared to the use of random fixations for a majority of visual cortical locations within both the dorsal and ventral streams.

## 5 Discussion

In summary, we introduce a new dual-stream neural network that incorporates the brain's mechanisms for parallel visual processing. The defining features of our model include 1) using retinal transformation to separate complementary inputs to each stream, 2) using different learning objectives to train each stream to learn either spatial attention or object recognition, and 3) controlling sequential fixations for active and interactive visual sensing and processing. We demonstrate that the combination of these features renders the model more akin to the human brain and better predictive of brain responses in humans freely engaged in naturalistic visual environments. Importantly, the two streams in our model differentially explain the two streams in the brain, contributing to the computational understanding as to how and why the brain exhibits and organizes distinct responses and processes along the structurally segregated dorsal and ventral visual pathways. Our findings suggest that the primary factor contributing to the dorsal-ventral functional segregation is the different goals of the dorsal and ventral pathways. That is, the dorsal pathway learns spatial attention to control eye movements [6, 13, 75, 49], while the ventral stream learns object recognition.

Although our model demonstrates initial steps to model parallel visual processing in the brain, it has limitations that remain to be addressed in future studies. For one limitation, the model uses different spatial sampling to generate the retinal inputs to the two streams but does not consider different temporal sampling that makes the dorsal stream more sensitive to motion than the ventral stream [58, 29, 53, 3]. For another limitation, the interaction between the two streams is limited to the common fixation that determines the complementary retinal input to each stream. Although attention-driven eye movement is an important aspect of human visual behavior shaping brain responses for both dorsal and ventral streams, the two streams also interact and exchange information at higher levels. The precise mechanisms for dorsal-ventral interactions remain unclear but may be important to understanding human vision or improving brain-inspired computer vision.

## 6 Acknowledgements

## References


[1] Daniela Adolf, Snezhana Weston, Sebastian Baecke, Michael Luchtmann, Johannes Bernarding, and Siegfried Kropf. Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. *Frontiers in neuroinformatics*, 8:72, 2014.

[2] Emre Akbas and Miguel P Eckstein. Object detection through search with a foveated visual system. *PLoS computational biology*, 13(10):e1005743, 2017.

[3] Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, 34:25164–25178, 2021.

[4] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

[5] William Berrios and Arturo Deza. Joint rotational invariance and adversarial training of a dual-stream transformer yields state of the art brain-score for area v4. *arXiv preprint arXiv:2203.06649*, 2022.

[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] Alyssa A Brewer, William A Press, Nikos K Logothetis, and Brian A Wandell. Visual areas in macaque cortex measured using functional magnetic resonance imaging. *Journal of Neuroscience*, 22(23):10416–10426, 2002.

[8] Minkyu Choi, Yizhen Zhang, Kuan Han, Xiaokai Wang, and Zhongming Liu. Human eyes inspired recurrent neural networks are more robust against adversarial noises. *arXiv preprint arXiv:2206.07282*, 2022.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[10] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016.

[11] Carol L Colby and Michael E Goldberg. Space and attention in parietal cortex. *Annual review of neuroscience*, 22(1):319–349, 1999.

[12] Michael Connolly and David Van Essen. The representation of the visual field in parvicellular and magnocellular layers of the lateral geniculate nucleus in the macaque monkey. *Journal of Comparative Neurology*, 226(4):544–564, 1984.

[13] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

[14] Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990.

[15] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

[16] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[17] Gustavo Deco and Edmund T Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision research*, 44(6):621–642, 2004.

[18] AM Derrington and P Lennie. Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *The Journal of physiology*, 357(1):219–240, 1984.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

[21] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. *arXiv preprint arXiv:1709.01889*, 2017.

[22] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.

[23] Ricardo Gattass and Charles G Gross. Visual topography of striate projection zone (mt) in posterior superior temporal sulcus of the macaque. *Journal of neurophysiology*, 46(3):621–638, 1981.

[24] 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.

[25] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.

[26] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[27] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[28] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[29] Umut Güçlü and Marcel AJ van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, 2017.

[30] Yiyou Guo, Jinsheng Ji, Xiankai Lu, Hong Huo, Tao Fang, and Deren Li. Global-local attention network for aerial scene classification. *IEEE Access*, 7:67200–67212, 2019.

[31] Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

[32] Anne Harrington and Arturo Deza. Finding biological plausibility for adversarially robust features via metameric tasks. In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

[33] James V Haxby, Cheryl L Grady, Barry Horwitz, Leslie G Ungerleider, Mortimer Mishkin, Richard E Carson, Peter Herscovitch, Mark B Schapiro, and Stanley I Rapoport. Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences*, 88(5):1621–1625, 1991.

[34] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[36] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

[37] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[38] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[39] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[40] Aditya Jonnalagadda, William Yang Wang, BS Manjunath, and Miguel P Eckstein. Foveater: Foveated transformer for image classification. *arXiv preprint arXiv:2105.14173*, 2021.

[41] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[42] Hyun-Chul Kim, Sangsoo Jin, Sungman Jo, and Jong-Hwan Lee. A naturalistic viewing paradigm using 360 panoramic video clips and real-time field-of-view changes with eye-gaze tracking. *NeuroImage*, 216:116617, 2020.

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

[44] Talia Konkle and George Alvarez. Deepnets do not need category supervision to predict visual system responses to objects. *Journal of Vision*, 20(11):498–498, 2020.

[45] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[46] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[48] Margaret Livingstone and David Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240(4853):740–749, 1988.

[49] John HR Maunsell and Stefan Treue. Feature-based attention in visual cortex. *Trends in neurosciences*, 29(6):317–322, 2006.

[50] William H Merigan and John HR Maunsell. How parallel are the primate visual pathways? *Annual review of neuroscience*, 16(1):369–402, 1993.

[51] A David Milner. How do the two visual streams interact with each other? *Experimental brain research*, 235(5):1297–1308, 2017.

[52] Juhong Min, Yucheng Zhao, Chong Luo, and Minsu Cho. Peripheral vision transformer. *arXiv preprint arXiv:2206.06801*, 2022.

[53] Patrick Mineault, Shahab Bakhtiari, Blake Richards, and Christopher Pack. Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, 34:28757–28771, 2021.

[54] Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417, 1983.

[55] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014.

[56] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[57] Jonathan J Nassi and Edward M Callaway. Parallel processing strategies of the primate visual system. *Nature reviews neuroscience*, 10(5):360–372, 2009.

[58] Reuben Rideaux and Andrew E Welchman. But still it moves: static image statistics underlie how we see motion. *Journal of Neuroscience*, 40(12):2538–2552, 2020.

[59] Giacomo Rizzolatti and Massimo Matelli. Two different streams form the dorsal visual system: anatomy and functions. *Experimental brain research*, 153:146–157, 2003.

[60] 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.

[61] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014.

[62] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual review of vision science*, 5:399–426, 2019.

[63] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[64] Keiji Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1):109–139, 1996.

[65] Chittesh Thavamani, Mengtian Li, Nicolas Cebron, and Deva Ramanan. Fovea: Foveated image magnification for autonomous navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15539–15548, 2021.

[66] Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *Advances in Neural Information Processing Systems*, 33:2135–2146, 2020.

[67] Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*, pages 2022–09, 2022.

[68] Binxu Wang, David Mayo, Arturo Deza, Andrei Barbu, and Colin Conwell. On the use of cortical magnification and saccades as biological proxies for data augmentation. *arXiv preprint arXiv:2112.07173*, 2021.

[69] Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, and Gao Huang. Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. *Advances in Neural Information Processing Systems*, 33:2432–2444, 2020.

[70] Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Scientific reports*, 8(1):3752, 2018.

[71] 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.

[72] Kevin Wu, Eric Wu, and Gabriel Kreiman. Learning scene gist with convolutional neural networks to improve object recognition. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2018.

[73] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[74] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[75] Steven Yantis and John T Serences. Cortical mechanisms of space-based and object-based attentional control. *Current opinion in neurobiology*, 13(2):187–193, 2003.

[76] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

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

## 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 [9]. 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) [10, 11, 13, 3, 2]. 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 [8, 16, 4, 14]. 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 [7, 6, 1]. Fig. S1 displays the full set of region labels, corresponding to Fig.3(a) from the main text. Among the parcellations by [9], regions including significantly predicted voxels either by the WhereCNN or WhatCNN are presented in Fig. S1.

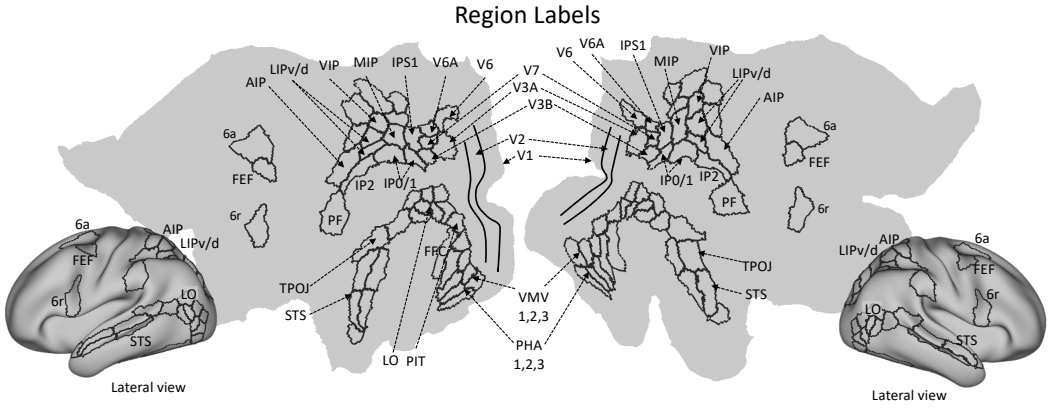


Figure S1: Region labels. Regions including significant voxels from Fig.3(a) in the main text are presented.

## 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 [15] for an image classification task over 120 epochs. A batch size of $1,024$ was employed, along with the Adam optimizer [12] (lr=$0.001$, $\beta_1$=0.9, $\beta_2$=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 [15] for an image classification task over 120 epochs, utilizing random fixations and the Adam optimizer (lr=$0.001$, $\beta_1$=0.9, $\beta_2$=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 [15] 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=$0.001$, $\beta_1$=0.9, $\beta_2$=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=$0.0001$, $\beta_1$=0.9, $\beta_2$=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 [18] 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, $\text{label}_c$ 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=$0.0002$, $\beta_1$=0.9, $\beta_2$=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. All codes and data will be made public.

## 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

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.

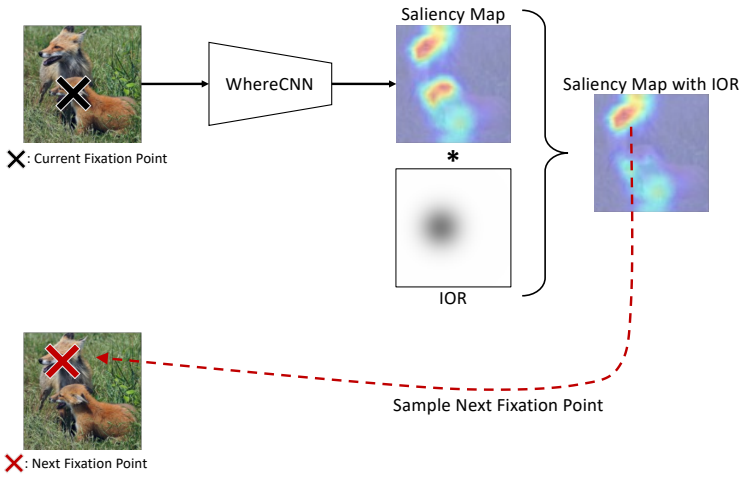


Figure S2: Process of determining the next fixation point. In IOR, white and black colors represent value 1 and 0, respectively.

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

where $G(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a 2D Gaussian function centered at $\boldsymbol{l}_\tau$ (prior fixations) with a standard deviation $\sigma$ at the $\tau$-th step. The Inhibition of Return (IOR) is initially created at a resolution of $224 \times 224$ 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.

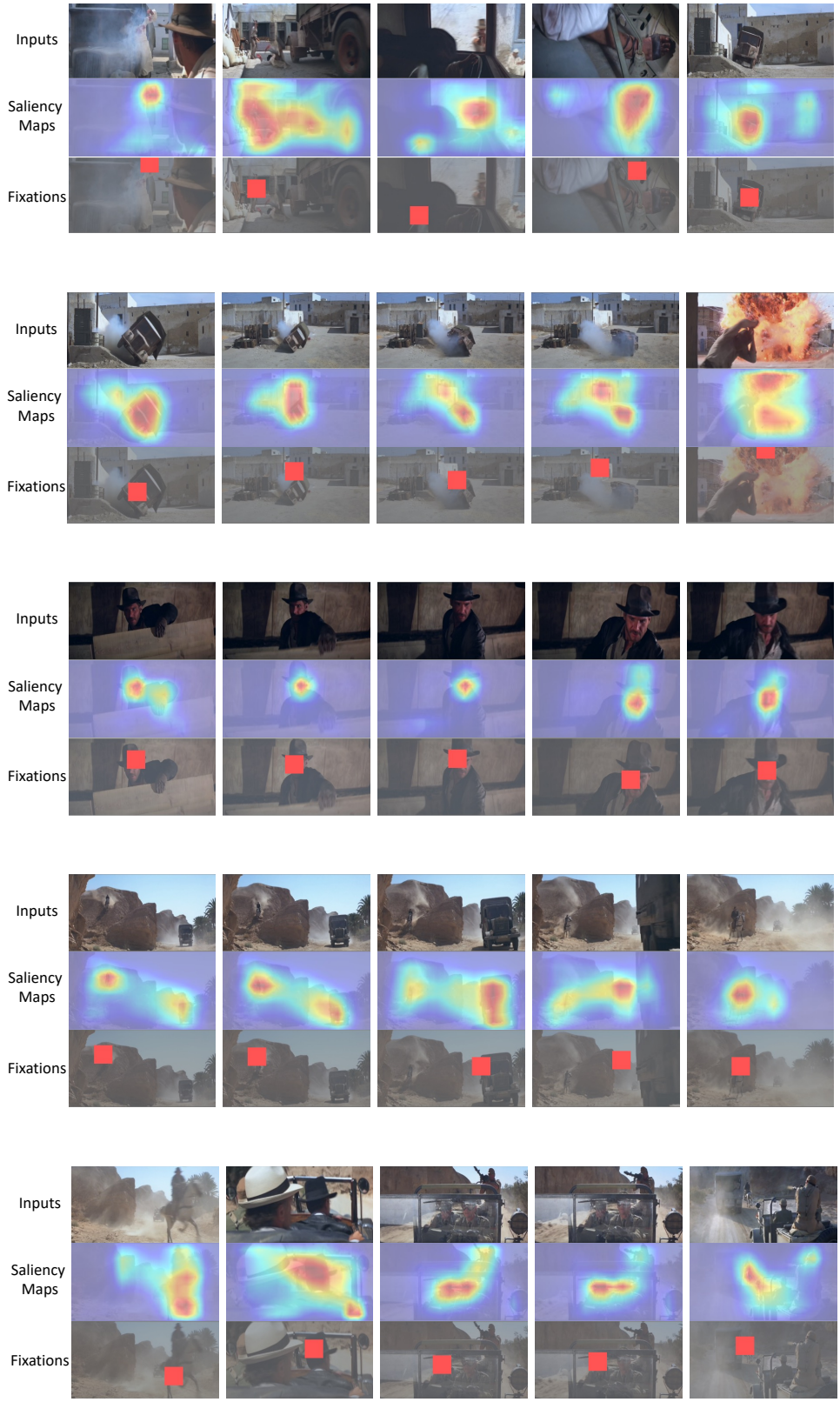


Figure S3: Given the movie frames (1st row), the WhereCNN generates saliency maps (2nd row) and fixations (3rd row). The red marker in the 3rd row presents the fixation point.

## 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 [5, 17]. 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.

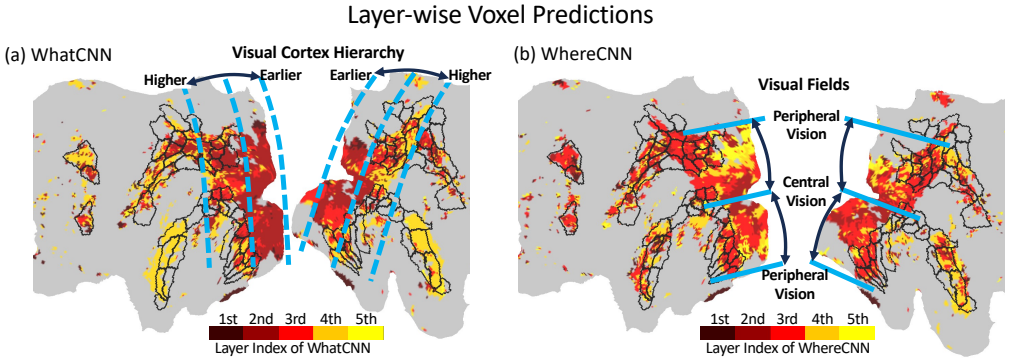


Figure S4: Each voxel is predicted by the features from a single layer from (a) WhatCNN and (b) WhereCNN. Layer indexes are color-coded so that the layer best predicting each voxel is presented.

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

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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