Human Eyes–Inspired Recurrent Neural Networks Are More Robust Against Adversarial Noises

Minkyu Choi
cminkyu@umich.edu
Department of Electrical Engineering and Computer Science, University of Michigan,
Ann Arbor, MI 48105, U.S.A.

Yizhen Zhang
yizhen.zhang@ucsf.edu
Department of Neurological Surgery, University of California, San Francisco,
CA 94143, U.S.A.

Kuan Han
kuanhan@umich.edu
Department of Electrical Engineering and Computer Science, University of Michigan,
Ann Arbor, MI 48105, U.S.A.

Xiaokai Wang
xiaokaiw@umich.edu
Department of Biomedical Engineering, University of Michigan, Ann Arbor,
MI 48109, U.S.A.

Zhongming Liu
zmliu@umich.edu
Department of Biomedical Engineering, University of Michigan, Ann Arbor,
MI 48109, U.S.A., and Department of Electrical Engineering and Computer Science,
University of Michigan, Ann Arbor, MI 48105, U.S.A.

Humans actively observe the visual surroundings by focusing on salient objects and ignoring trivial details. However, computer vision models based on convolutional neural networks (CNN) often analyze visual input all at once through a single feedforward pass. In this study, we designed a dual-stream vision model inspired by the human brain. This model features retina-like input layers and includes two streams: one determining the next point of focus (the fixation), while the other interprets the visuals surrounding the fixation. Trained on image recognition, this model examines an image through a sequence of fixations, each time focusing on different parts, thereby progressively building a representation of the image. We evaluated this model against various benchmarks in terms of object recognition, gaze behavior, and adversarial robustness.

Our findings suggest that the model can attend and gaze in ways similar to humans without being explicitly trained to mimic human attention and that the model can enhance robustness against adversarial attacks due to its retinal sampling and recurrent processing. In particular, the model can correct its perceptual errors by taking more glances, setting itself apart from all feedforward-only models. In conclusion, the interactions of retinal sampling, eye movement, and recurrent dynamics are important to human-like visual exploration and inference.

1 Introduction

CNNs and brains are different in both front and back ends. In the front end, they use an input layer to evenly sample a grid of pixels. The input size and computational complexity increase quadratically with the dimension of the visual field. In contrast, the brain uses the retina to obtain nonuniform and fisheye-style samples distributed around where the eyes are fixated (Bashivan et al., 2019; Curcio et al., 1990; Watson, 2014). Retinal sampling is much denser and more accurate in the fovea than in the periphery, as the size of the retina's receptive fields increases proportionally with eccentricity, becoming larger as they move away from the fovea (Connolly & Van Essen, 1984; Curcio et al., 1990; Derrington & Lennie, 1984). This nonlinear sampling density allows the brain to keep up with a very wide visual field despite restricted memory and computation (Gattass et al., 1981, 1988). In the back end, CNNs process the input pixels all at once through a single feedforward pass and arrive at a single perceptual decision per each image input. In contrast, humans may take multiple glances through eye movements and use dual visual streams to guide gaze behavior and support dynamic perception. The magnocellular subcortical pathway and its extension onto the dorsal cortical pathway direct where to look at each glance, while the parvocellular subcortical pathway and its extension onto the ventral cortical pathway transform the retina samples into abstract representations for recognition (Merigan & Maunsell, 1993; Mishkin et al., 1983). It is, however, unclear whether and how such front-end and back-end differences explain the gaze behavior and perceptual robustness of computer versus human vision.

Humans use spatial attention to drive eye movements. For rapid recognition, they tend to look at salient regions while disregarding less informative or insignificant parts. This is because the eyes collect more samples from the fovea such that downstream visual processing is also more dedicated to regions around the fixation. This bias in sampling and processing makes "where to look" an important decision that the brain has to make. To inform this decision, the brain uses a wide view enabled, including both central and peripheral vision, to locate the salient object and use this spatial information as overt attention to direct eye movement (Deubel & Schneider, 1996;

Wiecek et al., 2012). This process involves the brain's "where" pathway and the oculomotor control pathway (Colby & Goldberg, 1999; Corbetta & Shulman, 2002; Rizzolatti & Matelli, 2003).

Under normal viewing conditions, humans move their eyes to collect more information and refine visual perception. Human vision is robust, unlike computer vision. A small amount of adversarial noise can deceive computer vision but appear trivial to humans (Goodfellow et al., 2014). However, when humans have an extremely restricted time (e.g., 70 ms or less) to observe an image, the brain may also make perceptual mistakes given the same adversarial noises (Elsayed et al., 2018). In this case, humans can only afford a single glance while the brain is limited to its feedforward processing rather than feedback or recurrent processing (Lamme & Roelfsema, 2000). The brain's feedforward visual processing takes 150 ms to reach a perceptual decision (Thorpe et al., 1996) for rapid object recognition (DiCarlo et al., 2012), akin to the mode of operation in CNNs. Therefore, it is likely that eye movement, feedback, and recurrent processing are necessary for robust human vision. The retina may place a key role in adversarial robustness. Depending on the distance from the point of fixation, or eccentricity, the same object bears different retinal patterns when it is in the fovea versus periphery (Balas et al., 2009; Bouma, 1970; Lettvin, 1976; Rosenholtz, 2016; Stewart et al., 2020). This dynamic sampling acts as a form of data augmentation, possibly making visual perception less vulnerable to minor distortions or adversarial perturbations (Vuyyuru et al., 2020).

Therefore, the retina has profound impacts on human vision but is rarely explored in computer vision. In this study, we draw inspirations from human eyes, dorsal and ventral streams, and spatial attention to design a recurrent neural network for computer vision and evaluate the model's attention and adversarial robustness. The model has three modules: an input sampler that takes retinal samples around the fixation, an attention network that mimics the brain's dorsal visual pathway and guides where to look next, and a recognition network that mimics the ventral visual pathway and represents the retinal samples recurrently for object recognition. As illustrated in Figure 1, we design three variations of such a model. Each of them uses a different strategy for sampling visual input with respect to the point of fixation, including cropping image patches with a single field of view, with two fields of view, or applying retinal foveation and nonuniform sampling. These models all attempt to mimic the brain's ability to engage in saccadic eye movement and recurrent neural processing (Kar et al., 2019; Lamme & Roelfsema, 2000). We hypothesize that the model using eyelike, retinal transformation learns attention behavior that is closer to human attention. Additionally, the model with retinal transformation can progressively refine the perceptual decision and improve robustness against adversarial perturbations as it allocates more time to take additional glances at an image under attack, allowing for a more refined perceptual decision.
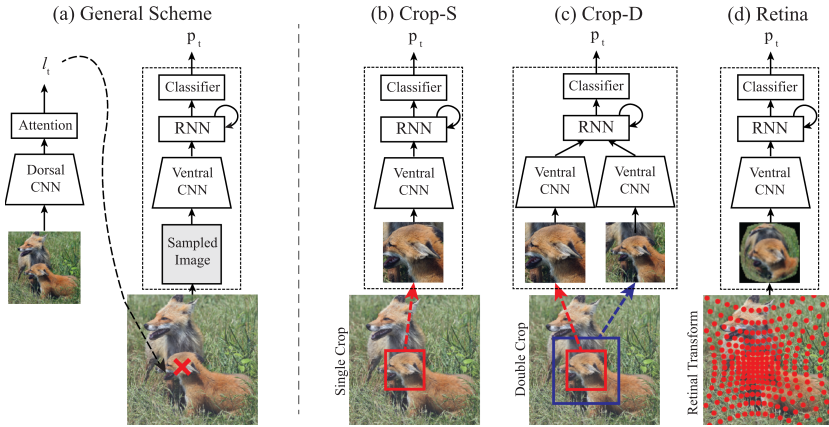
(a) General Scheme (b) Crop-S (c) Crop-D (d) Retina

p_t p_t p_t p_t

l_t

| Attention | | Classifier | | Classifier | | Classifier | | Classifier |

| Dorsal CNN | | RNN | | RNN | | RNN | | RNN |

| Ventral CNN | | Ventral CNN | | Ventral CNN | | Ventral CNN | | Ventral CNN |

| Sampled Image |

Single Crop Double Crop Retinal Transform

Figure 1: Recurrent models for eye movement and visual recognition. (a) The general two-stream model architecture: The model includes two streams: the dorsal (left) and ventral (right) streams. The dorsal stream has a broad field of view and generates a fixation point (l_t) at a step (or glance). The ventral stream takes selective samples around the fixation, extracts a representation, and accumulates the representations across multiple glances for object recognition (p_t). Using this general scheme, we design and test three implementations of the ventral stream: Crop-S, Crop-D, and Retina, illustrated in panels b to d. (b) Crop-S crops a small image region (shown as the red box) around the fixation. (c) Crop-D crops two regions (shown as the red and blue boxes) around the fixation and samples them with different resolutions such that the same number of samples is extracted from either region. (d) Retina applies retinal transformation and extracts nonuniform samples with respect to the fixation. Those models use the same architecture for the dorsal stream, for which the weights are learned separately alongside the different ventral stream models.

2 Related Work

2.1 Two-Stream Architecture. In the field of computer vision, focusing on distinct aspects of data, such as spatial and temporal features, often proves beneficial for solving specific tasks like object recognition and action recognition. To handle these distinct features, two or more parallel processing streams specialized for each feature can be employed.

In object recognition, Esteves et al. (2017), Guo et al. (2019), Sermanet et al. (2014), and Wang et al. (2020) use dual-stream architectures to support both global and local feature processing. In these works, one stream processes the entire image area, extracting global features that provide a broad understanding of the scene. Based on these global features, the models determine where to allocate more resources to obtain detailed local features. Subsequently, the other stream processes these local features, allowing for

more precise object recognition. By processing only the informative image areas, these models save computational costs while achieving comparable recognition accuracies to single-stream models with larger sizes. For video processing, dual-stream architectures specialized for temporal and spatial information processing demonstrate improved action recognition performance (Feichtenhofer et al., 2017; Simonyan & Zisserman, 2014; Wang et al., 2016). Choi et al. (2023) proposed two-stream architectures to model the dorsal and the ventral visual streams of human brains. The two streams are trained to perform distinct functions of saliency prediction and object classification. The results demonstrate that the two streams resemble the representations of the brain's dorsal and ventral visual streams, respectively.

Our study also employs a two-stream architecture for local and global feature processing. However, unlike previous works that simply crop patches from the given images (Guo et al., 2019; Sermanet et al., 2014; Wang et al., 2020), our model uses retina-like image sampling, enabling the model to effectively capture detailed local features while not ignoring the global context of the images. Our work shares similarities with the study by Choi et al. (2023) in terms of utilizing two streams and retinal transformation. However, a key distinction lies in our approach to modeling visual attention. Unlike Choi et al. (2023), who employ a human saliency data set to mimic human attention, our model is exclusively trained for object recognition without relying on any human saliency data set. Consequently, in our model, the emergence of human-like visual attention occurs as a natural by-product of the object recognition task rather than being directly trained to replicate human attention patterns.

2.2 Foveated Visions. In primate vision, the visual acuity is highest at the center of the gaze, a region called the fovea, due to the dense concentration of cone photoreceptors. This phenomenon, known as foveated vision, enables high-resolution and color perception in the central visual field. As the distance from the fovea increases, visual acuity decreases, reflecting the natural distribution of photoreceptors in the retina (Curcio & Allen, 1990; Curcio et al., 1990; Weber & Triesch, 2009). The concept of foveated vision has been extensively explored in the field of machine vision (Cheung et al., 2016; Deza & Konkle, 2020; Pramod et al., 2022). Researchers have investigated its applications across various domains, such as image rendering for virtual reality, predicting human scan paths, and improving object recognition performance.

Specifically, in virtual reality, low-resolution rendering of peripheral image regions based on foveated vision has been shown to significantly reduce rendering latency due to the reduced accuracy of human peripheral vision. Several studies have explored this approach and demonstrated its effectiveness in reducing computational requirements while maintaining the quality of the rendered image (Jabbireddy et al., 2022; Kaplanyan et al., 2019).

Furthermore, the foveated vision has been widely used in predicting human scan paths by analyzing eye movement patterns during visual perception. Researchers have developed models that can predict the order in which humans fixate on different regions of an image or scene, which can be useful in various applications such as advertising, design, and robotics (Bao & Chen, 2020; Berga & Otazu, 2020; Wang et al., 2017).

In addition to predicting scan paths, foveated vision has been shown to improve object recognition performance or reduce the computational costs of recognition systems. By using a high-resolution foveal region to focus processing resources on important areas of the image and a low-resolution peripheral region for less important areas, recognition systems can achieve higher accuracy with lower computational requirements. This approach has been explored in several studies and has shown promising results in improving recognition accuracy while reducing computational costs (Jonnalagadda et al., 2021; Min et al., 2022; Thavamani et al., 2021; Wang et al., 2021).

2.3 Recurrent Attention. A crucial difference between human and machine vision is that humans explore images or scenes by directing attention with eye movements (overt attention) or without eye movements (covert attention). Prior works has attempted to make machines solve downstream tasks by teaching them where to look through a recurrent process. Given an attention focus (or the fixation), the attended region can be cropped at various resolutions or scales (Mnih et al., 2014; Sermanet et al., 2014; Wang et al., 2020; Xu et al., 2015). Arguably better than hard cropping, the attended region may be subject to retinal transformation (Akbas & Eckstein, 2017; Bashivan et al., 2019; Vuyyuru et al., 2020) or polar transformation (Esteves et al., 2017) inspired by the primate retina (Coletta & Williams, 1987; Curcio et al., 1990; Geisler & Hamilton, 1986; Thibos et al., 1987). For covert attention, models generate and apply soft weightings for all features and locations depending on their relative importance (Jaegle et al., 2021; Xu et al., 2015; Zoran et al., 2020).

These recurrent attention models are similar to the human brain (Kietzmann et al., 2019). Because of this similarity, understanding recurrent attention models would provide insights into the mechanisms of adversarial robustness in human vision. However, despite their importance, recurrent attention models have rarely been the focus of research compared to their counterparts, feedforward CNNs. In the current work, we design and test different recurrent attention models and show how they are affected by adversarial noise as more recurrent steps are deployed.

2.4 Adversarial Attacks. An adversarial attack refers to an attempt to deceive machine learning models by adding carefully designed perturbations to an input image (Athalye et al., 2018; Carlini & Wagner, 2017; Goodfellow et al., 2014; Madry et al., 2017; Szegedy et al., 2013). The perturbation, known as adversarial noise, can be optimized with the projected

gradient descent (PGD) (Madry et al., 2017), the fast gradient sign method (Goodfellow et al., 2014) and Carlini and Wagner's (2017) attack, among others. To counteract adversarial noise, defensive methods have been proposed (Goodfellow et al., 2014; Gu & Rigazio, 2014; Papernot et al., 2016; Xie et al., 2017) but remain vulnerable (Athalye et al., 2018; Carlini & Wagner, 2017; He et al., 2017).

Unlike computer vision models, humans do not seem to suffer from the same vulnerability, especially when enough time is available for them to observe the image under attack (Elsayed et al., 2018). This distinction has motivated prior work to take inspiration from human vision for computer vision. For example, Luo et al. (2015) have found that allowing models to fixate on different image regions can alleviate the effect of adversarial noise. More recently, Dapello et al. (2020) show that attaching a block with properties of V1 at the front of feedforward CNNs can make CNNs more robust. Huang et al. (2020) demonstrate that a neural network with predictive coding improves adversarial robustness. Berrios and Deza (2022) show that their adversarially robust model achieves higher explainable variance for some brain areas compared to other biologically plausible neural networks, showing the possible link between the adversarially trained transformer and the explainability in the brain.

Vuyyuru et al. (2020) demonstrate that nonuniform spatial sampling and varying receptive fields that mimic the retinal transformation in the primate retina can also improve the robustness against adversarial attacks. Harrington and Deza (2021) show that the representation robust against adversarial noises is more attributable to processing information from the periphery, as opposed to the fovea.

In line with the related work, we also explore biologically inspired computational mechanisms for adversarial robustness. Similar to Vuyyuru et al. (2020), we use a retina-like front end to generate fixation-dependent retinal input for image recognition. A notable distinction is that Vuyyuru et al. (2020) prefix the points of fixation, whereas in our study, the fixation is adaptive and sequentially inferred via an attention module that learns where to look next. As such, our model iteratively and sequentially samples an image into time-varying retinal patterns, similar to retinal input to the brain during saccadic eye movement.

3 Method

In this section, we detail the methodology behind our approach, which integrates elements of human visual processing into our computer vision models. As illustrated in Figure 1a, the dorsal stream has a wide field of view that covers the whole image sampled with lower resolution. It learns spatial attention, predicts where to look next, and passes the predicted fixation (l_t) to the ventral stream. The ventral stream has a narrow field of view around the fixation, samples the input, learns to represent the samples for each

fixation, recurrently accumulates the representation across different fixations, and outputs the probability in image classification (p_t).

Specifically, the dorsal stream has two modules: `DorsalCNN` and `Attention`, and the ventral stream includes `VentralCNN`, `RNN`, and `Classifier`. Three model variations shown in Figures 1b to 1d all share the same architecture for the dorsal stream, but they use different sampling strategies for the ventral stream: a single-cropped field of view (Crop-S), double-cropped field of views with different resolution but the same matrix size (Crop-D), and retinal sampling (Retina). Further details of these configurations are provided below.

3.1 Image Sampling.

Single Crop: The model Crop-S crops a rectangular patch around the fixated location as the input to the ventral stream (see Figure 1b).

Double Crop: The model Crop-D processes two rectangular patches around the same fixation point, as shown in Figure 1c, with different scales and resolution. They are further resized to have the same patch size.

Retinal Sampling: We use retinal sampling with the knowledge of biological visual systems. In primate visual systems, both retinal ganglion cells and neurons at early visual areas have increasingly larger receptive fields yet lower resolution at higher eccentricity relative to where the eyes are fixated in the visual field (Freeman & Simoncelli, 2011; Gattass et al., 1981, 1988). In addition, more cells and neurons are devoted to the central vision than to peripheral vision (Bashivan et al., 2019; Curcio et al., 1990; Watson, 2014). Inspired by these properties, we design an input layer with two steps: foveated imaging and non-uniform sampling.

Foveated imaging (Duchowski et al., 2004; Perry & Geisler, 2002) varies image resolution and acuity along eccentricity: higher resolution around the fixation and progressively lower resolution toward the periphery. We implement this by applying a larger gaussian kernel to peripheral regions but a smaller kernel to the central region, resulting in greater smoothness in the periphery. More details are included in appendix A. Since each foveated image is transformed from the original image with a varying extent of spatial blurring, the feedforward convolutional layers effectively have eccentricity-dependent receptive fields. This effect is similar to how biases from different receptive fields in the retina are passed to downstream visual areas.

Retinal sampling collects nonuniform discrete samples from the foveated image. Figure 1d shows an example of retinal sampling points and the resulting retinal image. Suppose that we want to sample an $N \times N$ foveated image to fill an $n \times n$ retinal image with $n < N$. We first calculate
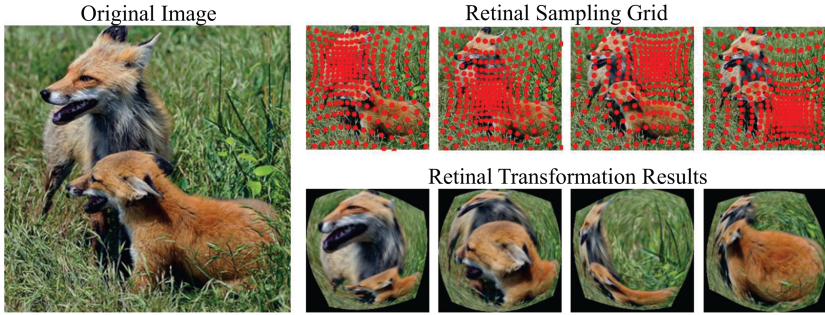
Figure 2: Examples of the retinal transformation. Left: Original image. Right: Retinal sampling grid and resulting retinal images.

the eccentricity $e(x, y; f_x, f_y)$ of any pixel location (x, y) with respect to the fixation point (f_x, f_y) in the foveated image (see equation 3.1). Similarly we calculate $\rho(i, j; f_i, f_j)$ as the distance from fixation (f_i, f_j) to pixel (i, j) (see equation 3.2) in the retinal image.

$$e(x, y; f_x, f_y) = \|(x, y) - (f_x, f_y)\|_2, \tag{3.1}$$

$$\rho(i, j; f_i, f_j) = \|(i, j) - (f_i, f_j)\|_2. \tag{3.2}$$

We then relate the eccentricity in the foveated image e to ρ of the retinal image through a nonlinear mapping function $g(\rho)$ (see equation 3.3):

$$e = g(\rho) = \frac{N \sinh(\rho \cdot b \frac{2}{N})}{2 \sinh(b \frac{n}{N})}. \tag{3.3}$$

Here, $b \in \mathbb{R}^+$ is a hyperparameter. Its value controls the degree of nonuniform sampling, which is set to 12.0 in our work. When there is no retinal sampling or b is very small, it can be considered that an image is under the fovea. Therefore, visual acuity is high everywhere sampled. As b becomes larger, the fovea area decreases and the periphery increases, while the retinal image is increasingly distorted relative to the original image (see Figures 2 and 5c for examples).

3.2 Dorsal Stream. In the preceding section, we explained how we implement retinal sampling to the visual input when a fixation point is present. Now, the next question is: How do we integrate an overt attention mechanism to determine where to look? Inspired by human vision, this process is facilitated by the dorsal stream in our proposed model, which is designed to generate a sequence of fixation points on the given images.
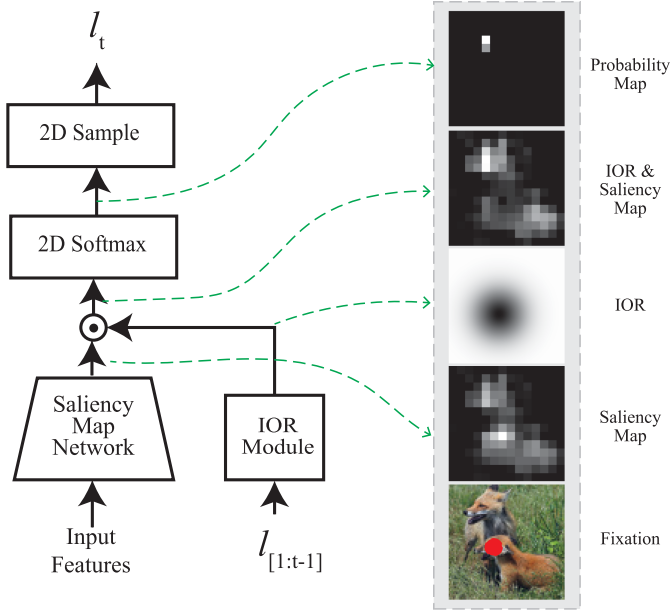
Figure 3: Details of the `Attention` module in the dorsal stream. `Attention` produces a 2D saliency map and fixation point at time step t (l_t). Intermediate representations from each module are shown on the right.

The dorsal stream includes `DorsalCNN` and `Attention` module. The input to the dorsal stream is an image with the full view but in a low resolution. The `DorsalCNN` is a stack of convolutional layers. `Attention` determines the next location to focus on. It uses the feature maps from all layers from `DorsalCNN` after they are resized to the same size and then concatenated along the channel dimension. Figure 3 shows the architecture of the attention module. To predict the next fixation point (l_t), the module maps salient regions with a `Saliency Map Network`. It includes a convolutional layer with 3×3 kernels and outputs a 2D saliency map (Itti et al., 1998; Koch & Ullman, 1987), highlighting the candidate regions for the next fixation.

To prevent future fixations returning to the previously attended regions, the inhibition-of-return (IOR) module (`IOR-Module`; Akbas & Eckstein, 2017; Itti & Koch, 2001; Najemnik & Geisler, 2005; Posner & Cohen, 1984) reduces the saliency of previously attended regions. Specifically, the IOR at time t is expressed as

$$\mathbf{IOR}(t) = \mathtt{IOR\text{-}Module}(l_{1:t}) = \mathbf{ReLU}\left(\mathbf{1} - \sum_{\tau=1}^{t} G(\boldsymbol{\mu} = l_\tau, \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I})\right). \quad (3.4)$$

Here, G is a normalized 2D gaussian function centered at l_τ with a standard deviation σ at the τth step. Its values are normalized so that the maximum equals 1. We sum the normalized gaussian kernels across all previous time steps to prohibit future fixations from going back to the previously visited locations. Subtracting the accumulated gaussian kernels from a matrix of ones creates a soft mask with high values at unattended regions but low values at previously attended regions. Applying this mask to the saliency map by element-wise multiplication followed by softmax gives rise to a 2D probability map based on which the next fixation is randomly sampled. Figure 3 shows typical examples of the saliency map, IOR, and the prediction of the next fixation point.

The dorsal stream is trained with reinforcement learning based on the `REINFORCE` algorithm (Williams, 1992). At time t, the fixation l_t generated by the attention module results in a new class prediction p_t by the recognition pathway (ventral stream). The reward r_t of choosing l_t as the fixation is calculated as the reduced classification loss relative to the previous time step $r_t = CE(p_{t-1}, \text{label}) - CE(p_t, \text{label})$, where CE is the cross-entropy loss. The goal of reinforcement learning 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 it is set as 0.8.

3.3 Ventral Stream.

The recognition pathway includes `VentralCNN`, `RNN`, and `Classifier` stacked as shown in Figure 1. `VentralCNN` consists of a stack of convolutional layers to extract features from the retinal samples at each time step. The extracted features initially go through global average pooling and then are channeled to a recurrent neural network (RNN) equipped with gated recurrent units (Chung et al., 2014). The `RNN` learns to accumulate information across different fixation points, which are then used for object recognition at the `Classifier` (a fully connected layer) at every time step t. The learning objective is to minimize the cross-entropy losses summed across T time steps.

3.4 Implementation Details.

Convolutional layers in `VentralCNN` and `DorsalCNN` have the same architecture (but distinct parameters), and they all have convolutional layers using 3×3 kernels with stride equal to 1. `VentralCNN` and `DorsalCNN` have `Conv(64)` $\times 2 \rightarrow$ `MaxPool` $\rightarrow$ `Conv(128)` $\times 2 \rightarrow$ `MaxPool` $\rightarrow$ `Conv(256)` $\times 2 \rightarrow$ `MaxPool` $\rightarrow$ `Conv(512)` $\times 2$, where `Conv(C)` $\times k$ represents k convolutional layers with C channels. `MaxPool` has a kernel size of 2 and a stride of 2. The sampled images for `VentralCNN` and `DorsalCNN` are in the size of 64×64 and 128×128, respectively. The dorsal and the ventral stream as shown in Figure 1 are trained together. For the Crop-D model, the same `VentralCNN` sharing the weights is used to extract features from both patches, and the resulting features are

summed to be forwarded to the next module. Code and data are publicly available.[1]

3.5 Training Details. We train these three two-stream models (Crop-S, Crop-D, and Retina) in three stages. First, we train models on single-label classification on ImageNet100. We randomly sample 100 classes from ILSVRC2012 (Deng et al., 2009) with 1000 classes in total to form ImageNet100 to save computational costs. Then the models are trained on multilabel classification tasks on MS-COCO (Lin et al., 2014). We include MS-COCO in the training process because of the complexity of the data set. Unlike ImageNet, images in MS-COCO include multiple objects, and the sizes of the objects are usually smaller than those of ImageNet, which is beneficial for training the dorsal stream for mimicking human attention. In the last stage, we fine-tune the models on ImageNet100 again. This stage is required to test our models on adversarial attacks because most of the adversarial attack algorithms are for single-label images.

In the first stage on ImageNet100, we train our models for 90 epochs using the Adam optimizer (Kingma & Ba, 2014) (lr $= 0.002$, $\beta_1 = 0.9$, $\beta_2 = 0.99$). The learning rate is decreased by 10 at the 30th, 60th, and 80th epochs. During training, models are allowed to deploy four fixations to explore images. In the second stage on MS-COCO, we train our models for 90 epochs using the Adam optimizer (lr $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.99$). The learning rate is decreased by 10 at the 30th, 60th, and 80th epochs. During training, models are allowed to deploy eight fixations to explore images. In the last stage on ImageNet100, we train our models for 20 epochs using the Adam optimizer (lr $= 0.0005$, $\beta_1 = 0.9$, $\beta_2 = 0.99$). The learning rate is decreased by 10 at the 10th epoch. During training, models are allowed to deploy four fixations to explore images.

4 Experiments

With the proposed models, we first test the performance of the models on single-label and multilabel classification tasks. Then we test how close the model attention is to human attention. At the same time, the effect of input sampling (retinal transformation, double crops, or single crop) on the generated attention is investigated. Finally, we test our models on the adversarial attacks to evaluate their adversarial robustness. As additional baseline models, we also trained a model with feedforward CNN (FF-CNN) and a model with recurrent attention but without overt eye movements (S3TA; Zoran et al., 2020). In the literature by Zoran et al. (2020), S3TA is trained with adversarial training. However, since our focus is on the model architecture instead of learning strategies, we use S3TA trained on

[1] https://github.com/minkyu-choi04/rs-rnn.

Table 1: Summary of the Models.

Model	ImageNet100 Top-1 Accs	MS-COCO F1 Scores	Input Type	Attention
Retina	76.6%	59.5	Retinal Image	Overt
Crop-D	81.2%	57.0	Double Crop	Overt
Crop-S	75.5%	47.6	Single Crop	Overt
S3TA	82.3%	58.8	Whole Image	Covert
FF-CNN	80.8%	57.0	Whole Image	–

single-label classification and multilabel classification tasks without adversarial training.

4.1 Model Performances on Classification Tasks. With the models trained on ImageNet100 and MS-COCO, we report their top-1 accuracy and F1 scores on the validation sets, which were not used for training the models. Table 1 includes the top-1 accuracy on ImageNet100's validation set and F1 scores for multi-label recognition on MS-COCO's validation set for all models. It also summarizes the attention properties of all models compared. For the ImageNet100 data set, models that receive full images (FF-CNN and S3TA) are generally better than the models with sampling images (Retina, Crop-D, and Crop-S), which may be attributable to the large size of the objects in the image. For image-sampling models (Retina, Crop-D, and Crop-S), a single glimpse may not capture the entire object. In contrast, FF-CNN models view the entire object, which contributes to their better performance. However, for MS-COCO, the model Retina's F1 score surpasses the other models, which implies that the complex image data sets with multiple small objects, such as MS-COCO, are more suitable for the models with human-like image sampling strategy, compared to simpler data sets with a single large object, such as ImageNet.

4.2 Model Attention Compared to Human Attention. In this batch of experiments, we quantitatively and qualitatively compare the generated attention maps with human attention. The proposed two-stream models, particularly the Retina model, successfully capture human-like attention without direct supervision from human saliency data. The models with explicit attention mechanisms (Retina, Crop-S, Crop-D, and S3TA) are trained to produce and use the attention maps for object classification tasks. Although FF-CNN lacks explicit attention mechanisms, we use class activation map (CAM, Zhou et al., 2016) to highlight image regions contributing to object label predictions. Figure 4 illustrates the attention maps from humans, Retina model, Crop-D, Crop-S, S3TA, and FF-CNN from the top row using the images from the validation set of SALICON (Jiang et al., 2015). In the visualized attention maps, the Retina model and Crop-D largely overlap with
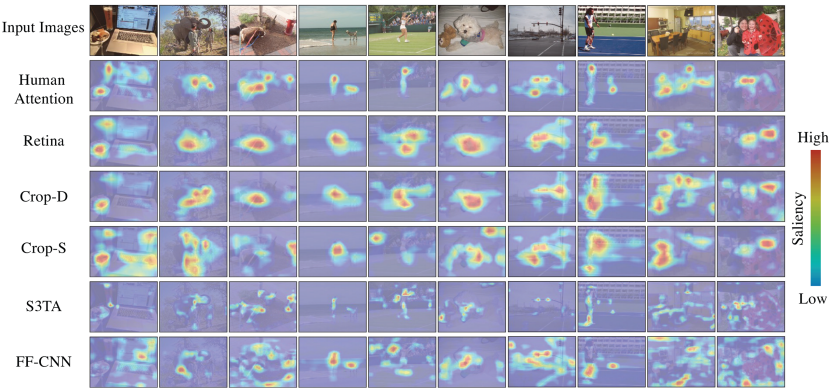
Figure 4: Saliency maps from top to bottom: humans, Retina, Crop-D, Crop-S, S3TA, and FF-CNN. Regions of higher saliency are highlighted in red, while areas of lower saliency are depicted in blue.

Table 2: Saliency Prediction Metrics on the SALICON Data Set.

	AUC ↑	NSS ↑	SIM ↑	CC ↑
Retina	**0.77± < 0.01**	**0.41± < 0.01**	**0.49± < 0.01**	**0.44± < 0.01**
Crop-D	0.75± < 0.01	0.38 ± 0.01	0.48 ± 0.01	0.41 ± 0.01
S3TA	0.72± < 0.01	0.29 ± 0.01	0.46± < 0.01	0.29± < 0.01
Crop-S	0.65 ± 0.02	0.23 ± 0.04	0.40 ± 0.02	0.26 ± 0.03
FFCNN	0.62± < 0.01	0.21 ± 0.01	0.40± < 0.01	0.20 ± 0.01

Note: Values in bold indicate the highest scores achieved across the models for each metric.

the human attention maps. Attention from S3TA is more selective and fragmented, but still overlaps with human attention. However, the attended areas from FF-CNN are fragmented and sometimes focus on nonessential parts of the images.

Quantitative evaluation of attention accuracy aligns with the visual observation. We report quantitative metrics of AUC (area under the curve), NSS (normalized scanpath saliency), SIM (similarity metric), CC (linear correlation coefficient) in Table 2 to measure how close the attention produced from the models is to the human attention. The result shows that the attention from the Retina model is the closest to the human attention, and Crop-D is the second closest compared to the other models. This implies that the input sampling schemes, with varying resolutions and resembling foveaperiphery, might be helpful in producing attention maps that are similar to human attention. On the other hand, FF-CNN produces the least humanlike attention maps.
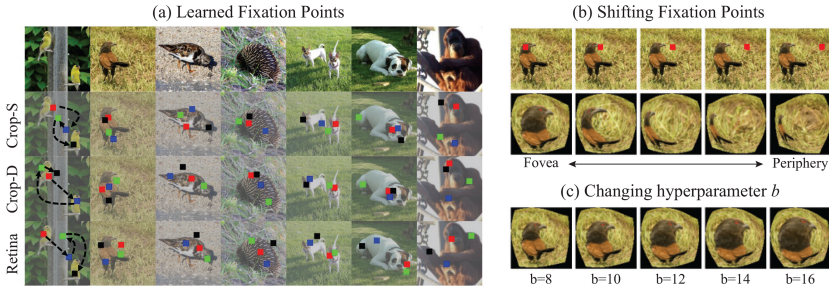
Figure 5: (a) Learned fixations from the first four steps. From the first fixation to the last fixations are marked as red-blue-green-black. The first column presents exemplar attention shifts. (b) Visualization of the retinal transformation as the object is shifted from the fovea to the periphery. (c) Examples of the retinal transformation as the hyperparameter b is changing from 8 to 16.

To validate the plausibility of the fixations, we visualize them in Figure 5a, where they are drawn from the attention maps generated by the models. As illustrated in Figure 3, fixations are sampled from the models' attention maps, and the inhibition of return (IOR) fosters models to explore new areas of images. The first four fixations are marked with red, blue, green, and black squares in order. Figure 5a demonstrates that the learned fixation points are placed on the foreground objects as opposed to focusing on background regions with no information.

4.3 Adversarial Attacks. Evaluating the robustness of computer vision models against adversarial attacks is crucial for understanding their resilience and reliability in real-world applications. Adversarial attacks aim to manipulate model predictions by introducing carefully crafted perturbations to the input images. These attacks can be broadly classified into two types: targeted and untargeted. In a targeted attack, the goal is to deceive an image-computable model, which correctly classifies an image, into labeling it as a different, attacker-specified class. In contrast, an untargeted attack strives to reduce the likelihood of the model recognizing the image as its original class, causing it to classify the image incorrectly.

We evaluate the adversarial robustness of the models trained on ImageNet100 using the attack success rate (ASR), which represents the ratio of successful attacks to total attack attempts. A model with a lower ASR is deemed more resistant to attacks. For attack algorithms, projected gradient descent (PGD) with L_∞ is used while varying the maximum perturbation budget allowed (ϵ). We iterate PGD 100 times with the step size at $\epsilon/20$. We also consider fast gradient signed method (FGSM) attack (Goodfellow et al., 2014) and SPSA attack (gradient-free; Uesato et al., 2018). All the sampling methods (retinal transformation and image crop) are fully differentiable.

Table 3: Attack Success Rates of the Models.

Model		Untargeted PGD				Targeted PGD			
	ϵ:	2e-3	3e-3	5e-3	7e-3	3e-3	5e-3	7e-3	1e-2
Retina		**28.2%**	**42.4%**	**65.2%**	**78.8%**	**4.7%**	**13.6%**	**31.7%**	**55.2%**
Crop-D		49.2%	73.2%	89.1%	96.1%	22.4%	56.6%	85.1%	96.1%
Crop-S		65.1%	81.1%	93.0%	97.8%	31.2%	70.7%	88.6%	95.9%
S3TA		94.4%	97.1%	99.8%	99.9%	72.2%	93.2%	99.2%	100.0%
FF-CNN		91.5%	96.5%	99.8%	99.8%	82.8%	98.8%	99.8%	99.8%

Note: Values in bold indicate the highest scores achieved across the models for each metric.

To understand how these attack strategies affect the dynamical nature of our models, we further explore their responses under both targeted and untargeted attack scenarios. The dynamical models (Crop-S, Crop-D, Retina, and S3TA) are set to take 12 gazes at each image under the untargeted/targeted attack to minimize/maximize the prediction probability of the true/target class from all gazes. To reflect the stochastic nature of eye movements (Burak et al., 2010; Kuang et al., 2012), gaussian noise (with 0 mean and 0.1 standard deviation) is added to the fixation points generated from the models (Crop-S, Crop-D, and Retina), and expectation over transformation (EOT; Athalye et al., 2018) is used to deal with the randomness. In our experiments, we average the gradients over 40 iterations. Considering the huge amount of computational costs from 100 iterations of PGD and 40 iterations from EOT, we further sample the validation set of ImageNet100 to include 1000 images to reduce computational costs. For an SPSA attack, we use a sample size of 4096.

4.3.1 Effects of the Retinal Transformation. Table 3 shows the ASR of dynamical models at the 12th step given adversarial examples with various adversarial noise levels, epsilon (ϵ). The result shows that the models with overt attention (Crop-S, Crop-D, and Retina) are more robust than the covert attention (S3TA). Table 4 shows the results from FGSM attack and SPSA attack. Consistent with previous findings, we observe that combining multiresolution patches (Crop-D) is more robust than a single-resolution patch (in Crop-S; Vuyyuru et al., 2020). The model Retina maintains the highest robustness compared to other baseline models and outperforms Crop-S and Crop-D by a large margin. This result suggests that nonuniform retinal transformation is a key mechanism for adversarial robustness among the three sampling strategies. In general, models using explicit fixations (Crop-S, Crop-D, and Retina) demonstrate greater robustness compared to S3TA, whose robustness is relatively low, and akin to that of FF-CNN when adversarial training is not applied.

Table 4: Attack Success Rates from FGSM and SPSA.

Model		FGSM				SPSA			
	ϵ:	2e-3	3e-3	5e-3	7e-3	7e-3	1e-2	2e-2	3e-2
Retina		**21.7%**	**31.7%**	**42.9%**	**52.6%**	**5.7%**	**13.5%**	**21.5%**	**27.3%**
Crop-D		33.3%	44.1%	59.8%	65.8%	21.5%	26.7%	46.5%	61.9%
Crop-S		52.2%	64.6%	77.9%	86.5%	65.9%	77.5%	89.5%	96.8%
S3TA		67.9%	76.5%	87.0%	90.6%	100.0%	100.0%	100.0%	100.0%
FF-CNN		82.0%	88.2%	93.5%	92.0%	100.0%	100.0%	100.0%	100.0%

Note: Values in bold indicate the highest scores achieved across the models for each metric.
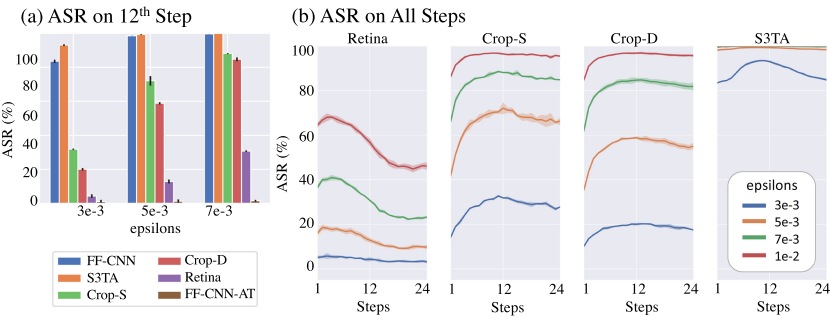

Figure 6: Results of targeted PGD attack. (a) Attack success rate (ASR) from 12th step. (b) ASR from all steps.

It should be noted that the performance of the S3TA model in our study, particularly its attack success rate, differs from what was reported in the original paper. This discrepancy arises from our focus on architectural design choices rather than learning strategies, such as adversarial training. In our experiments, we employed S3TA to contrast the effects of overt attention (eye movements) and covert attention (soft attention) without the adversarial training used in the original study. As a result, the S3TA model in our research, lacking this adversarial training component, exhibits a higher attack success rate than its counterpart in the original study.

4.3.2 Effects of the Number of Recurrent Steps. Unlike feedforward CNNs, dynamical models can vary and unroll their computational graphs with increasing time. Hence, we evaluate these models' ASR as a function of time at each recurrent inference step. Figure 6a visualizes ASR at the 12th step of the targeted attack from Table 3, including adversarially trained FF-CNN (FF-CNN-AT), which achieves the highest robustness. Figure 6b shows ASR for all 24 steps from the targeted PGD attack. All models except for the model Retina appear to maintain or increase ASR by taking more glances
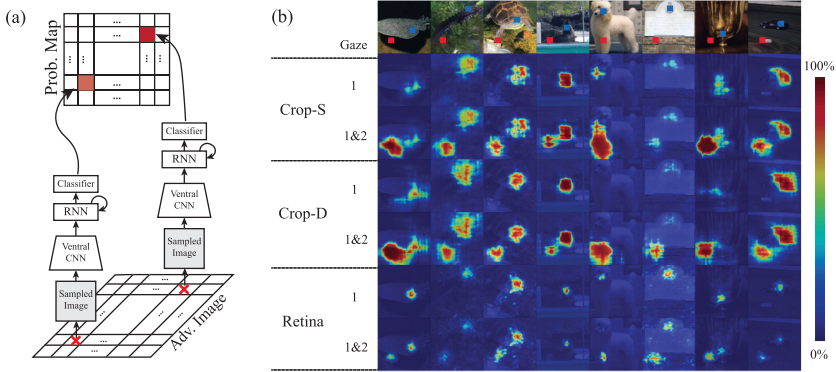
Figure 7: (a) Methodology for generating adversarial probability maps: These maps are created by systematically altering the fixation point across all regions of the image and recording the model's output probability of identifying the target class used in a targeted adversarial attack scenario. (b) Visualization of the experiment: The first row shows original images with fixations used during attack generation (blue dot: first fixation; red dot: second fixation). Subsequent rows present adversarial probability maps for each model under two scenarios during the attack: (1) using only the first fixation and (2) incorporating both first and second fixations. These scenarios reflect different levels of exposure to adversarial influence. The color gradient on the maps signifies the model's likelihood of misclassifying the image as the adversarial target class based on the fixation: warmer colors (toward red) suggest higher probabilities and cooler colors (toward blue) for lower probabilities. This indicates how the fixation strategy affects the model's vulnerability to adversarial attacks.

for object recognition. In contrast, the model Retina increases its robustness during the steps both under 1 to 12 steps and after 12 to 24 steps the attack.

4.3.3 Effect of Sequential Eye Movements. To intuitively understand this difference, we visualize the regions that are affected by the adversarial attack for each model (see Figure 7). For each model (Crop-S, Crop-D, and Retina), we generated two sets of adversarial images: the first set using only the model's initial, learned fixation point, reflecting the model's natural point of focus in an adversarial context; and the second set incorporating an additional, artificially designated fixation, allowing us to examine the impact of a two-step fixation sequence. In this sequence, the first adversarial image is obtained using the model's natural first fixation, while the second image employs both the original and a new, predefined fixation in the lower left corner, effectively simulating a more complex viewing pattern. The second fixation is manually selected at the lower left corner to ensure that the first fixation point will be located in the periphery of the

second fixation. The first (blue) and the second (red) fixations are marked on image examples in the first row of Figure 7b. Two adversarial images from each model are generated under the targeted PGD attack: one deploying only the first fixation (blue) and the other deploying both fixations. The resulting adversarial images are not shown in Figure 7. The same gaussian noise and the targeted attack used for Table 3 and Figure 6 are used again for the experiment.

Once the adversarial images are generated, the models take them as input and perform a recognition task by making fixations on all the possible locations on the images as illustrated in Figure 7a. In this experiment, our goal is to assess the model's response to adversarial attacks by analyzing its performance across a range of potential fixation points. To this end, we overlay a 32×32 virtual grid over each adversarial image, as depicted at the bottom of Figure 7a. This grid represents an array of possible fixation locations, allowing us to simulate how the model might focus on different parts of the image. Each grid point serves as a hypothetical fixation site, enabling a methodical exploration of the model's responses to a diverse set of focal points.

For each grid location, we direct the model's fixation to that point and record its prediction probability for a randomly selected target class used in the targeted attack. These probabilities are mapped to the corresponding locations on the grid, as shown in Figure 7a, top. This approach provides a detailed perspective on how the model processes adversarial images, revealing its susceptibility or resilience to attacks based on where it focuses.

Each fixation is treated as an isolated event in this experiment, meaning the recurrent neural network (RNN) in the model's ventral stream is reset for every new fixation. This enables us to consider the model's object recognition capabilities as a series of independent, one-step tasks. By systematically varying the fixation points across the grid and observing the model's classification outcomes, we gain valuable insights into how its visual attention and processing strategies influence its robustness against adversarial attacks.

Figure 7b displays the probability maps for the target class from the models. In these maps, color indicates the likelihood of the models classifying the image as the targeted class in the adversarial attack, with red signifying a higher probability and blue a lower one. This means that fixations placed on red regions in an adversarial image are likely to lead the model to classify the image as the target class. Conversely, fixations on blue regions tend to result in the model not classifying the image as the target class. Thus, the probability maps effectively represent the model's vulnerability to the adversarial attack, considering all possible fixation locations on the image.

As shown in Figure 7b, the locations near the attacked fixation points are more vulnerable. However, although the fixations are the same across all models, the area that is highly affected by the adversarial attack is different for distinct models. The Retina model shows the smallest area affected

by the adversarial attack compared to the Crop-S and Crop-D models. By comparing the probability maps from a single fixation and two fixations, we observe that the attacked area near the first fixation is persistently affected during the second attack. However, for the Retina model, the area affected by the attack at the first fixation is reduced after including the second fixation.

This observation can be attributed to the fact that the convolutional operation on the retinal images does not maintain the property of translational equivariance, while the convolution on the images in a regular grid does. When the fixation in Crop-S or Crop-D changes a little bit, the representations after convolutional operations will be simply shifted to the new fixation location. However, for the Retina model, small changes in the fixation will result in large changes in the retinal images after foveation and nonuniform sampling. This may further alter its representation significantly in the stacked CNN layers. This fixation-dependent representation is not a simple translation but a change of the whole layout. Therefore, the adversarial noise generated at the certain fixation does not necessarily fool the Retina model when a fixation point is made on other locations.

This is in line with the human brain, where foveal vision and peripheral vision have different functional roles. The foveal vision focuses on object recognition, while the peripheral vision has better adversarial robustness (Harrington & Deza, 2021; Logothetis et al., 1995; Riesenhuber & Poggio, 1999). In Figure 7, when the first fixation is made on the object, the adversarial noise is calculated based on the object in the foveal vision. Then when the second fixation is set at the corner, the object is placed in the peripheral vision. Because of this difference, the adversarial noise from the foveal vision (from the first fixation) fools the model less effectively when it is placed on the peripheral vision (from the second fixation).

5 Discussion

This study introduces brain-inspired mechanisms into computer vision models, specifically focusing on retinal sampling, adaptive eye movement, and recurrent processing. Our models can produce more human-like visual attention and be more robust against adversarial attacks. Through computational experiments, we demonstrate that (1) visual attention emerges as a strategy that optimizes eye movement for efficient object recognition and (2) nonuniform retinal sampling and varying fixations contribute to adversarial robustness.

We found that the Retina model and the Crop-D model exhibit superior performance in mimicking human visual attention compared to baseline models. A key feature of these models is their fovea/periphery approach, which involves high-resolution foveal viewing combined with low-resolution peripheral processing. This approach mirrors the human vision system where both foveal and peripheral views are integral for

attention. The high-resolution foveal view in these models allows for detailed analysis of specific regions of interest, paralleling the human use of fovea to focus on and scrutinize central vision details. In contrast, the low-resolution peripheral processing effectively manages the broader scene context, assisting in overall interpretation and situational awareness, similar to human peripheral vision.

This dual approach to visual processing, inspired by human vision, enables these models to effectively learn and prioritize visual attention. By dynamically shifting focus between detailed and broader views, the models more accurately identify relevant image areas, enhancing their ability to mimic human attention patterns. This capability aligns with the concept of attention mining (Li et al., 2018; Wei et al., 2017) in computer vision, wherein the models, akin to human visual processing, distribute attention across a wider range of relevant features, reducing biases toward certain dominant features. The retina model, in particular, excels in this aspect by adaptively focusing on various parts of an image, thus offering a more balanced and comprehensive analysis.

The implications of these findings are significant, suggesting that integrating elements of human visual processing, especially the balance between foveal and peripheral attention, can create more intuitive and effective AI vision systems. These systems would excel not only in recognizing fine details but also in understanding the broader context of a scene. Such advancements hold promise for applications in areas like autonomous navigation, surveillance, and human-computer interaction, where nuanced visual perception is crucial.

In this study, models incorporating retinal transformation have emerged as the most robust among all tested. This superior robustness can be attributed to retinal sampling, which complicates adversarial attacks. Adversarial perturbations tailored for a specific retinal pattern, associated with a given fixation, prove difficult to generalize across different retinal patterns. This is because as the eyes move and fixate on different locations, the retinal patterns of the same objects are represented with varying features, resolutions, and scales. Unlike models that process images on a regular grid, where perturbations at one location can easily be transferred to other areas, retinal sampling introduces an element of unpredictability for attackers, thereby enhancing robustness.

Building on the advantages of retinal transformation, models with sequential eye movement further augment this robustness. In scenarios without time constraints, such models have the capacity to scan an image multiple times, each time attending to a different region. This adaptability, especially when eye movements are stochastic, introduces an additional layer of complexity for attackers. If attackers lack knowledge of the model's subsequent fixation points, our experiments suggest that the model can use additional glances to rectify errors induced by adversarial images. Notably, models with overt eye movement or overt attention

consistently demonstrate greater robustness compared to the model with covert attention (S3TA). This finding underscores the significance of variable attention mechanisms in enhancing model resilience. The integration of these brain-inspired approaches—retinal transformation and sequential eye movements—not only bolsters the models' defenses against adversarial attacks but also aligns their functioning more closely with human visual processing. Such advancements hold promise for future developments in AI, potentially leading to more intuitive, efficient, and resilient computer vision systems.

Building on the strengths and insights gained from our research, it is important to also acknowledge some of the limitations and areas for further exploration. In advancing the work of Vuyyuru et al. (2020), we integrated an attention mechanism in our models that learns and adapts fixation points dynamically. This approach marks a departure from the static fixation strategies used in prior studies. However, our investigation into the impact of this learned fixation strategy on model robustness revealed complex dynamics. We observed no significant difference in robustness between learned and random or predefined fixations (results are not included in the main text). However, a noticeable decrease in robustness occurred when models operated without the inhibition of return (IOR), suggesting the importance of varied and shifting fixation points in enhancing adversarial robustness.

The influence of the data set characteristics on our findings is also noteworthy. Our evaluations primarily used the ImageNet data set, characterized by images with large, prominent objects. In such a context, both random and learned fixations often focus on these dominant objects, reducing the distinction between the two fixation strategies. A clearer difference might emerge with data sets featuring smaller objects, where the precision of fixations could play a more significant role. Absence of IOR led to static fixations, allowing for the persistence of adversarial patterns. In contrast, models with IOR or adaptive fixation strategies, which involve dynamic eye movements, showed an ability to mitigate the impact of adversarial patterns over time, enhancing overall robustness.

In our exploration of brain-inspired mechanisms within computer vision models, we have only scratched the surface of potential applications. While our study primarily focused on retinal transformation and adaptive eye movements, other mechanisms prevalent in the brain, such as top-down and recurrent connections, hold significant promise. In our models, recurrent connections were utilized at the end of the `VentralCNN`, but top-down connections, crucial for integrating higher-level cognitive processes with sensory inputs, were not implemented. Delving deeper into these mechanisms could pave the way for developing more sophisticated and resilient computer vision models. This exploration has the potential not only to enhance AI technology but also to contribute to our understanding of human

vision processes and behaviors, bridging the gap between computational modeling and biological reality.

It is also essential to address some limitations in our study. One notable area is the object recognition accuracy of models using retinal transformation, particularly in single-label classification tasks. This approach, while mimicking the human eye's foveal-peripheral dynamics, introduces a nonlinear distortion to images. The high-resolution focus in the fovea and significant downsampling in peripheral areas can affect the efficiency of convolutional operations, especially since the same kernels process regions of varying scale. Consequently, our model exhibits suboptimal performance compared to other baseline methods when dealing with large objects that extend into the peripheral field. Conversely, for images with smaller objects or more complex scenes, the retinal transformation demonstrates superior performance, underscoring its suitability for handling intricate visual information.

Appendix A: Foveated Imaging

In the foveated image, the foveated region maintains high acuity, but the periphery has reduced acuity. To implement the eccentricity-dependent acuity, we apply different levels of gaussian blur based on the eccentricity. Here, we assume that the fixation is at the center of an image. Similar procedures can be generalized to noncentered fixation.

To illustrate a varying blurring effect, we use three 2D isotropic gaussian kernels, K_{b1}, K_{b2}, and K_{b3}. The kernel K_{bi} ($i \in [1, 2, 3]$) is in size 7×7 and has a variance σ_{bi}^2. The σ_{b1}, σ_{b2}, and σ_{b3} are set to 1, 3, and 5, respectively, in this study.

First, the gaussian kernels K_{b1}, K_{b2}, and K_{b3} are separately applied to the original image, resulting in blurred images, A_{b1}, A_{b2}, and A_{b3}. The original image (A) and the blurred images are shown in Figure 8a.

Second, region masks are produced. The region masks are used to confine the regions with a specific level of blurring. The masks appear as concentric rings. They act as filters to pass the image with the desired resolution to the target area. Applying such masks to the blurred images makes the blurring dependent on the eccentricity. For a foveated image, a region mask designed to pass the fovea region is applied to an image with original resolution (see Figure 8b, top row). For the periphery, region masks for the peripheral regions are applied to the blurred images (shown in Figure 8b in the second, third, and fourth rows from the top).

To generate region masks, 2D isotropic gaussian kernels are first generated. The gaussian kernels are denoted as T_{t1}, T_{t2}, and T_{t3}, and they are in size 224×224. T_{t1}, T_{t2}, and T_{t3} have means at the fixated location and variances σ_{t1}^2, σ_{t2}^2, and σ_{t3}^2, respectively. (In this study, $\sigma_{t1} = 40$, $\sigma_{t2} = 70$, $\sigma_{t3} = 90$.) The generated gaussian kernels are normalized to make
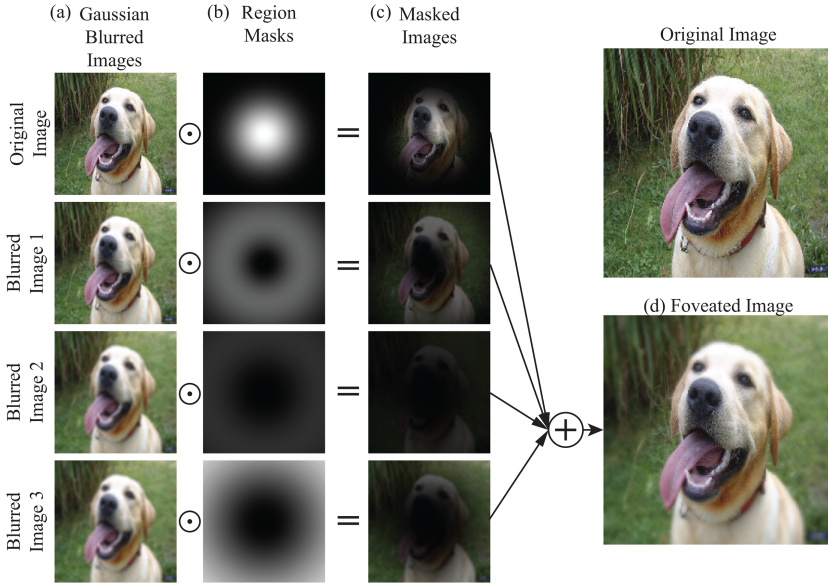
Figure 8: Illustration of the foveated image generated from an example image, when the fixation is at the image center. (a) The original image is blurred by applying different gaussian kernels, with (b) region masks applied to the blurred images. The black and white colors represent the range of values (0, 1). (c) Results of element-wise multiplication of panels a and b. (d) The foveated image by summing up the masked images.

their maximal amplitude 1.0. The region masks M_1, M_2, M_3, and M_4 are generated as equations A.1a to A.1d

$$M_1 = T_{t1}, \tag{A.1a}$$

$$M_2 = T_{t2} - T_{t1}, \tag{A.1b}$$

$$M_3 = T_{t3} - T_{t2}, \tag{A.1c}$$

$$M_4 = 1 - T_{t3}, \tag{A.1d}$$

where 1 is a 2D matrix filled with ones whose size is identical to T_{t3} The region masks are shown in Figure 8b, and it can be checked that each region mask passes distinct image regions based on the eccentricity.

Once the blurred images and the region masks are obtained, the region masks are applied to the corresponding original and blurred images as element-wise multiplications (see equations A.2a to A.2d):

$$A_{m1} = A \odot T_{t1}, \tag{A.2a}$$

$$A_{m2} = A_{b1} \odot T_{t2}, \tag{A.2b}$$

$$A_{m3} = A_{b2} \odot T_{t3}, \tag{A.2c}$$

$$A_{m4} = A_{b3} \odot T_{t4}, \tag{A.2d}$$

where the operation $\odot$ is an element-wise multiplication. The first region mask T_{t1} designed to pass the center of fixation is applied to the original image so that the foveated image maintains high acuity in the fixated region. On the contrary, the region masks assigned to the peripheral regions are applied to the corresponding blurred images. The masked images are shown in Figure 8c.

Finally, the masked images are summed together to produce the foveated image:

$$\text{Foveated image} = A_{m1} + A_{m2} + A_{m3} + A_{m4}. \tag{A.3}$$

The resulting image is shown in Figure 8d. In the foveated image, the region near the fixation maintains the high acuity, and the periphery regions are gradually blurred as the eccentricity grows.

Appendix B: ImageNet100

The randomly selected 100 classes from 1000 classes of ILSVRC2012 (Deng et al., 2009) are:

```
[n01496331   n01756291   n01833805   n02025239   n02100583   n02137549
n02480495   n02808440   n03124043   n03291819   n03770679   n03902125
n04201297   n04371430   n09246464   n01531178   n01768244   n01843383
n02028035   n02102480   n02138441   n02492660   n02834397   n03127747
n03388183   n03781244   n03956157   n04251144   n04399382   n09472597
n01630670   n01797886   n01847000   n02077923   n02105412   n02172182
n02504458   n02892201   n03131574   n03443371   n03785016   n04037443
n04275548   n04505470   n01644900   n01806143   n01871265   n02087046
n02106166   n02190166   n02640242   n02963159   n03180011   n03494278
n03796401   n04040759   n04325704   n04536866   n01667778   n01807496
n01872401   n02088632   n02108089   n02233338   n02701002   n02971356
n03216828   n03662601   n03837869   n04049303   n04335435   n04589890
n01669191   n01817953   n01968897   n02089078   n02113712   n02317335
n02787622   n02977058   n03249569   n03710193   n03877472   n04146614
n04355338   n04591713   n01694178   n01824575   n01980166   n02094258
n02119789   n02410509   n02795169   n03100240   n03272010   n03733131
n03877845   n04162706   n04356056   n07614500]
```

Acknowledgments

References

Akbas, E., & Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLOS Computational Biology*, *13*(10), e1005743. 10.1371/journal .pcbi.1005743

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In *Proceedings of the International Conference on Machine Learning* (pp. 284–293).

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), art. 13.

Bao, W., & Chen, Z. (2020). Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing*, *404*, 154–164. 10.1016/j.neucom.2020.03.060

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439). 10.1126/science.aav9436

Berga, D., & Otazu, X. (2020). Modeling bottom-up and top-down attention with a neurodynamic model of V1. *Neurocomputing*, *417*, 270–289. 10.1016/j.neucom .2020.07.047

Berrios, W., & Deza, A. (2022). *Joint rotational invariance and adversarial training of a dual-stream transformer yields state of the art brain-score for area V4*. arXiv:2203.06649.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*(5241), 177–178. 10.1038/226177a0

Burak, Y., Rokni, U., Meister, M., & Sompolinsky, H. (2010). Bayesian model of dynamic image stabilization in the visual system. *Proceedings of the National Academy of Sciences*, *107*(45), 19525–19530. 10.1073/pnas.1006076107

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy* (pp. 39–57).

Cheung, B., Weiss, E., & Olshausen, B. (2016). *Emergence of foveal image sampling from learning to attend in visual scenes*. arXiv:1611.09430.

Choi, M., Han, K., Wang, X., Zhang, Y., & Liu, Z. (2023). *A dual-stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains*. arXiv:2310.13849.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling*. arXiv:1412.3555.

Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, *22*(1), 319–349. 10.1146/annurev.neuro.22.1.319

Coletta, N. J., & Williams, D. R. (1987). Psychophysical estimate of extrafoveal cone spacing. *JOSA A*, *4*(8), 1503–1513. 10.1364/JOSAA.4.001503

Connolly, M., & Van Essen, D. (1984). 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. 10.1002/cne.902260408

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. 10.1038/nrn755

Curcio, C. A., & Allen, K. A. (1990). Topography of ganglion cells in human retina. *Journal of Comparative Neurology*, *300*(1), 5–25. 10.1002/cne.903000103

Curcio, C. A., Sloan, K. R., Kalina, R. E., & Hendrickson, A. E. (1990). Human photoreceptor topography. *Journal of Comparative Neurology*, *292*(4), 497–523. 10.1002/cne.902920402

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., & DiCarlo, J. J. (2020). *Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations*. bioRxiv.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Derrington, A., & Lennie, P. (1984). Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *Journal of Physiology*, *357*(1), 219–240. 10.1113/jphysiol.1984.sp015498

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837. 10.1016/0042-6989(95)00294-4

Deza, A., & Konkle, T. (2020). *Emergent properties of foveated perceptual systems*. arXiv:2006.07991.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. 10.1016/j.neuron.2012.01.010

Duchowski, A. T., Cournia, N., & Murphy, H. (2004). Gaze-contingent displays: A review. *Cyberpsychology and Behavior*, *7*(6), 621–634. 10.1089/cpb.2004.7.621

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). *Adversarial examples that fool both computer vision and time-limited humans*. arXiv:1802.08195.

Esteves, C., Allen-Blanchette, C., Zhou, X., & Daniilidis, K. (2017). *Polar transformer networks*. arXiv:1709.01889.

Feichtenhofer, C., Pinz, A., & Wildes, R. P. (2017). Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4768–4777).

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201. 10.1038/nn.2889

Gattass, R., Gross, C., & Sandell, J. (1981). Visual topography of V2 in the macaque. *Journal of Comparative Neurology*, *201*(4), 519–539. 10.1002/cne.902010405

Gattass, R., Sousa, A., & Gross, C. (1988). Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of Neuroscience*, *8*(6), 1831–1845. 10.1523/JNEUROSCI.08-06-01831.1988

Geisler, W. S., & Hamilton, D. B. (1986). Sampling-theory analysis of spatial vision. *JOSA A*, *3*(1), 62–70. 10.1364/JOSAA.3.000062

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv:1412.6572.

Gu, S., & Rigazio, L. (2014). *Towards deep neural network architectures robust to adversarial examples*. arXiv:1412.5068.

Guo, Y., Ji, J., Lu, X., Huo, H., Fang, T., & Li, D. (2019). Global-local attention network for aerial scene classification. *IEEE Access*, *7*, 67200–67212. 10.1109/ACCESS.2019.2918732

Harrington, A., & Deza, A. (2021). Finding biological plausibility for adversarially robust features via metameric tasks. In *Proceedings of the SVRHM 2021 Workshop@ NeurIPS*.

He, W., Wei, J., Chen, X., Carlini, N., & Song, D. (2017). Adversarial example defense: Ensembles of weak defenses are not strong. In *Proceedings of the 11th USENIX Workshop on Offensive Technologies*, *17*.

Huang, Y., Gornet, J., Dai, S., Yu, Z., Nguyen, T., Tsao, D., & Anandkumar, A. (2020). Neural networks with recurrent generative feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems*, *33* (pp. 535–545). Curran.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. 10.1038/35058500

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. 10.1109/34.730558

Jabbireddy, S., Sun, X., Meng, X., & Varshney, A. (2022). *Foveated rendering: Motivation, taxonomy, and research directions*. arXiv:2205.04529.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021). Perceiver: General perception with iterative attention. In *Proceedings of the International Conference on Machine Learning* (pp. 4651–4664).

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1072–1080).

Jonnalagadda, A., Wang, W. Y., Manjunath, B., & Eckstein, M. P. (2021). *Foveater: Foveated transformer for image classification.* arXiv:2105.14173.

Kaplanyan, A. S., Sochenov, A., Leimkühler, T., Okunev, M., Goodall, T., & Rufo, G. (2019). Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics*, *38*(6), 1–13. 10.1145/3355089.3356557

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. 10.1038/s41593-019-0392-5

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863. 10.1073/pnas.1905544116

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization.* arXiv:1412.6980.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In S. Goldstein, D. Princiotta, & J. Naglieri (Eds.), *Matters of intelligence* (pp. 115–141). Springer.

Kuang, X., Poletti, M., Victor, J. D., & Rucci, M. (2012). Temporal encoding of spatial information during active visual fixation. *Current Biology*, *22*(6), 510–514. 10.1016/j.cub.2012.01.050

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579. 10.1016/S0166-2236(00)01657-X

Lettvin, J. Y. (1976). On seeing sidelong. *Sciences*, *16*(4), 10–20. 10.1002/j.2326-1951.1976.tb01231.x

Li, K., Wu, Z., Peng, K.-C., Ernst, J., & Fu, Y. (2018). Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9215–9223).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European Conference on Computer Vision*, 740–755.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552–563. 10.1016/S0960-9822(95)00108-4

Luo, Y., Boix, X., Roig, G., Poggio, T., & Zhao, Q. (2015). *Foveation-based mechanisms alleviate adversarial examples*. arXiv:1511.06292.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). *Towards deep learning models resistant to adversarial attacks.* arXiv:1706.06083.

Merigan, W. H., & Maunsell, J. H. (1993). How parallel are the primate visual pathways? *Annual Review of Neuroscience*, *16*(1), 369–402. 10.1146/annurev.ne.16.030193.002101

Min, J., Zhao, Y., Luo, C., & Cho, M. (2022). *Peripheral vision transformer.* arXiv:2206.06801.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. 10.1016/0166-2236(83)90190-X

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). *Recurrent models of visual attention.* arXiv:1406.6247.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391. 10.1038/nature03390

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In *Proceedings of the 2016 IEEE Symposium on Security and Privacy* (pp. 582–597).

Perry, J. S., & Geisler, W. S. (2002). Gaze-contingent real-time simulation of arbitrary visual fields. *Human Vision and Electronic Imaging*, *7*, *4662*, 57–69.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and Performance X: Control of Language Processes*, *32*, 531–556.

Pramod, R., Katti, H., & Arun, S. (2022). Human peripheral blur is optimal for object recognition. *Vision Research*, *200*, 108083. 10.1016/j.visres.2022.108083

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. 10.1038/14819

Rizzolatti, G., & Matelli, M. (2003). Two different streams form the dorsal visual system: Anatomy and functions. *Experimental Brain Research*, *153*, 146–157. 10.1007/s00221-003-1588-0

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, *2*, 437–457. 10.1146/annurev-vision-082114-035733

Sermanet, P., Frome, A., & Real, E. (2014). *Attention for fine-grained categorization.* arXiv:1412.7054.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, 27. Curran.

Stewart, E. E., Valsecchi, M., & Schütz, A. C. (2020). A review of interactions between peripheral and foveal vision. *Journal of Vision*, *20*(12), 2. 10.1167/jov.20.12.2

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks.* arXiv:1312.6199.

Thavamani, C., Li, M., Cebron, N., & Ramanan, D. (2021). Fovea: Foveated image magnification for autonomous navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15539–15548).

Thibos, L., Cheney, F., & Walsh, D. (1987). Retinal limits to the detection and resolution of gratings. *JOSA A*, *4*(8), 1524–1529. 10.1364/JOSAA.4.001524

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. 10.1038/381520a0

Uesato, J., O'Donoghue, B., Kohli, P., & Oord, A. (2018). Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of the International Conference on Machine Learning* (pp. 5025–5034).

Vuyyuru, M. R., Banburski, A., Pant, N., & Poggio, T. (2020). Biologically inspired mechanisms for adversarial robustness. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems*, *33* (pp. 2135–2146). Curran.

Wang, B., Mayo, D., Deza, A., Barbu, A., & Conwell, C. (2021). *On the use of cortical magnification and saccades as biological proxies for data augmentation.* arXiv:2112.07173.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision* (pp. 20–36).

Wang, Y., Lv, K., Huang, R., Song, S., Yang, L., & Huang, G. (2020). *Glance and focus: A dynamic approach to reducing spatial redundancy in image classification.* arXiv:2010.05300.

Wang, Y., Wang, B., Wu, X., & Zhang, L. (2017). Scanpath estimation based on foveated image saliency. *Cognitive Processing*, *18*, 87–95. 10.1007/s10339-016 -0781-6

Watson, A. B. (2014). A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision*, *14*(7), 15. 10.1167/ 14.7.15

Weber, C., & Triesch, J. (2009). Implementations and implications of foveated vision. *Recent Patents on Computer Science*, *2*(1), 75–85.

Wei, Y., Feng, J., Liang, X., Cheng, M.-M., Zhao, Y., & Yan, S. (2017). Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1568–1576).

Wiecek, E., Pasquale, L. R., Fiser, J., Dakin, S., & Bex, P. J. (2012). Effects of peripheral visual field loss on eye movements during visual search. *Frontiers in Psychology*, *3*, 472. 10.3389/fpsyg.2012.00472

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3–4), 229–256.

Xie, C., Wang, J., Zhang, Z., Ren, Z., & Yuille, A. (2017). *Mitigating adversarial effects through randomization.* arXiv:1711.01991.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning* (pp. 2048–2057).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921–2929).

Zoran, D., Chrzanowski, M., Huang, P.-S., Gowal, S., Mott, A., & Kohli, P. (2020). Towards robust image classification using sequential attention models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9483–9492).