Retinotopy-Inspired Transformations Boost Robustness and Object Localization in Convolutional Neural Networks for Categorization Tasks

Jean-Nicolas Jérémie[1], Emmanuel Daucé[1,2], Laurent U Perrinet[1*]

1 Institut de Neurosciences de la Timone, Aix-Marseille Université - CNRS UMR 7289, Marseille, France
2 Ecole centrale Méditerranée, Marseille, France

* laurent.perrinet@univ-amu.fr

Abstract

The visual systems of numerous animal species, including humans, exhibit a distinctive characteristic known as a fovea. This fovea represents an area within the retina where the highest level of visual acuity is concentrated at the center of gaze. In contrast, peripheral vision possesses lower resolution compared to the fovea's sharpness. Retinotopy plays a crucial role in biological vision but remains underutilized in current machine learning applications. This study investigates whether retinotopic mapping can be leveraged to enhance deep convolutional neural networks (CNNs) on the ImageNet image categorization and localization task, consisting of approximately one million images and 1000 labels. By integrating retinotopic mapping into the input layers of standard ResNet architectures and retraining these models using log-polar mapped images, we achieved comparable performance in classification. Moreover, this approach improved the network's ability to handle arbitrary image zooms and rotations while addressing vulnerabilities exhibited by classical models. Crucially, efficiency depended highly on the point of fixation, with output likelihood serving as a proxy for localizing objects of interest. This study suggests that retinotopic mapping in foveated vision implements prior knowledge of visual object geometry, providing an efficient solution to address the visual search problem, an essential function for predators.

Author summary

The fovea is a specialized structure found in the eyes of many animal species, including humans, that concentrates high visual acuity at the center of gaze. Our goal was to explore whether this unique feature could be leveraged through deep learning architectures to enhance object detection and localization tasks in computer vision. By incorporating retinotopic mapping into convolutional neural networks (CNNs), we observed improved performance in identifying and locating visual objects while maintaining resilience against common geometrical transformations such as rotations and zooms. This study demonstrates the potential of foveated vision to advance machine learning and computer vision, potentially bringing these models closer to human-like efficiency and adaptability in processing complex visual scenes. Furthermore, utilizing these computational models allows us to better understand these fundamental neurobiological features.

Introduction: Properties of the visuo-motor system endowing visual search

This introduction explores the foundations of visual search, drawing upon historical and modern studies while emphasizing its significance in predator evolution. Efficient prey detection relies heavily on a predator's ability to swiftly move their gaze across the environment by displacing the fovea, which contains a high concentration of photoreceptors responsible for sharp central vision [1]. This defines visual search as the process by which organisms examine their visual environment to locate and identify objects of interest. For predators like raptors [2] and sharks [3], efficient prey detection is vital for survival. The systematic study of visual search began with pioneering work by Yarbus [4] and Noton et al. [5], who introduced the concept of a "scan path." A scan path refers to the trajectories of eye movements produced during visual exploration, which are not random but guided by an individual's objectives as demonstrated by Yarbus through experiments involving different exploration goals and realistic paintings or photographs. Noton et al.'s methodologies for recording and analyzing these movements provided a solid foundation for understanding the mechanisms underlying visual search. Their approach involves non-invasive monocular measurement of eye movements using the diffuse scleral reflection technique, allowing ecological observation of eye movements during recognition tasks [5]. Crucially, moving the eyes allows to displace the fovea and thus, visual search is tightly dependent on retinotopy. As a consequence, visual search plays a crucial role in predator survival. In this context, it would be interesting to explore how different species have evolved their visual systems and strategies for efficient prey detection, considering factors such as habitat type, prey behavior, and predator adaptations.

Retinotopy: A non-linear radial organization

In order to understand the underlying mechanisms that determine the generation of scan paths in visual search, it is essential to examine retinotopy along the visual pathways. The visual system in humans and many mammals is characterized by a substantial resolution disparity between the central area of the visual field (fovea) and the peripheral regions, where the number of photoreceptors decreases exponentially with eccentricity [6]. This radial structure of the retina, with maximal visual acuity at the center and decreasing in the periphery, plays a key role in this exploration. This organization is moreover maintained across the different stages of visual processing in the brain. Theoretical advances have been supported and confirmed by experiments using modern imaging techniques, notably by using functional magnetic resonance imaging (fMRI) to map visual responses in the human cortex. A first study by Tootell et al [7] focuses on spatial attention and explores the response when displaying a bar stimulus in a visual field reduced to a quadrant. A retinotopic mapping of the visual field is found in the early visual area, V1, V2 and V3. Dougerthy et al [8] then worked on the representation in these same areas when displaying stimuli varying in eccentricity or angular direction in direct connection with the log-polar referential. These studies demonstrated how different regions of the visual cortex are activated based on the position of stimuli in the visual field, corroborating previously proposed models. Such topographic maps, which transform the spatial relationships of sensory inputs, are a fundamental component of processing in species such as carnivores or primates [9], suggesting that they confer evolutionary advantages, although the precise computational benefits remain to be found.

In addition, these precise retinotopic mappings are thought to facilitate efficient parallel processing of spatial features. Such a topographic organization also facilitates

connections within and between processing areas that respect the geometry of the sensory epithelium and minimize global wiring length. However, the function of this topography is still debated, with one possibility being that it is merely an artifact of the scaffolding that operates during development [10].

Eye movements and the sequential analysis of the visual scene

A natural question is what advantages these retinotopic visual inputs confer on information processing. Numerous hypotheses have been proposed regarding the role of this mapping of the visual field. One primary explanation is the coupling of foveal inputs with visual exploration: a retina with a fovea allows efficient visual processing if the eye can actively move and focus attention on specific points of interest. This may explain the coupling of foveal inputs with visual exploration : a retina with a fovea allows for efficient visual processing when the eye actively moves and focuses on specific objects.

The fovea is thus associated with a set of oculomotor behaviors aimed at positioning objects of interest at the center of the retina (saccades, smooth pursuit, vestibulo-ocular reflex...), maximizing access to visual information for those objects. When we visually explore a scene, our eyes perform saccades, rapid eye movements that reposition the fovea onto successive points of interest. Between these saccades, fixations allow for precise visual information processing. This cycle of saccades and fixations form the basis of our scan path, influencing how we perceive and interact with our environment. One primary hypothesis thus posits that the cone distribution reflects the probability of an object of interest being placed at the center of the retina, given the oculomotor system [11, 12].

Understanding how the visual search process functions in humans involves examining the underlying mechanisms that guide our eye movements and visual attention. Research shows that our visual exploration is influenced by various factors, including the saliency of objects in our visual field [13]. Studies using eye tracking have revealed that our eyes do not move randomly but follow predictable patterns that maximize the efficiency of visual information gathering [14]. Recent studies have shown that this combination of saccades and foveal input, coupled with an effective mechanism for detecting points of interest, significantly enhances visual acuity [15–17]. This supports functions such as the integration of local feature analysis into global perceptual representations.

Biologically-inspired retinotopic mapping in computer vision

Observations on visual acuity and eccentricity notably led to the development of novel artificial vision models. Sandini et al [18] were among the first to develop these models, drawing inspiration from the structures and functions observed in the human eye. In their study, they propose a model that samples the visual scene with a resolution dependent on the eccentricity, already demonstrating the contribution of this kind of transformation to compress visual information. This characteristic was brought to light in the field of computer vision applied to robotics, where the exploration of larger visual scene could be made possible with a reduced computational cost. Advances in neurology and cognitive psychology have led to the development of artificial vision algorithms that mimic human visual mechanisms, one of the most robust models for representing this topography is the log-polar model, which is described in detail by Arujo et al [19]. Indeed, the model is organized around two polar coordinates: azimuth (angle) and eccentricity (distance from the center). There is an over-representation around the center, the fovea. This area of high resolution is contrasted with the periphery. In addition to information compression at the periphery, the analysis of the various transformations, such as rotation and zoom, is equivalent to translation in the new

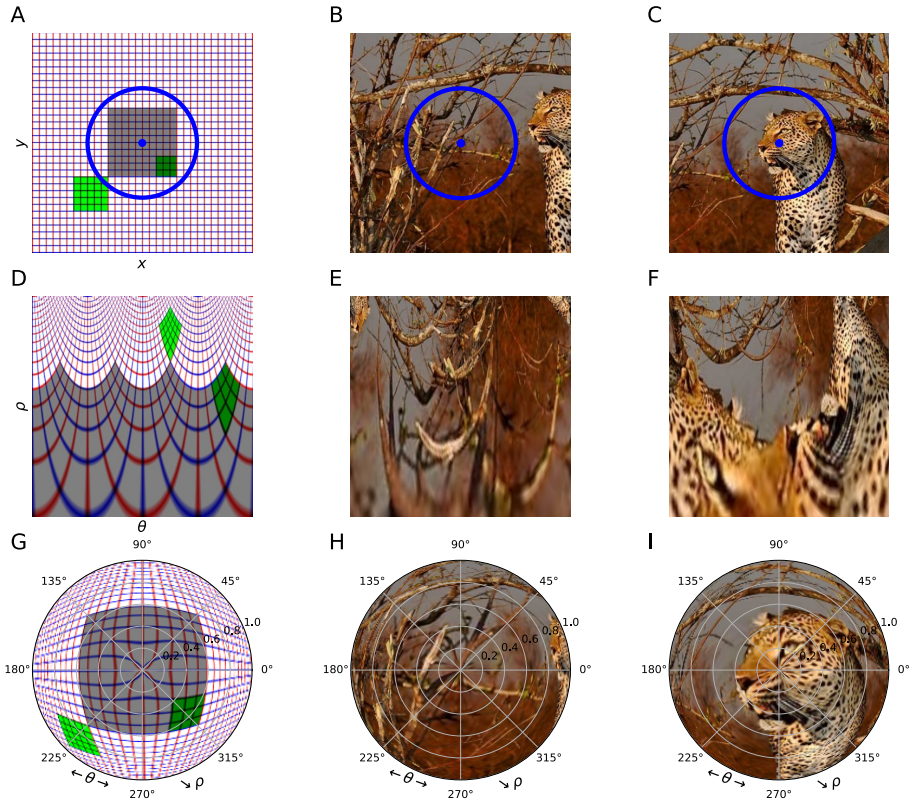

Fig 1. We illustrate the process of mapping input images defined in Cartesian coordinates to retinotopic space using a log-polar transformation with the fixation point (marked by a blue disk) and the approximate area of the fovea (blue circle). In **(A)**, the input image is defined as a regular grid representing the Cartesian coordinates (x, y) by vertical (red) and horizontal (blue) lines. As shown in **(D)**, by applying the log-polar transform to this image, the coordinates of each pixel with respect to the fixation point are transformed based on its azimuth angle θ (abscissa) and the logarithm of its eccentricity $\rho = \log(\sqrt{x^2 + y^2})$ (ordinates). This transformation results in an over-representation of the central area and a deformation of the visual space. Note that the green square is translated in retinotopic space when it is scaled and rotated. When the transformation is applied to a natural image, as shown in **(B)**, there is a noticeable compression of information in the periphery, as shown in **(E)**. As shown in **(F)**, this representation is highly dependent on the fixation point, as indicated by the shift shown in **(C)** when the fixation point is moved to the right and down.

reference frame. The definition of the model based on the log polar referential allowed the development of more biological approach such as [20] where they proposed strategies to exploit these characteristics, paving the way for applications in artificial vision and robotics, where efficient visual perception is crucial.

The implementation of the principles of visual search using retinotopic sensors nevertheless presents important challenges related to the issue of information retrieval: it is difficult to estimate in advance (before actuating an eye movement) which future position contains the most relevant information. Thus, in the analysis of a scene, the eye must guess in advance the regions of interest, even before they are positioned at the center of the retina. Two broad categories of approaches can be defined to characterize this anticipatory property of the visual system. The first family of methods (called bottom-up) considers only low-level features to characterize the regions that attract the eye (regions that significantly deviate from the average statistics of the scene). This is the approach proposed, for example, by Itti and Koch [13]. Conversely, top-down approaches use prior knowledge about visual content to direct attention to potentially interesting regions which are not yet explored [21]. There are few works utilizing this principle of "predictive coding" in the case of retinotopic sensors, see for instance [22, 23].

Visual search has varied applications, ranging from artificial vision and robotics to cognitive psychology and neurobiology. Models inspired by human vision, such as the log-polar model, are used to develop artificial vision systems capable of processing visual information efficiently and robustly [13]. These systems found applications in areas such as image recognition, autonomous robot navigation, and even in human-computer interfaces. In cognitive psychology, visual search helps to better understand the processes of perception, attention, and visual memory. Studies on eye movements and visual search patterns provide valuable insights into how our brain processes visual information and how these processes can be altered in certain pathological conditions, such as attention disorders or neurodegenerative diseases [24]. An intriguing aspect of visual search is how natural vision models inspire developments in artificial intelligence [REF?].

Our proposal: a biologically inspired mapping to CNN inputs

Despite the growing attention paid to this field and the rapid evolution of computer science in recent years, the contribution of retinotopic mapping to visual search remains uncertain. A significant challenge remains in understanding the impact of different aspects of vision, including eye movements, on visual search performance. The objective of this study is to pursue the integration of these approaches with the intention of transferring the biological aspects of neuroscience to the computational domain. We examine the potential integration of retinotopic mapping into classical computational architectures that model biological visual pathways, with a particular focus on deep Convolutional Neural Networks (dCNNs). In particular, we aim at improving image categorization and localization performance and robustness.

Here, we propose to take advantage of this biologically inspired preprocessing approach. Our goal is to build zoom and rotation invariance directly into the architecture of the network. Inspired by retinotopic mapping in visual pathways, we apply a transformation to input images such that translations in the transformed domain correspond to rotations or scale changes in the original spatial domain. By then using a convolutional network to process images in this transformed space, the local connectivity of the convolutions inherently imparts relative rotation and scale invariance to the model. Rather than expecting learned filters alone to discover such invariance post hoc, our approach maps rotations and zooms to translations at the input level. We hypothesize that this biological mapping front-end will improve the CNN's robustness

to controlled geometric perturbations of images.

Contributions of the paper

This paper makes several contributions:

- We show that popular off-the-shelf CNNs can be substantially perturbed by a simple rotation.
- We introduce a biologically inspired mapping to CNN inputs, inspired by retinotopic mapping in visual pathways, to build zoom and rotation invariance directly into the architecture of the network.
- We demonstrate the effectiveness of transfer learning models pre-trained on standard datasets to then classify inputs transformed in this way, without the need for retraining from scratch.
- We show that our approach induces significant robustness to controlled geometric perturbations compared to baseline models, confirming that the bank of convolutions in this transformed domain confers relative invariance as intended.
- We show improved object localization despite distortions, suggesting that retinotopic transformation may benefit tasks beyond classification.
- We show that this biologically inspired mapping offers promising directions for future research in geometric deep learning, robust vision systems, and computational models of early visual processing.

Implementing a biologically-inspired retinotopic mapping

Retinotopic mapping in humans results from the combined effect of the arrangement of photoreceptors in the retina and their output convergence via the optic nerve. This causes nearby regions of the visual field to activate adjacent neural structures as signals travel from the retina to the brain, while giving more resolution to the central field of view. Simple parameterized transformations computationally model this biologically inspired retinotopic mapping.

The log-polar transform as a robust retinotopic mapping

We implement the retinotopic mapping by transforming the Cartesian pixel coordinates into log polar coordinates. As shown in Figure 1, the location of each pixel in the log-polar domain (ρ, θ) is defined by

$$\rho = \log_2 \sqrt{x^2 + y^2} \quad (1)$$

$$\theta = \arctan(\frac{y}{x}) \quad (2)$$

Where (x, y) are the Cartesian coordinates. The increasing size of the receptive fields with eccentricity ρ also parallels the decreasing retinal ganglion cell density in the periphery of the visual field. This retinotopic mapping has a unique and important property - zooms and rotations in Cartesian space become vertical and horizontal shifts, respectively, in log-polar space. This mapping transforms these two geometric transformations into two independent dimensions, since rotation only changes the

azimuthal angle θ, while zooming affects the logarithmic eccentricity ρ on the elevation axis.

Crucially, by applying convolutions such as those used in CNNs after the log-polar transform, the model inherently acquires relative rotational and scale invariance without having to explicitly learn these transformations[1]. Together, these properties allow our model to take advantage of the translational invariance of convolutions while respecting the geometry of retinotopic mapping, which is computationally advantageous in biological vision systems. This type of transformation is commonly used in computer vision, especially for template matching [19, 25, 26] or robotics [27, 28], but is largely underutilized in deep learning, especially in CNNs.

This transformation is performed with the PyTorch library [29] through the use of the grid_sample() function, as it applies a grid to the pixels of the image in Cartesian coordinates. This function has been implemented particularly efficiently for its use in spatial transformer networks [30]. Thus, with such a log-polar grid, each pixel in Cartesian space is assigned a new position in retinotopic space. We set the number of sampled angles (N_θ) and the number of sampled eccentricities (N_ρ) to 224 to obtain an output image with a resolution of 224×224, which was also used during training. The resolution of the grid was set to the same dimension as the input images in the classical model to avoid bias in the evaluation due to image resolution. All θ values are within a linear distribution in $[0; 2\pi]$, while ρ values are within a logarithmic distribution within $[r_{\min}; r_{\max}]$. In practice, we use a log-polar grid with an outer log-radius of $r_{\max} = 0$ and an inner log-radius of $r_{\min} = -5$. This choice of parameters allows us to focus on the central area of the image, which is the most informative part of the image.

Data sets

We have selected two datasets for our study. The first dataset is that from the ImageNet [31] challenge, which is widely used due to its extensive collection of images and associated labels. This dataset offers rich semantic links, enabling the construction of task-specific datasets. However, it is worth mentioning that ImageNet exhibits some biases, particularly with objects being centered in many images. This feature enables the application of a log-polar transformation, whereby the information is concentrated at the fixation point, which is regarded as the center of the image during training. Notwithstanding the aforementioned, it can be seen that this a priori centered position assumption is not sufficient to generate a dataset that is perfectly suitable for retinotopic transformation. A second step is to utilize the information provided by the ImageNet dataset, specifically the bounding box. Consequently, a sample was selected that encompasses the smallest square containing the aforementioned bounding box, resulting in the generation of a fixation point dataset defined as the center of the bounding box for the label of interest from the ImageNet dataset [31]. This novel dataset was subsequently employed to train the retinotopic and Cartesian networks.

Despite its advantages, ImageNet has limitations for localization tasks. For example, the dataset lacks multi-labels, with only one label per image. Additionally, the proportion of bounding boxes relative to the image size is relatively large, which may limit the impact of certain analyses. To address these limitations, we also utilize the Animal 10k dataset, as referenced in Yu (2021) [32]. The dataset provides a series of key points for each animal joint present in an image. By fitting a Gaussian to the aforementioned key points, heat maps can be generated, centered around the label of interest, which in this case is 'animal' (for further details, please refer to Figure 9 in the Appendix). This approach enables more effective localization and the analysis of visual

[1]We further introduced a circular padding in the convolutions, however we controlled this had little impact overall (not shown).

distribution of animals in images.

Retraining, transfer learning and evaluation of results

Previously, it has been successfully demonstrated that using transfer learning to re-train networks like VGG16 [33] allows its application to different tasks using the semantic link that underlies IMAGENET's labels. It has also been shown that it is possible to predict the likelihood of a trained network on a newly defined task (such as categorizing an animal) using this semantic link that connects the outputs of a pre-trained network to a library of labels, in particular to learn to categorize images containing or not an animal [34]. Therefore, based on these findings, we used these networks to perform a categorization task on the second dataset, ANIMAL 10K [32], seeking animals in the image. We extended the study by retraining a deep CNN RESNET (with 18 to 101 layer groups and provided by the PYTORCH library) on the categorization of the 1000 IMAGENET labels. This deeper networks showed improved robustness, but at the cost of increased computational complexity [35].

We explore two network configurations: one with a retinotopic mapping at the input and one without. To evaluate the efficiency of introducing the retinotopic mapping, both networks are retrained (or fine-tuned) with retinotopic inputs and compared to the legacy networks on the IMAGENET dataset. In a first generation, images were transformed by setting the fixation center to the center of the image, exploiting the fact that images in the IMAGENET dataset are *a priori* more likely to be centered. While giving good results, a number of visual objects were not centered, giving raise to some variability. As information degrades rapidly with eccentricity from the fixation point, the dataset was pre-processed by using the bounding boxes provided with the IMAGENET dataset to center the images. We used a fixation point defined as the center of the bounding box of the label of interest. These new images were used to train these networks in a second generation. This approach is more robust to the position of the visual object, but requires reliable bounding boxes. This procedure yielded more accurate results, and we used this procedure in the paper. It should be noted that even when using the full image, the networks were able to efficiently categorize the 1000 labels, demonstrating the specific robustness of the retinotopic referential, although the localization of these labels was not optimal.

Since the network is asked to perform inference over 1000 labels during training, we implemented loss using the cross-entropy loss from the PYTORCH library. We used the stochastic gradient descent (SGD) optimizer from the PYTORCH library and validated parameters such as batch size, learning rate, and momentum by performing a sweep of these parameters for each network. During the sweep, we varied each of these parameters over a given range while leaving the others at their default values for 1 epoch on 10% of the entire IMAGENET training dataset. We chose the parameter values that gave the best average accuracy on the validation set: batch size = 80, learning rate = 0.001, momentum = 0.9. We retrained the networks during 5 epochs of the full training dataset, keeping all learning parameters identical.

Localization tools for CNNs

One widely recognized technique for evaluating the performance of Convolutional Neural Networks (CNNs) in localization tasks is the Class Activation Mapping method (CAM). CAM operates by analyzing the CNN's output in relation to the targeted class, assigning weights to activations in each spatial feature map. This process generates a heat map, highlighting significant areas of the image based on their contribution to the prediction. Building upon the foundation of CAM, several derivative methods have emerged, including Grad-CAM [**?**], Score-CAM [36], and Opti-CAM [37].

In an effort to fairly quantify the contribution of each method, many quantification techniques have been developed. Here, we select some of them to compare the models using the retinotopic or Cartesian referential. Energy-Based Pointing Game: The localization is a success when the peak activation of the heatmap of a given label is located inside the ground true mask (or box). Mean activation IN: Mean activation of the heatmap of a given label inside the ground true mask. Mean activation OUT: Mean activation of the heatmap of a given label outside the ground true mask. Mean activation Ratio: Ratio of the activation inside and outside the box, the higher, the greater is the heatmap to indicate efficiently the position for a given label. Intersection over Union (IoU): Ratio between the area of overlap of the heat map and the ground truth and the area of union of the heat map and the ground truth. Peak-IoU and Peak-Threshold: With a modulation of a threshold on the heatmap, the Peak-IoU is the maximal IoU value reached at the Peak-Threshold.

Empirical results

Training on Transformed Images

We fine-tuned the pre-trained ResNet networks on the ImageNet dataset with retinotopic transformed images. We compared the performance of the networks trained on the original dataset (with a Cartesian mapping) with the performance of our retrained network on the transformed dataset defined in retinotopic space. We found that the networks retrained on the transformed images had similar categorization accuracy as the network pre-trained on regular images (see Figure 2-A & B). This is surprising given that the networks used in this retraining process were previously trained on regular images and that images with a log-polar transformation show a high degree of distortion, in particular a compression of visual information around the fixation point and a degradation of textures in the periphery, see Figure 1. One hypothesis is that the degradation of texture during the frame of reference change may cause the network to rely on shape rather than texture. However, it remains to be investigated whether there is a qualitative change in the features underlying the categorization.

Robustness of Convolutional Neural Networks

Deep learning algorithms have made tremendous progress in the last few years. For some visual recognition tasks, such as the ImageNet challenge [31], they have achieved superhuman accuracy. However, one limitation that remains is their vulnerability to adversarial attacks. Studies have shown that these learned models can be fooled by imperceptible modifications to images that are undetectable to humans [38]. Yet these small distortions cause the algorithms to incorrectly classify examples with high confidence [39]. This vulnerability makes deep networks unstable and unsuitable for use in safety-critical domains like medicine, autonomous vehicles, or other life-or-death situations. Before deep learning can be relied upon for such applications, researchers must find ways to make these models more resilient to adversarial examples and introduce human-level robustness to ensure that mistakes do not have dangerous consequences in the real world.

Attacking classical CNNs with a Geometrical Rotation

While previous adversarial attacks have focused on imperceptible pixel perturbations, such approaches introduce changes in many independent dimensions without reflecting the perturbations that occur in natural environments. We present a more ecologically

valid attack: controlled image rotations. Just as real-world objects appear in different orientations, we evaluate the robustness of models to rotations of test inputs. Compared to modifying independent pixels, rotation represents a coherent, whole-image transformation controlled by a single parameter, the rotation angle. Our goal is to investigate attacks that simulate realistic rotations in order to understand vulnerabilities and inform robust algorithm development.

To evaluate such a rotation-based attack, we examined popular off-the-shelf CNNs pre-trained on standard, large image datasets. For each test image, we first computed the unperturbed baseline accuracy. We then systematically rotated the images and tracked the change in model loss. To perform such an attack on a model m, we follow this simple procedure. Given an image I and the output of the model $\boldsymbol{p} = m(I)$, which returns a probability vector over $K = 1000$ classes, the loss function $\mathcal{L}$ is defined as the cross-entropy between the predicted probability vector and the ground truth label y that we denote as $\mathcal{L}(m(I), y)$. This is the loss that was minimized during the training procedure using gradient descent. By denoting a rotation of the image by an angle θ as $\text{rot}(I; \theta)$, we define the rotation-based attack as the following heuristic on each image of the dataset:

$$\bar{\theta} = \arg\max_{\theta} \mathcal{L}(m(\text{rot}(I;\theta)), y) \tag{3}$$

$$\hat{y} = \arg\max_{k}(\bar{p}_k) \text{ with } \bar{\boldsymbol{p}} = m(\text{rot}(I;\bar{\theta})) \tag{4}$$

More specifically, our approach consists first to choose the rotation angle that maximizes the loss, then to infer the most probable label for that particular angle. As a result, we can compute the concordance between the predicted label $\hat{y}$ for the image rotated at the angle $\bar{\theta}$ with the worst loss with respect to the ground truth label y. Using this procedure, we calculated the overall accuracy on the entire test set, quantifying the network's brittleness to natural image rotations.

Our experiments showed that while ResNet101 reached nominally 74% baseline accuracy, rotation attacks significantly reduced this performance. The application of the maximally deceptive rotation to each image reduced the accuracy to 39%. These results were confirmed by performing the same procedure on the Animal 10k [32] dataset (see Appendix Figure 7). In addition, we computed for any given rotation the average accuracy over the test dataset, showing that this average accuracy dropped sharply compared to baseline accuracy, confirming that this simple geometrical transformation mislead the networks. The decline was steepest between two cardinal rotations, demonstrating limited rotational invariance compared to humans [40].

We tested each network either with raw images or with retinotopic mapping (Cartesian or Retinotopic). We observed a symmetry with respect to the horizontal flips and only show the positive angles here. This shows that Cartesian network has a degraded performance compared to ResNet101, and notably that rotating images may have an adversarial effect on categorization performance, an effect which is less observed for the retinotopic network.

Noteworthy, these results show that while the Cartesian network retrained and tested on regular images showed some degradation for different rotations, the categorization results were much more invariant for the network including a retinotopic mapping (see Figure 2-C). This phenomenon is a consequence of the translation invariance imposed by the structure of the CNNs. Applied to retinotopic mapping, this translation invariance in retinotopic space is translated into rotation and zoom invariance in visual space [19].

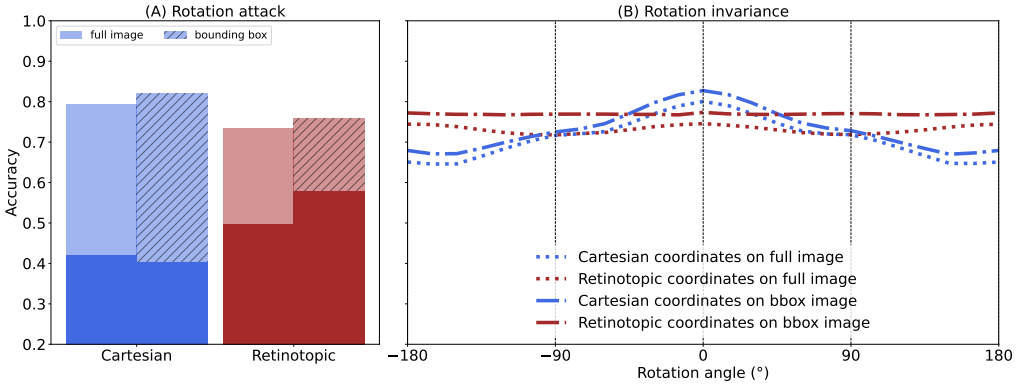

Fig 2. To illustrate the brittleness of classical CNNS, we show the effect of rotation-based attack by plotting accuracy over a sample of 1000 images from the IMAGENET [31] dataset, for **(A)** For all the networks, we plot the accuracy over the raw data set (without rotation, in light color), or on the image rotated at the angle $\bar{\theta}$ with the worst loss (rotation-based attack, in full color). **(B)** Average accuracy over a sample of 1500 images from the IMAGENET [31] dataset, shown for both retrained and pre-trained networks with different input image rotations. The rotation is applied around the fixation point with an angle ranging from $-180°$ to $+180°$ (in steps of $15°$).

Visual Object Localization : Protocol

Deep convolutional networks output a vector of real numbers, often interpreted as the log-probability on the label space ("logit"), as implemented through cross-entropy loss. Applying the softmax function allows the output to be interpreted as a probability vector, assigning a probability of presence to each of the 1000 labels (likelihood score). This allows us to make a binary decision ("present" or "not present"), e.g. by selecting the label corresponding to the highest likelihood. In our setting, we can also take different views of a large image and compute the likelihood for each of them, allowing us to compare which view provides the best likelihood. Views can consist, for example, of focusing on regions of the image centered on different fixation points, with the fixation points aligned on a regular grid in visual space.

We used two parameters to define these maps: the first parameter is the resolution of the fixation point grid. The second is the size of the samples clipped at each of these positions, defined as the ratio of the log-polar grid radius of the input to the total input size or Cartesian grid size, since the grid is a square for Cartesian samples (for an illustration of this simple process, see Appendix Figure 8).

This sample is then transformed or not by the retinotopic mapping before being used as input for the corresponding network, see Appendix Figure 8-B & D. Conveniently, a collection of samples for different fixation points can be processed as a single batch. This protocol defines a likelihood map for any given network as the likelihood of categorizing the presence of a label of interest inferred at regularly spaced fixation points in the image. As can be observed qualitatively in the example image, the Cartesian representation induces uniform activation across the whole image, while the retinotopic mapping induces differential activation inside and outside the object of interest.

Visual Object Localization : Quantification

We sub sampled each image with a grid of equidistant viewpoints at a resolution of 11×11. At each viewpoint, the largest possible sample is cropped. Thus a minimum sample with a 30% ratio at the border and the whole image at the center. From the

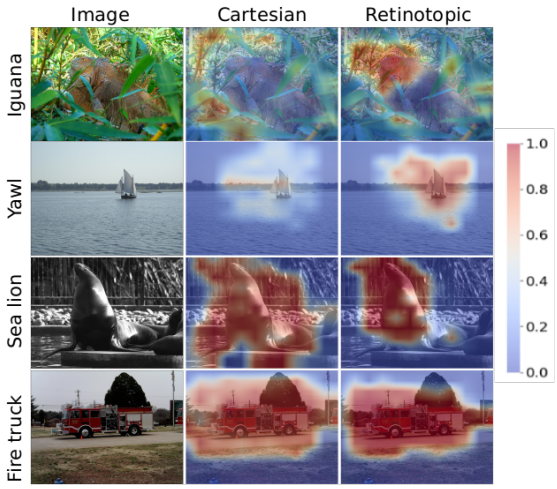

Fig 3. Likelihood maps computed on a representative image using 11×11 points of fixation with networks, trained and tested either on regular images (**Left**) or on images mapped on the retinopic space using a log-polar grid (**Right**).

Cartesian or retinotopic reference frame, this sample is then resized to a 224×224 resolution to match the optimal size for the CNN before processing. We selected some key metrics to compare the retinotopic or Cartesian referential. An indicator of the correct map activation position is the Energy-Based Pointing Game accuracy [41] where the localization is a success when the peak activation of the heat map of a given label is located inside the ground truth bounding box. In addition, we chose to track the ratio of the activation inside versus outside the box: the higher the ratio and the greater is the contrast of the heat map.

Using the heat map extracted from the key points of the Animal 10k [32] dataset as ground truth, "in" represents coordinates inside an animal (and "out" coordinates outside an animal, see Appendix Figure 9). For each point in the 11×11 grid, a likelihood value is obtained (probability of the presence of an animal). Next, we calculate the average likelihood for all points that are inside the zone corresponding to the animal (likelihood "in"), as well as the average likelihood for the zone that does not contain the animal (likelihood "out"). Next, we compare the values obtained in the "in" zone with those obtained in the "out" zone. A higher contrast indicates a better ability of the network to identify regions of interest in an image. If we consider a good categorization to be a high average probability on "in" coordinates (or a low probability on "out" coordinates), then in general networks using Retinotopic meshes tend to be more contrasted than networks using Cartesian meshes, which is more apparent in the case of ResNet101, see Table 1).

Despite the greater diversity of poses, environments, and intra-class variation in Animal 10k compared to ImageNet, we found that they accurately highlighted regions containing the target animals. This suggests that our models learned general object-level features rather than dataset-specific cues.

Accuracy after "saccades" protocol

In this part of the study, we focused on finding a label of interest in including a large number of fixation points per image. We applied this protocol to the 50,000 images in the validation set of the ImageNet data set. If we only stop at the best position

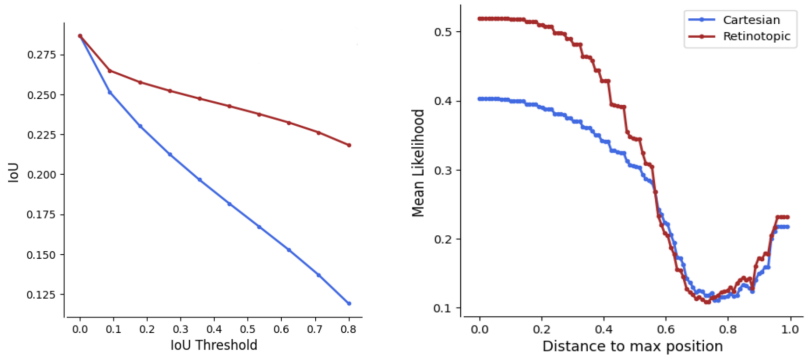

Fig 4. Quantifying the localization performance of CNNs

Table 1. Likelihood maps results for the ResNet networks and as computed on the ImageNet [31] dataset. In: Average probability for a fixation point in a position where an animal is present. Out: Average probability for a fixation point outside the position where an animal is present. Ratio: Average probability in positions where there is an animal on the average probability outside positions where there is an animal (In/Out). PG: In this context, the term "PG" is used to denote the degree of accuracy achieved in the Pointing Game. IoU-Peak: Represent the maximum value of the Iou reached by the network. IoU-T: In this context, "IoU-T" refers to the IoU Threshold, that is to say, the value at which the IoU reaches zero for each network. We highlight the mapping which reaches the best performances for all three conditions for each network.

	ResNet18		ResNet50		ResNet101	
	Cartesian	Retinotopic	Cartesian	Retinotopic	Cartesian	Retinotopic
In	**0.92**	0.76	0.73	**0.80**	0.73	**0.80**
Out	0.77	**0.49**	0.69	**0.54**	0.73	**0.80**
Ratio	1.19	**1.59**	1.06	**1.50**	0.73	**0.80**
PG	1.19	**1.59**	1.06	**1.50**	0.73	**0.80**
IoU-Peak	1.19	**1.59**	1.06	**1.50**	0.73	**0.80**
IoU-T	1.19	**1.59**	1.06	**1.50**	0.73	**0.80**

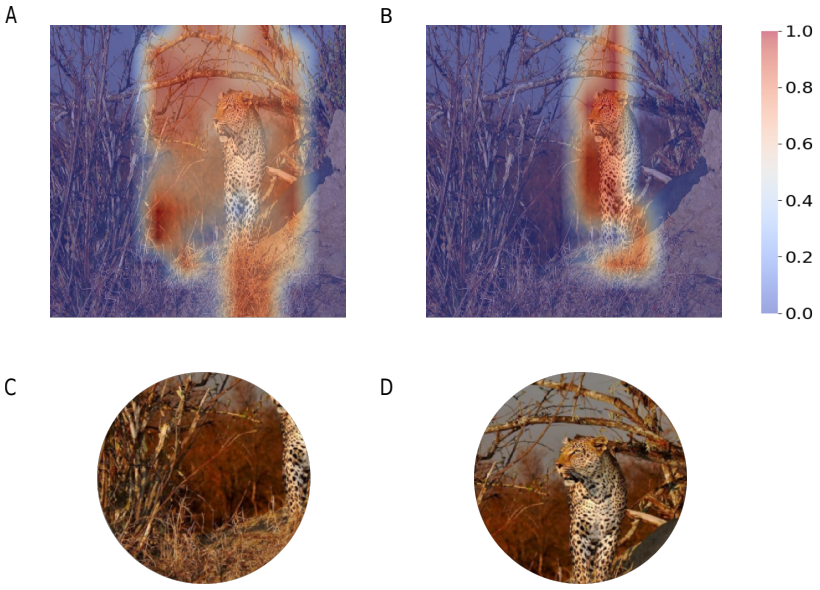

Fig 5. Sample saccade to maximum likelihood position

(Top 1), the performance of the networks is degraded compared to their accuracy without saccades, and the same is true for Top 5 (compared to the performance of Top 5 without saccades, not shown here). On the other hand, by adding a simple saccade selection strategy (Top Choice), we find that the accuracy of all networks exceeds their baseline level.

Task Dependence of Localization

We note that the likelihood maps generated by the model will vary depending on the label or class being predicted for a given input image. Specifically, the spatial distribution of high likelihood regions in the map differs depending on the label the model is trying to identify, or a class of objects defined by a set of labels. This suggests that the model learns to focus on different discriminative regions for different classes. The different likelihood maps provide insights into how the model's spatial attention changes according to the visual patterns that drive each classification decision. We tested this hypothesis on an example image by showing likelihood maps for the classes "animal", "dog", and "cat". As expected, the different regions are separated accordingly (see Figure 6).

The observation that likelihood maps vary by predicted class provides an exciting opportunity for future work. By modeling visual search as a path through these conditional spatial distributions, we may gain insight into how human gaze is affected during visual tasks. Attentional mechanisms have long been shown to guide eye movements in humans [4]. This work therefore presents promising opportunities to use deep networks as a means to unravel cognitive processes through computational modeling.

Discussion

Retina-Inspired Mapping Enhances the Robustness of CNNs

A first and main result of this study is to demonstrate the excellent ability of standard deep CNNs to deal with retinotopic inputs, even though this mapping represents a profound transformation of the visual inputs. The ResNet networks adapt easily to inputs where a large part of the image is highly compressed in the periphery and the spatial arrangement is significantly disturbed. The accuracy rates achieved with retinotopic inputs are equivalent to those of the original models.

In addition, the log-polar transformation has the advantage of better invariance to zoom and rotation. However, this invariance comes at the cost of reduced invariance to translations. For images that are not centered on the region of interest, one would have to shift the fixation point to the region of interest, similar to eye saccades.

The integration of a retinotopic mapping approach holds great promise for improving the efficiency and accuracy of image processing tasks. Our results are consistent with physiological data on ultrafast image categorization, which show that human accuracy in recognizing briefly flashed images of animals is robust to rotation [42]. The log-polar compression used in our approach allows a seamless extension to larger images without a significant increase in computational cost.

Relation to Pre-attentive Mechanisms

As a second result, the definition of likelihood maps based on scanning the visual scene at a limited number of fixation points allows us to gain insight into the specificities of retinotopic processing: this transformation provides a more focused view, thus better

separating the different elements of the image when focusing on its specific parts. As a result, it provides a proxy for the measurement of saliency with respect to the respective labels in the image.

In our case, it seems that the retinotopic mapping allows for a more precise localization of the category of interest, e.g., an animal, compared to off-the-shelf pre-trained networks using a Cartesian representation. It also gives us insight into the features on which our networks actually rely. Such information can be compared with physiological data [43], used to design better CNNs, and ultimately allow physiological tests to be proposed to further explore the features needed to classify a label of interest. In particular, by focusing on the point of fixation with the highest probability in the likelihood maps, we could consider refining the training of the network to our retinotopic mapping.

Introducing Eye Movements

Building on these observations, simulating human saccadic eye movement patterns during visual tasks provides an exciting opportunity to gain further insights into these mechanisms using such networks. A protocol that iteratively classifies image patches corresponding to foveated regions in a manner that mimics eye movements could reveal how network performance is spatially modulated across the visual field. Comparing classification accuracy under different saccade strategies, such as selecting the most uncertain or most likely location at each step, would provide valuable information about how network attention operates. This framework also allows the modeling of popular biological saccade strategies, allowing direct comparison with human visual search behavior.

Overall, implementing foveated classification with algorithmic saccades would provide a powerful method for validating existing attentional mechanisms in these networks, as well as inspiring new architectural innovations through embodied, task-driven visual attention modeling. This line of research should be particularly well-fitted for a dual-pathway model which is well adapted to infer ego-motions [44, 45]. Finally, the implementation of this robust categorization, coupled with refined localization of a label of interest and optimal saccade selection, could allow us to extend this study to a more complex task. One such task is visual search (i.e., the simultaneous localization and detection of a visual target), and the likelihood maps could provide the underlying pre-attentive mechanisms on which its effectiveness seems to depend. For more information, see **??**.

Acknowledgments

Authors received funding from the ANR project number ANR-20-CE23-0021 ("AgileNeuroBot") and from the french government under the France 2030 investment plan, as part of the Initiative d'Excellence d'Aix-Marseille Université – A*MIDEX grant number AMX-21-RID-025 "Polychronies". This work was granted access to the HPC resources of Aix-Marseille Université financed by the project Equip@Meso (ANR-10-EQPX-29-01) of the program "Investissements d'Avenir" supervised by the Agence Nationale de la Recherche. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Polyak SL. The Retina.;.

2. Mitkus M, Potier S, Martin GR, Duriez O, Kelber A. Raptor Vision. In: Oxford Research Encyclopedia of Neuroscience;.Available from: https://oxfordre.com/neuroscience/display/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-232?ref=PDF.

3. Collin SP. Scene through the Eyes of an Apex Predator: A Comparative Analysis of the Shark Visual System;101(5):624–640. doi:10.1111/cxo.12823.

4. Yarbus A. Eye Movements during the Examination of Complicated Objects;6(2):52–56.

5. Noton D, Stark L. Scanpaths in Eye Movements during Pattern Perception;171(3968):308–311. doi:10.1126/science.171.3968.308.

6. Anstis SM. A Chart Demonstrating Variations in Acuity with Retinal Position;14(7):589–592. doi:10.1016/0042-6989(74)90049-2.

7. Tootell RBH, Hadjikhani N, Hall EK, Marrett S, Vanduffel W, Vaughan JT, et al. The Retinotopy of Visual Spatial Attention;21(6):1409–1422. doi:10.1016/S0896-6273(00)80659-5.

8. Dougherty RF, Koch VM, Brewer AA, Fischer B, Modersitzki J, Wandell BA. Visual Field Representations and Locations of Visual Areas V1/2/3 in Human Visual Cortex;3(10):1. doi:10.1167/3.10.1.

9. Kaas JH. Topographic Maps Are Fundamental to Sensory Processing;44(2):107–112. doi:10.1016/S0361-9230(97)00094-4.

10. Weinberg RJ. Are Topographic Maps Fundamental to Sensory Processing?;44(2):113–116. doi:10.1016/S0361-9230(97)00095-6.

11. Lewis A, Garcia R, Zhaoping L. The Distribution of Visual Objects on the Retina: Connecting Eye Movements and Cone Distributions;3(11):21–21. doi:10.1167/3.11.21.

12. Lewis A, Garcia R, Zhaoping L. Understanding Cone Distributions from Saccadic Dynamics. Is Information Rate Maximised?;58:807–813. doi:10.1016/j.neucom.2004.01.131.

13. Laurent Itti CK. Computational Modelling of Visual Attention;.

14. Rothkopf CA, Ballard DH, Hayhoe MM. Task and Context Determine Where You Look;7(14):16. doi:10.1167/7.14.16.

15. Daucé E, Albiges P, Perrinet LU. A Dual Foveal-Peripheral Visual Processing Model Implements Efficient Saccade Selection;20(8):22–22. doi:10.1167/jov.20.8.22.

16. Daucé E, Perrinet L. Visual Search as Active Inference. In: Verbelen T, Lanillos P, Buckley CL, De Boom C, editors. Active Inference. Communications in Computer and Information Science. Springer International Publishing;. p. 165–178.

17. Dabane G, Perrinet LU, Daucé E. What You See Is What You Transform: Foveated Spatial Transformers as a Bio-Inspired Attention Mechanism. In: 2022 International Joint Conference on Neural Networks (IJCNN). IEEE;. p. 1–8.

18. Sandini G, Tagliasco V. An Anthropomorphic Retina-like Structure for Scene Analysis;14(4):365–372. doi:10.1016/0146-664X(80)90026-X.

19. Araujo H, Dias JM. An Introduction to the Log-Polar Mapping;(1):139–144. doi:10.1109/CYBVIS.1996.629454.

20. Lim FL, West GAW, Venkatesh S. Use of Log Polar Space for Foveation and Feature Recognition;144(6):323. doi:10.1049/ip-vis:19971457.

21. Wolfe JM. Guided Search 2.0 a Revised Model of Visual Search;1:202–238. doi:10.3758/BF03200774.

22. Najemnik J, Geisler WS. Optimal Eye Movement Strategies in Visual Search;434(7031):387–391. doi:10.1038/nature03390.

23. Daucé E. Active Fovea-Based Vision through Computationally-Effective Model-Based Prediction;12:76. doi:10.3389/fnbot.2018.00076.

24. Carrasco M. Visual Attention: The Past 25 Years;51(13):1484–1525. doi:10.1016/j.visres.2011.04.012.

25. Sarvaiya JN, Patnaik S, Bombaywala S. Image Registration Using Log-Polar Transform and Phase Correlation. In: IEEE Region 10 Annual International Conference, Proceedings/TENCON;. p. 1–5.

26. Maiello G, Chessa M, Bex PJ, Solari F. Near-Optimal Combination of Disparity across a Log-Polar Scaled Visual Field;16(4):e1007699. doi:10.1371/journal.pcbi.1007699.

27. Javier Traver V, Bernardino A. A Review of Log-Polar Imaging for Visual Perception in Robotics;58(4):378–398. doi:10.1016/j.robot.2009.10.002.

28. Antonelli M, Igual FD, Ramos F, Traver VJ. Speeding up the Log-Polar Transform with Inexpensive Parallel Hardware: Graphics Units and Multi-Core Architectures;10(3):533–550. doi:10.1007/s11554-012-0281-6.

29. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, family=Buc pu given=F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.;. p. 8024–8035.

30. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial Transformer Networks;doi:10.48550/arXiv.1506.02025.

31. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge;115:211–252. doi:10.1007/s11263-015-0816-y.

32. Yu H, Xu Y, Zhang J, Zhao W, Guan Z, Tao D. AP-10K: A Benchmark for Animal Pose Estimation in the Wild;.

33. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition;.

34. Jérémie JN, Perrinet LU. Ultrafast Image Categorization in Biology and Neural Models;7(2):29. doi:10.3390/vision7020029.

35. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition;doi:10.1109/CVPR.2016.90.

36. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, et al. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks;. Available from: http://arxiv.org/abs/1910.01279.

37. Zhang H, Torres F, Sicre R, Avrithis Y, Ayache S. Opti-CAM: Optimizing Saliency Maps for Interpretability;. Available from: http://arxiv.org/abs/2301.07002.

38. Huang S, Papernot N, Goodfellow I, Duan Y, Abbeel P. Adversarial Attacks on Neural Network Policies;. Available from: http://arxiv.org/abs/1702.02284.

39. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing Properties of Neural Networks;.

40. Rousselet GA, Macé MJM, Fabre-Thorpe M. Is It an Animal? Is It a Human Face? Fast Processing in Upright and Inverted Natural Scenes.;3:440–455. doi:10.1167/3.6.5.

41. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization;128(2):336–359. doi:10.1007/s11263-019-01228-7.

42. Fabre-Thorpe M. The Characteristics and Limits of Rapid Visual Categorization;2. doi:10.3389/fpsyg.2011.00243.

43. Crouzet SM. What Are the Visual Features Underlying Rapid Object Recognition?;2. doi:10.3389/fpsyg.2011.00326.

44. Bakhtiari S, Mineault P, Lillicrap T, Pack CC, Richards BA. type [;]Available from: http://biorxiv.org/lookup/doi/10.1101/2021.06.18.448989.

45. Mineault PJ, Bakhtiari S, Richards BA, Pack CC. type [;]Available from: http://biorxiv.org/lookup/doi/10.1101/2021.07.09.451701.

Appendix

Robustness to rotations on the Animal 10k dataset

To evaluate the robustness of our models to geometric transformations such as rotations beyond the ImageNet dataset (see Figure 2), we performed additional experiments on the Animal 10k dataset. This dataset contains 10,000 images from a variety of animal species and presents a challenging classification task due to high intra-class variability in pose, size, orientation, and environment.

As with ImageNet, we first obtained baseline classification accuracy on Animal 10k images without perturbation. We then systematically rotated each test image through the same range of rotation angles. At each rotation value, we computed the average accuracy over the entire test set in order to characterize performance as a function of viewing angle.

The results on Animal 10k broadly mirror those on ImageNet, with standard CNNs showing a sharp decline in accuracy for even small rotations away from the trained orientation, particularly away from the cardinal angles (see Figure 7). However, models incorporating our retinotopic preprocessing approach maintain significantly higher performance over the full range of orientations. This demonstrates that the benefits of rotational robustness generalize beyond the original dataset and image characteristics.

Testing on Animal 10k, with its greater pose diversity and difficulty, reinforces the conclusion that our method improves invariance to natural, ecologically valid geometric transformations beyond what standard CNNs alone can achieve.

Creating Multiple Views for Likelihood Maps

To create likelihood maps to test our trained models, we generated batches of transformed input views for each full-scale test image. Specifically, we created a grid of fixation points as a grid of linearly spaced points on the horizontal and vertical axes.

We then cropped a sample of the image centered on each fixation point, with the size of the sample defined by the ratio of the log-polar grid radius to the total input size (or the Cartesian grid size, since the grid is a square for Cartesian samples). We then transformed each sample with or without the retinotopic mapping and ran the batch of transformed images through the model to obtain a likelihood map for that image-label pair(see Figure 9). We repeated this process for each image-label pair in the test set (see Figure 10).

Ground Truth Heat Maps

To quantify how accurately our models are able to localize objects within images, we generated ground truth heatmap labels for the datasets. For each image from Animal 10k containing a set of keypoints, we created a Gaussian heatmap centered on those points, with the peak value set to 1 and values falling off with a standard deviation proportional to object size. This results in heatmap labels between 0 and 1 that capture the true spatial extent and location of the target object within each image (see Figure 9).

Fig 6. Likelihood maps generated by RESNET101 networks trained and tested on retinotopic space using the log-polar grid. Searching for the (A) "animal", (B) "dog", or (C) "cat" label in an image.

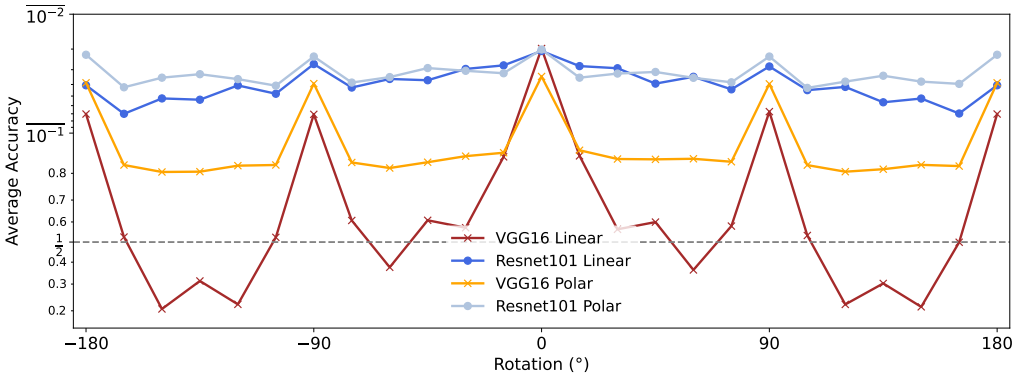

Fig 7. Average accuracy over the ANIMAL 10K [32] dataset, shown for both retrained and pre-trained networks with different input image rotations. The rotation is applied around the fixation point with an angle ranging from −180° to +180° (in steps of 15°). We tested each network either with raw images in Cartesian coordinates or with retinotopic mapping. The dotted line represents chance level.

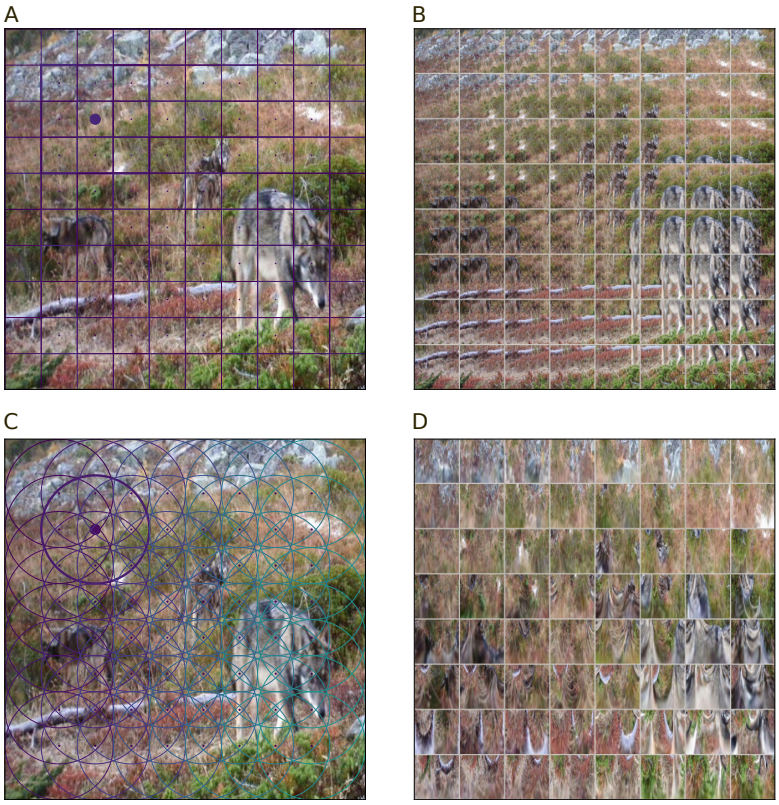

Fig 8. Generating different views of a single image to compute likelihood maps. **(A)** For the networks with Cartesian inputs, we used a regular grid of 8×8 fixation points, which allows samples to be cropped, highlighting a particular view. As shown in **(B)**, this produces a batch of images that can be used to generate likelihood maps. **(C)**) We used a similar grid to generate batches of retinotopic inputs, as shown in **(D)**). In **(B)** & **(D)** each sample corresponds to 30% of the input.

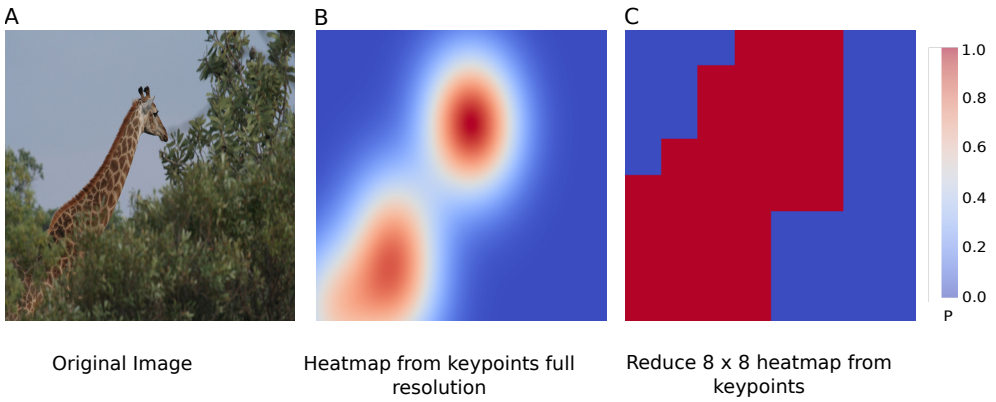

Fig 9. (A) The original image of the Animal 10k dataset. (B) A heat map constructed by fitting Gaussians to the key points of the Animal 10k dataset. (C) The heat map constructed in (B) is normalized and reduced to a resolution of 8×8 to be used as ground truth when evaluating the heat map. A threshold (0.2) is applied to reduce the heat map field to the assumed contour of the animal.

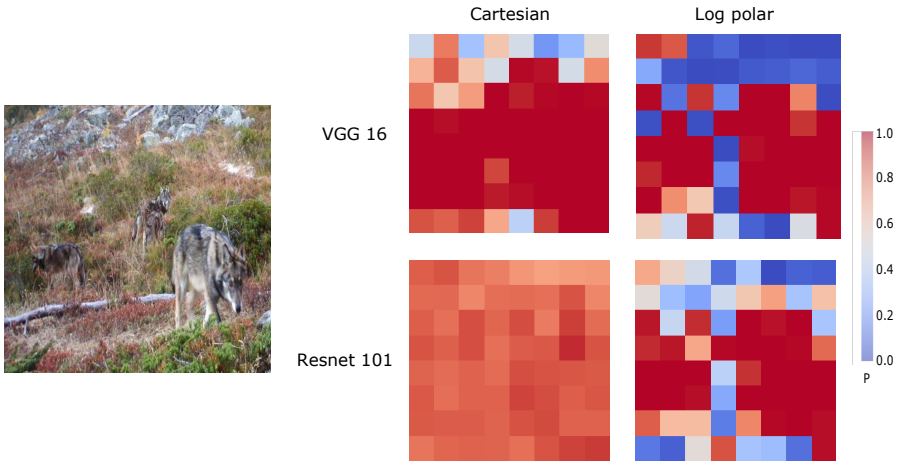

Fig 10. Likelihood maps computed on a representative image using 8×8 points of fixation, (as computed on the whole ANIMAL 10K [32] dataset), (**top**) with a VGG16, or (**bottom**) a RESNET101) network, trained and tested on either regular images (**left**) or images mapped in retinopic space using a log-polar grid (**right**). Red predicts the presence of the label 'animal', blue its absence.