Foveated Retinotopy Improves Classification and Localization in Convolutional Neural Networks

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 École Centrale Méditerranée, Marseille, France

* laurent.perrinet@univ-amu.fr

Abstract

From a falcon detecting prey to humans recognizing faces, many species exhibit extraordinary rapid visual abilities in both localization and classification. These are made possible by a specialized retinal region called the fovea, which provides high acuity at the center of vision while maintaining lower resolution in the periphery. This distinctive spatial organization, preserved along the early visual pathway through retinotopic mapping, is fundamental to biological vision, yet remains largely unexplored in machine learning. Our study investigates how incorporating foveated retinotopy may benefit deep convolutional neural networks (CNNs) in image classification tasks. By adding a foveated retinotopic transformation to the input layer of standard off-the-shelf models and re-training them, we maintain comparable classification accuracy while enhancing the network's robustness to scale and rotational perturbations. Although this architectural modification introduces increased sensitivity to shifts of the fixation point, we demonstrate how this apparent limitation becomes advantageous: variations in classification probabilities across different gaze positions serve as effective indicators for object localization. Our findings suggest that foveated retinotopic mapping encodes implicit knowledge about visual object geometry, offering an efficient solution to the visual search problem, a capability crucial for many living species.

Author summary

Many animals, among them falcons, hummingbirds, sharks or humans, have a remarkable structure in their eyes called the fovea – a small region that provides sharp, detailed vision at the center of our gaze, like a natural zoom lens. When we want to see something clearly, we move our eyes to focus on this high-resolution area directly at it. This efficient design has evolved over millions of years, yet modern artificial intelligence systems process images very differently, examining each pixel with equal weight. Our study asks: what if we teach computers to see more like animals do? By adding a fovea-like mechanism to artificial neural networks, we create systems that not only maintain their ability to recognize objects but also gain new strengths. Like animals, these modified networks become better at handling images from different distances and angles. They also develop an appealing new ability: by sampling different viewpoints, they can effectively search for and locate objects of interest – much as we scan a scene with our eyes. This work suggests that borrowing designs from nature can make artificial vision systems more efficient and capable.

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

For predators such as birds of prey [1] or sharks [2], efficient detection of prey is essential for survival. In a broader context, visual search is defined as the essential function by which organisms examine their visual environment to locate and identify objects of interest. This can apply to predators such as falcons, sharks, or humans, but also to species such as leaf-eating monkeys such as howler monkeys, fruit bats (which feed primarily on fruit and nectar), nectar-feeding birds such as hummingbirds, or seed-eating birds such as parrots. Most of these species have a visual organization that exhibits an inhomogeneous representation of the visual field which is denser around one center, commonly referred to as foveated retinotopy [3]. Visual search is thought to be largely based on this organization.

Foveated Retinotopy: A non-linear, radial organization

While many species have a rather uniform retinotopy which gives equal weights in all direction from the visual field (notably preys such as rabbits), the mapping of the visual field in humans and many mammals is characterized by a foveated retinotopy, with a substantial disparity in resolution between a central area and the peripheral regions. Its center defines the point of fixation. Already at the retina, the mapping is mostly radial and the density of photoreceptors decreases exponentially with eccentricity [4], defining a so called *log-polar* mapping. This radial organization of the retina, with highest acuity in the center and decreasing acuity in the periphery, is largely maintained across the different stages of visual processing in the brain, as shown in the human cortex by functional magnetic resonance imaging (fMRI) to map visual responses [5,6]. A first study by Tootell et al [5] focused on the phenomenon of spatial attention and sought to elucidate the response exhibited when a bar stimulus is presented in a subdivided visual field. As a result, a foveated retinotopic mapping of the visual field is observed in the early visual areas, specifically V1, V2, and V3. Dougerthy et al [6] then worked out the representation in the same areas when stimuli varying in eccentricity or angular direction were projected in direct association with the log-polar reference. These studies showed how different regions of visual cortex are activated depending on the position of stimuli in the visual field, confirming previously proposed models.

The precise evolutionary advantage of this retinotopic organization is still under debate. For example, it is thought to facilitate efficient parallel processing of spatial features in both primates [7] and humans [8]. It may also facilitate connections within and between processing areas that respect the geometry of the sensory epithelium and minimize global wiring length [9]. However, it has also been suggested that this is merely an artifact of the scaffolding that operates during development [10]. Another hypothesis is that foveated retinotopy implements prior knowledge of visual object geometry with respect to the species' range of behaviors and allows efficient use of resources by prioritizing high-resolution vision in areas where it matters most (i.e., the center of gaze) while sacrificing some resolution in peripheral vision. Eagles or bottlenose dolphins are remarkable in this regard because they have a dual fovea, one for fine lateral vision and one for high acuity straight ahead vision. This suggests that each retinotopy provides an efficient solution to the visual search problem for the given species within its respective ecological niche. More generally, this raises the problem of how the visual system builds an egocentric model of local space from the parcellated retinal input [11].
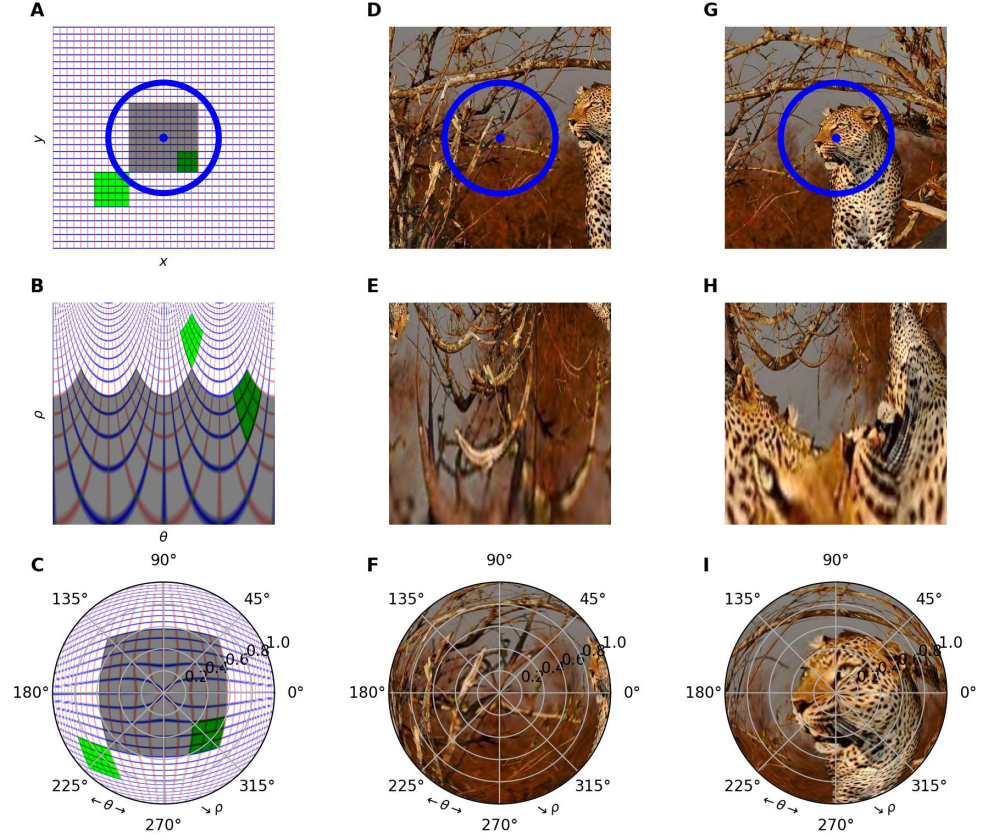
Fig 1. We illustrate the process of mapping input images defined in Cartesian coordinates to a foveated retinotopic space using a log-polar transformation. The fixation point is marked by a blue disk and the approximate area of the fovea by a blue circle. In **(A)**, the input image is defined as a regular 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 a fine-grained 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. The third row **(G, H, I)** illustrates the reconstruction of the image in question, exhibiting a over-representation around the point of fixation in the Cartesian reference frame. When the transformation is applied to a natural image, as shown in **(B, C)**, there is a noticeable compression of information in the periphery in the log-polar referential (see **(E, F, H, I)**). Also, this representation is highly dependent on the fixation point, as indicated by the shift shown in **(C, F, I)** when the fixation point is moved to the right and up.

Eye movements and the sequential analysis of the visual scene

A natural question is what computational advantages these retinotopic visual inputs confer on information processing. Numerous hypotheses have been proposed regarding the role of this nonuniform visual field mapping. One primary explanation is that foveal input facilitates 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. More generally, the fovea is 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, ...), thus maximizing access to visual information for these objects. In particular, when we visually explore a scene, our eyes perform saccades (in humans, 2 to 3 times per second on average, corresponding to about 150, 000 movements per day), rapid eye movements that move the fovea to successive points of interest (the typical duration of a saccade is very short, about 200 milliseconds). In between these saccades, fixations allow precise processing of visual information in a short period of time averaging 200 to 300 milliseconds. This alternation between periods of saccades and fixations forms the basis of our scan path and influences how we perceive and interact with our environment. An interesting hypothesis is that the distribution of photoreceptors in foveated retinotopy reflects the probability that an object of interest is located at the center of the retina, given the dynamic parameters of the oculomotor system [12, 13], thus allowing efficient visual exploration.

To better understand how visual search works in humans, it is necessary to study the underlying mechanisms that control our eye movements and visual attention. It is influenced by several factors, including the saliency of objects [14]. The systematic study of visual search began with the pioneering work of Yarbus [15] and Noton et al. [16], who introduced the concept of a "scan path" as the trajectory of eye movements produced during visual exploration. In the original experiment conducted by Yarbus, participants were presented with a series of different exploration targets applied to realistic paintings or photographs. The results showed that the scan paths that participants followed were not random, but had a degree of structure that was consistent with their individual goals. Noton et al. further improved the methods for recording and analyzing these movements by developing an approach that involves non-invasive monocular measurement of eye movements using the diffuse scleral reflection technique, allowing ecological observation of eye movements during recognition tasks [16]. Since then, eye-tracking studies have shown that our eyes follow predictable patterns that maximize the efficiency of visual information acquisition [17]. Recent studies have shown that this combination of saccades and foveal input, coupled with an effective point-of-interest detection mechanism, significantly enhances visual acuity [18–20] and supports functions such as the integration of local feature analysis into global perceptual representations. This method provides a promising basis for further investigation of eye movements and the function of foveated retinotopy.

The log-polar model in computer vision

These observations have led to the development of novel models of artificial vision using foveated retinotopy. Sandini et al [21] were among the first to develop such 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 as a function of eccentricity. It already demonstrates the contribution of this type of transformation to the compression of visual information, emphasizing the visual information in the center and shrinking visual information in the periphery. The most widely used model to represent this topography is the log-polar model, described in detail by Arujo et al [22]. It is organized around two polar coordinates: azimuth (angle) and eccentricity (distance

from center) on a logarithmic scale. A log-polar sensor would have a resolution (pixel density) that decays exponentially with distance from the center of the image. Recent developments in neuromorphic computing have also shown that a foveated transformation can be implemented in hardware, allowing real-time processing of visual information [23].

The log-polar transformation has important consequences for the representation of the geometry of objects in the image. It produces important distortions of the original image, with a strong enlargement of the objects present in the center and a shrinking of the peripheral ones (see figure 1). In addition, the changes produced by movements of the visual sensor, such as translation, rotation, or zoom, have different consequences in a log-polar reference frame. Any movement of the visual sensor that results in a shift of the center of fixation profoundly alters the distribution of pixels, enlarging the central region and blurring the visual information at the periphery (see Figure 1-E&F). Other camera movements have a more limited effect. In particular, any rotation of the sensor around the central fixation point results in a translation in polar space. Similarly, zooming in or out has the effect of a simple translation in the polar dimension [22]. The log-polar mapping transforms the rotations and zooms into two independent (orthogonal) displacements, since a rotation only changes the azimuth angle θ, while a zoom affects the logarithmic eccentricity ρ on the radial axis. This corresponds to a unique and important property - zooms and rotations in Cartesian space become vertical and horizontal shifts, respectively, in log-polar space (see the green square in Figure 1-A,D&G).

The log-polar retinotopic mapping was successfully applied in computer vision, especially for template matching [22, 24–26] or robotics [27, 28]. However, implementing visual processing with retinotopic sensors poses significant challenges in terms of information retrieval: Due to the low peripheral resolution, it is difficult to estimate in advance (before an eye movement is made) which part of the image contains the most relevant information. When analyzing a scene, the eye must infer 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 deviate significantly from the average statistics of the scene). This is the approach proposed, for example, by Itti and Koch [14]. Conversely, top-down approaches use prior knowledge about visual content to direct attention to potentially interesting regions that have not yet been explored [29]. Only a few works exploit this principle of "predictive coding" in the case of retinotopic sensors; see for example [30, 31].

Recent studies have also explored the potential of foveal retinotopy; Cao et al [32] designed a new LPNet architecture that includes an internal layer using the log-polar reference. They showed that rotation invariance during categorization can be improved by adding a layer that encodes the input image in the log-polar reference. Da Costa et al [33] used a CornetZ to investigate the reorganization of the intermediate layer when training the network with retinal ganglion cell (RGC) sampling, a method that produces Cartesian images that mimic foveal retinotopy. It is shown that the use of such inputs has an impact on CNN receptive field mapping. Finally, also using a sampling method, Lukanov et al [34] shows that this non-uniform distribution may be beneficial when considering the attentional mechanism for implementing saccades, as it could improve the categorization accuracy of the networks. It seems that the community is increasingly interested in foveal retinotopy and its benefits for the accuracy and robustness of CNNs. These promising results pave the way to a better understanding of this organization, but they overlook an important point: they all rely on a Cartesian reconstruction of the foveal retinal distribution instead of focusing on the log-polar reference frame.

Convolutional neural networks and translational invariance

Deep learning algorithms have made tremendous progress in recent years. For some visual recognition tasks, such as the IMAGENET challenge [35], convolutional neural networks (CNNs) have made huge strides in computer vision, so much so that they have now surpassed human accuracy in this task [36]. CNNs are particularly well suited to image processing tasks because they allow the operations performed by a neural network to be factorized, while limiting the number of parameters dedicated to each processing layer. Convolutional kernels are adaptive and are trained on many examples for image classification. Although CNNs are trained according to a plasticity rule that is not biologically plausible [37], they nevertheless have properties that are close to some exhibited by biological visual systems [38], in particular (i) massively parallel processing and (ii) the ability to increase the size of the receptive field (the size of a region of the original image) as a function of layer depth. In addition, the "pooling" operator also allows a CNN to be less sensitive to the position (or displacement) of objects in the image: we speak of translation invariance, which can be achieved by alternating convolutional layers and pooling layers with increasingly large receptive fields. CNNs also show striking similarities to the ability of humans to quickly classify images, such as whether an image contains an animal or not [39].

However, one limitation that remains is their vulnerability to adversarial attacks. Studies have shown that these learned models can be fooled by modifications to images that are imperceptible to humans [40]. These small distortions cause the algorithms to misclassify examples with high confidence [41]. This vulnerability makes deep networks unstable and unsuitable for use in safety-critical domains such as medicine, autonomous vehicles, or other life-or-death situations, and requires proper care to generate more robust responses [42]. 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. One emerging hypothesis is that insights from biological neuroscience would be critical to achieving this goal [43].

Despite the growing influence of biological and computational neuroscience on computer vision and machine learning, the contribution of foveated retinotopic mapping to visual processing remains scarce. In particular, it should be highly advantageous to exploit the translational invariance within the feature maps of CNNs as an advantage over the geometry of retinotopic mapping. Indeed, it has been shown that the architecture of any deep learning architecture acts as a "deep prior" to constrain the range of achievable representations, so that at one extreme one could use CNNs without learning [44]. In this context, the log-polar transformation could be seen as a way to constrain the range of transformations that are possible in the input space, thus providing a more efficient way to learn the underlying structure of the visual world.

Paper contributions

Here, we bridge biological and artificial vision by implementing a log-polar transform which we feed to the input of off-the-shelf deep convolutional neural networks (CNNs), creating a foveated mapping that mimics the spatial organization of biological retinas. Our key contributions are:

- We demonstrate that conventional CNNs, despite their effectiveness, exhibit significant vulnerability to basic geometric transformations such as rotations and zooms, highlighting a fundamental limitation in current architectures.

- We introduce a biologically-inspired foveated architecture by incorporating a log-polar transformation at the network input, embedding rotation and scale

invariance into the model's structure.

- We validate our approach using transfer learning on standard off-the-shelf architectures and benchmarks, demonstrating that the benefits of foveated vision can be achieved without compromising classification performance.

- We show how our foveated architecture leverages existing convolutional layers in novel ways to enhance robustness against geometric transformations, providing insights into the relationship between network architecture and invariance properties.

- We reveal how the log-polar transformation's sensitivity to fixation point creates an unexpected advantage: by analyzing the probability at the output of the network, we observe variations across multiple viewpoints, such that our system achieves effective object localization without additional training.

- We establish connections between our computational findings and biological vision, suggesting new directions for both machine learning architectures and our understanding of natural visual processing.

Results

Training on transformed images

We retrain pre-trained RESNET [36] networks on different variants of the IMAGENET-1K dataset [35] (see Methods "Datasets and Dataset Transformations"), either using a simple circular mask applied on the raw images (hereafter called the Cartesian dataset) or using log-polar transformed images (hereafter called the "retinotopic" dataset). Two variants of the training sets are also considered. In a first case (called the "regular" case), the mask or the log-polar transformed images are applied to the "regular" images. In a second case (called the "focus" case), the mask or the log-polar transformation is focused on the center of the *bounding boxes* surrounding the objects of interest and which are provided with the data set. The pre-trained networks are first re-trained on the "regular" version of the dataset, generating a first generation of networks, and then a second generation is trained by fine-tuning these networks on the "focus" dataset. We then compare the classification accuracy of the original and the different re-trained networks.

The baseline classification accuracy of the standard pre-trained RESNET network, which we will call the "raw" network (as it has no mask, retinotopic transformation nor training), averaged on the Cartesian test dataset is 81.7%. In comparison, our re-trained networks show respectively accuracies of 78.5% and 74.3% on the "regular" Cartesian and retinotopic datasets, and 82.5% and 77.4% on the "focus" Cartesian and retinotopic datasets. As we can observe, when fine-tuning the networks on the "focus" dataset, despite reducing the image resolution by cropping the image, the accuracy is improved by 4.0% and 3.1% respectively. This demonstrates that this cropping tends to suppress distractors in the periphery and enhances average accuracy.

Furthermore, our networks re-trained on the log-polar transformed images have a slightly lower categorization accuracy than those re-trained on Cartesian images (see Figure 2-A & Figure 3-A). This result was expected because the log-polar transformation discards fine-grained information in the periphery without increasing central resolution. In fact, this rather limited loss of accuracy is quite remarkable for such a massive loss of information. One reason could be a general "photographer's bias" for the "regular" images that tends to place the main object in the central region of the

image. It is confirmed by the increase of accuracy observed when centering the point of fixation in the "focus" dataset.

Robustness of CNNs

To what extent can these re-trained networks be relied upon in practical applications? Indeed, a persistent challenge for deep neural networks is their lack of reliability in adversarial situations. For example, it is well documented that a minor alteration to the inputs of these networks, if well-designed, may result in a significant decline in classification accuracy. In particular, a classical robustness test applied to deep networks, called an "adversarial attack," consists of perturbing independently the pixels in each image in order to maximize the error rate in the test phase (see the Methods "Attacking classical CNNs with a geometrical rotation"). However, these modifications are typically subtle and perceptually resemble identically distributed independent noise, rendering them unlikely to happen in natural conditions. In contrast, biological agents interacting with their environment undergo significant visual perturbations, including large scene pans and tilts due to head and body movements.

As real-world objects appear in different orientations, we assess the resilience of our re-trained networks to geometric alterations of their inputs e.g. rotations and zooms. In comparison to modifying individual pixels, a rotation or a zoom represents a coherent, whole-image transformation controlled by a single parameter, namely the rotation angle or scaling factor. We thus choose to investigate for each network a simple attack scenario that maximizes the loss for different rotation angles for each image individually and then evaluates the classification result for this "worst" angle. Then, we compute the accuracy averaged over a sample of $50,000$ images from the IMAGENET validation dataset.
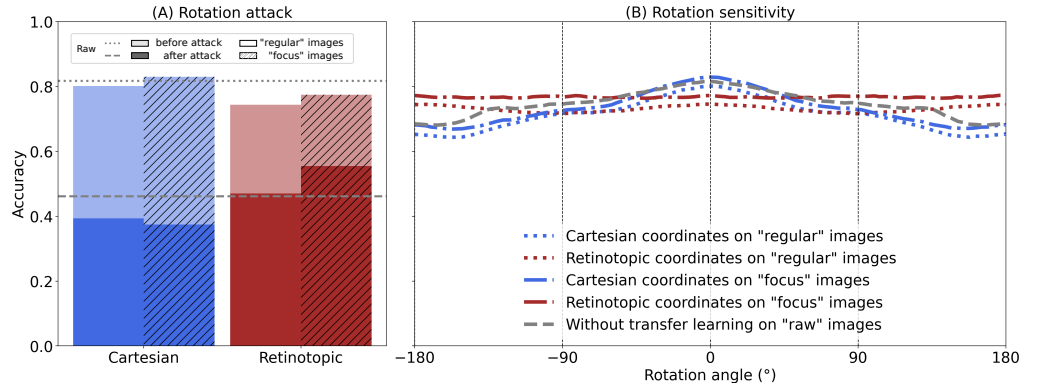

Fig 2. (A) For all the networks, we plot the accuracy averaged over the dataset without rotation (in light color), or for each image rotated at the angle $\bar{\theta}$ with the worst loss (rotation-based attack, in full color). No shading : regular dataset. Diagonal shading : focus dataset. Gray dashed lines: accuracy of the "raw" network (without any transformation nor training). **(B)** The average accuracy is shown for both Cartesian or retinotopic re-trained networks, and the "raw" network, with different image rotations. The rotation is applied around the central fixation point with an angle ranging from $-180°$ to $+180°$ (in steps of $15°$).

Our experiments show that while the original RESNET 101 achieves a nominal baseline accuracy of 81.7% on unperturbed images, a rotation attack significantly reduces the performance of the model. Applying the maximally deceptive rotation to each image reduces the average accuracy to 46.1% (see gray lines in Figure 2-A). This is
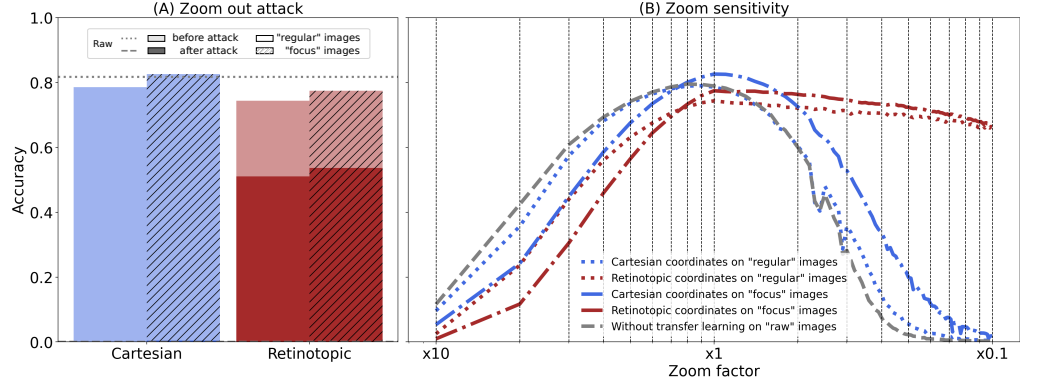
Fig 3. (A) For all networks, we plot the accuracy averaged over the unzoomed dataset (in light color), or for each image zoomed out to the scale with the worst loss (zoom out attack, in full color). The gray dashed lines represent the accuracy of the "raw" network (without re-training). **(B)** Average accuracy over a sample of 50,000 images from the IMAGENET validation dataset, shown for both re-trained and pre-trained networks with different input image zooms. The zoom is applied at the central fixation point with a zoom factor geometrically-spaced from ×10 to ×0.1.

also true for our Cartesian network, with accuracy dropping to 39.3% for the Cartesian network re-trained on "regular" images, and even to 37.4% for the network fine-tuned on the "focus" dataset. In contrast, the retinotopic networks show a lower sensitivity to rotation attacks, with an accuracy reduced to 47.0% when using the "regular" images, and only to 55.4% when using the "focus" dataset.

This difference between the two types of networks is even more manifest in Figure 2-B, where we compute the average accuracy over the test dataset for each single rotation angle. We first observe that the Cartesian networks show a decrease in accuracy with respect to the angle of rotation (with a symmetry with respect to horizontal flips), with a significant drop around 160°, "raw" and "focus" accuracies were degraded to 66.8% and 68.1%, i.e. a drop of 16.1% and 13.4% (64.3% in "regular", i.e. a drop of 15.8%). Strikingly, this effect is nearly absent in our retinotopic networks (see Figure 2-B), which show a flat (invariant) accuracy over the whole range of rotation angles, with a minimal degraded accuracy at 76.4%, i.e. a drop of 0.8% (71.6% in "regular", i.e. a drop of 2.9%). This marked difference can be interpreted as a consequence of the horizontal translation invariance found in classical CNNs. When applied to the retinotopic input space, this invariance transforms seamlessly into rotation invariance in visual space [22] (see [45] for a proof). Note that for all networks, no maximum accuracy exceeds the value obtained without rotation (i.e. 0° rotation angle) by more than 0.01%.

Analogous to rotation, zooming in and out is equivalent to a translation in log-polar space, and this property is expected to induce a similar invariance in the retinotopic networks. Similar experiments were therefore performed to test the effect of a zoom (see Figure 3), with a zoom ranging from ×10 to ×0.1, divided into "zoom-in" (×10 to ×1 range) and "zoom-out" (×1 to ×0.1 range). In Figure 3-A, we applied the zoom attack only for the zoom-out case, i.e. ratios between ×1 and ×0.1, because zoom-in attack scenarios always lead to the maximum zoom-in, where the accuracy approaches zero. The response of all Cartesian networks (i.e. "raw", "regular" or "focus") reach chance-level performance when submitted to a zoom-out attack. Only the retinotopic networks keep a discriminative capacity in this case (51.0% on "regular" images and 53.6% on "focus" images), illustrating the importance of foveal information in

categorization, especially when the peripheral information is scarce or deceptive.

Figure 3-B illustrates the average accuracy across the entire zoom range, offering a comprehensive perspective on the effects of zooming in and out across our different networks. Unlike the previous case, a pronounced asymmetry emerges in both scenarios. Specifically, the various Cartesian networks tested (both the original and re-trained versions) exhibit an approximate symmetry relative to the logarithmic scale of zoom-in and zoom-out factors. In contrast, the retinotopic network displays an asymmetric response, characterized by a slightly declining plateau in the zoom-out direction and a more pronounced accuracy loss in the zoom-in direction compared to the Cartesian networks. When zooming in, both the retinotopic and Cartesian networks exhibit heightened sensitivity, with mean accuracy dropping near chance levels at a ×10 zoom. Notably, the Cartesian networks achieve marginally higher accuracies in this scenario, suggesting a potential architectural advantage under extreme zoom-in conditions. However, this finding should be interpreted cautiously, as it likely stems from the loss of visual detail and the blurring effect inherent to applying extreme zoom levels on low-resolution original images.

In contrast to the case of rotations, our fine-tuning on the "focus" dataset has a nuanced effect, as evidenced by its impact on the zoom optimal point. For the Cartesian network, the zoom corresponding to maximum accuracy is around ×2 in both its pre-trained and re-trained states. However, fine-tuning on the "focus" dataset shifts this optimal zoom to ×1. In comparison, the optimal zoom for the retinotopic network remains consistently at ×1, regardless of evaluation on "regular" or "focus" data. From an ecological perspective, zoom-out invariance is likely one of the most advantageous features in natural vision. This capability facilitates the detection of predators or prey at a distance, even in complex and cluttered environments where survival-critical information may appear at varying sizes. While this feature enhances the interpretation of visual scenes, it comes with a trade-off: the necessity to position objects of interest centrally on the retina.

Finally, based on these observations and the fact that translations in Cartesian space induce a significant, nonlinear transformation in retinotopic space (see Figure 1-A&B), we investigate the effect of translations in Cartesian space on retinotopic networks. We thus investigate the effect of a rigid full-field translation by applying a roll function to the input image and place the fixation point at different positions in the image. Specifically, the fixation points are linearly distributed on an 11×11 grid. We then plot the mean accuracy when systematically selecting the worst position (based on the minimum loss as in the scenarios for a rotation Figure 2-A or zoom-out attack Figure 3-A) for the target label (see Figure 4-A), or the mean accuracy of the networks as a function of the position repositioned in the center (see Figure 4-B).

During this attack (see Figure 4-A), the accuracy of the retinotopic networks is degraded to 10.1%, i.e. a drop of 67.3% (8% in "regular", i.e. a drop of 66.3%), while the Cartesian "raw" and "focus" accuracies were degraded to 55.9% and 57.5%, i.e. a drop of 24% and 25.5% (52% in "regular", i.e. a drop of 26.9%). As expected, in contrast to previous attacks, the resilience of retinotopic networks is below that of Cartesian networks. Looking at the average accuracy maps (see Figure 4-B), we can see that the fixation points around the centre of the map have higher accuracy values than the fixation points around the periphery. This could be an artefact due to the photographer's bias explained earlier. However, this effect is not observed on the Cartesian maps, which shows similar accuracy for all positions examined. Indeed, the robustness of our retinotopic networks to zoom-out and rotation comes at the cost of a high sensitivity to image translations. This increased sensitivity, although detrimental for classification tasks, is associated, as we will see, with an important capability to localize objects of interest in visual space, providing a basis for spatial processing in the
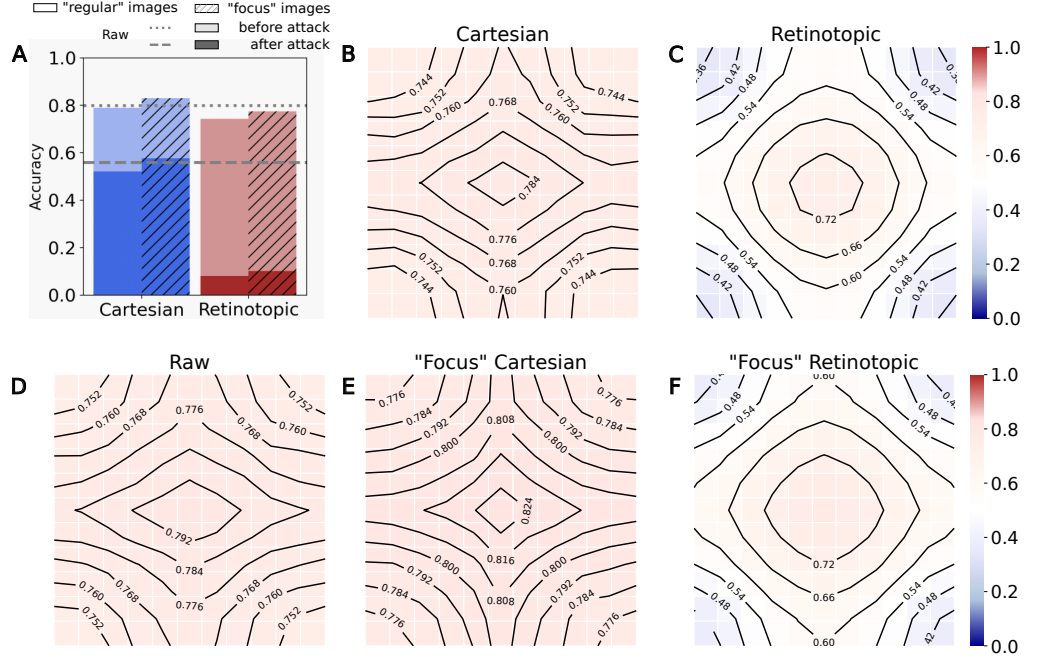
Fig 4. (A) For all networks, we plot the accuracy averaged over the validation dataset (in light color), or for each image translated (rolled) to the position with the worst loss (translation attack, in full color). The gray dashed lines represent the accuracy of the "raw" network (without re-training). **(B-F)** Average accuracy over a sample of $50,000$ images from the IMAGENET validation dataset, shown for both re-trained and pre-trained networks with different input image translation roll. The translation is applied from the central fixation point and defined a linear grid of 11×11 point of fixations

brain.

Visual object localization : Likelihood maps

Protocol

To quantify the contribution of this sensitivity to translation, we consider here a new task, i.e. the *visual search* task, in which a visual object (of which the label is known in advance) needs to be localized over the entire image. Is is for instance known that such task allows for the quick retrieval of an image label [46]. We design a protocol for each network to allow us to compare different visual shifts, each one corresponding to a potential fixation point, and to generate a map of the expected (or actual) accuracy as a function of the fixation point (see Methods "Visual object localization : Protocol"). In this protocol, a set of 11×11 fixation points is defined (with the coordinate $(5, 5)$ being the center of the image), and at each coordinate of the grid, a *likelihood* value is computed for the label of interest (see figure 13). In practice, a likelihood is given at each location from a softmax calculation over the different labels, as in the classical RESNET classifier, providing a value between 0 and 1 for the label of interest. This projects the network output onto a 11×11 Bernoulli probability space corresponding to the likelihood of detecting the given label at each position, finally providing a "heat map" on our 11×11 grid in a way that is compatible with other localization protocols. We tested our likelihood protocol on the "regular" validation dataset (see Fig 5 for some examples).

Mean likelihood maps

In Figure 6, we calculated the likelihood maps for all the images of the validation set, and re-centered them to place the viewpoint with the highest likelihood at the center of the grid. During the re-centering process, the spots outside the grid were assigned a `Nan` value to facilitate boundary management, and the average "center maps" were generated using the `nanmean()` function, providing the likelihood profile of the label of interest as a function of the distance from the most salient position. Our recentered likelihood maps are shown on Figure 6. All three maps show a similar 2D bell-shaped activation, indicating a clear object positioning capability for all networks, retinotopic or not, when applying our "visual search" protocol (i.e. without rolling the image borders). The activation level is higher at the center ("near" the object of interest), and lower at the periphery. The likelihood values are different though, being higher on average for the Cartesian network, lower on average for the "raw" network, and more contrasted for the retinotopic network. To further quantify this contrast difference, we calculate the map difference on the second row of Figure 6, along with the log-odd ratio in the third row of Figure 6, considering one-to-one comparisons, i.e. Cartesian vs. "raw", retinotopic vs. "raw", and Cartesian vs. retinotopic.

Let's analyze Figure 6. The second row shows the difference maps. The Cartesian minus "raw" map shows only positive values, while the retinotopic minus "raw" network has both positive and negative values. The retinotopic minus Cartesian map has only negative values, indicating on average a higher likelihood level in the Cartesian case. The area around the center remains close to zero though, reflecting a sharper slope towards the peripheral region in the retinotopic case. The log-odd ratio maps (third row) look quite similar at first sight to the difference maps, except in the center where the differences are more manifest. Of particular interest is the comparison of the first and second column, i.e. Cartesian vs. "raw" (first column), and Retinotopic vs. "raw" (second column). The radius of the central spot appears clearly different in the two cases, with a large central spot in the Cartesian vs. raw case, reflecting a loosy spatial
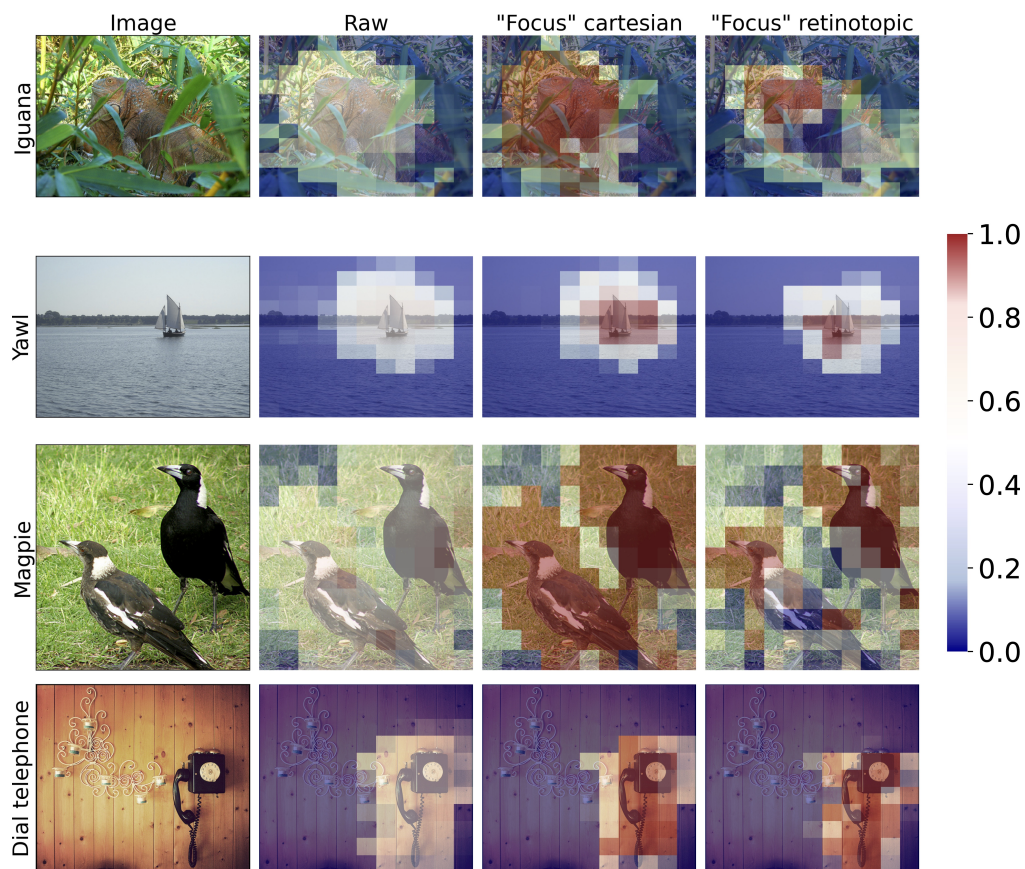
Fig 5. Likelihood maps computed on some prototypical images using 11×11 fixation points with the "raw" network, that is the original RESNET classifier with no re-training (second column), the network re-trained on images with circular mask on the "focus" dataset (third column), and the one re-trained on log-polar tranformed inputs (last column). The map displays the likelihood for the label of interest in the image.
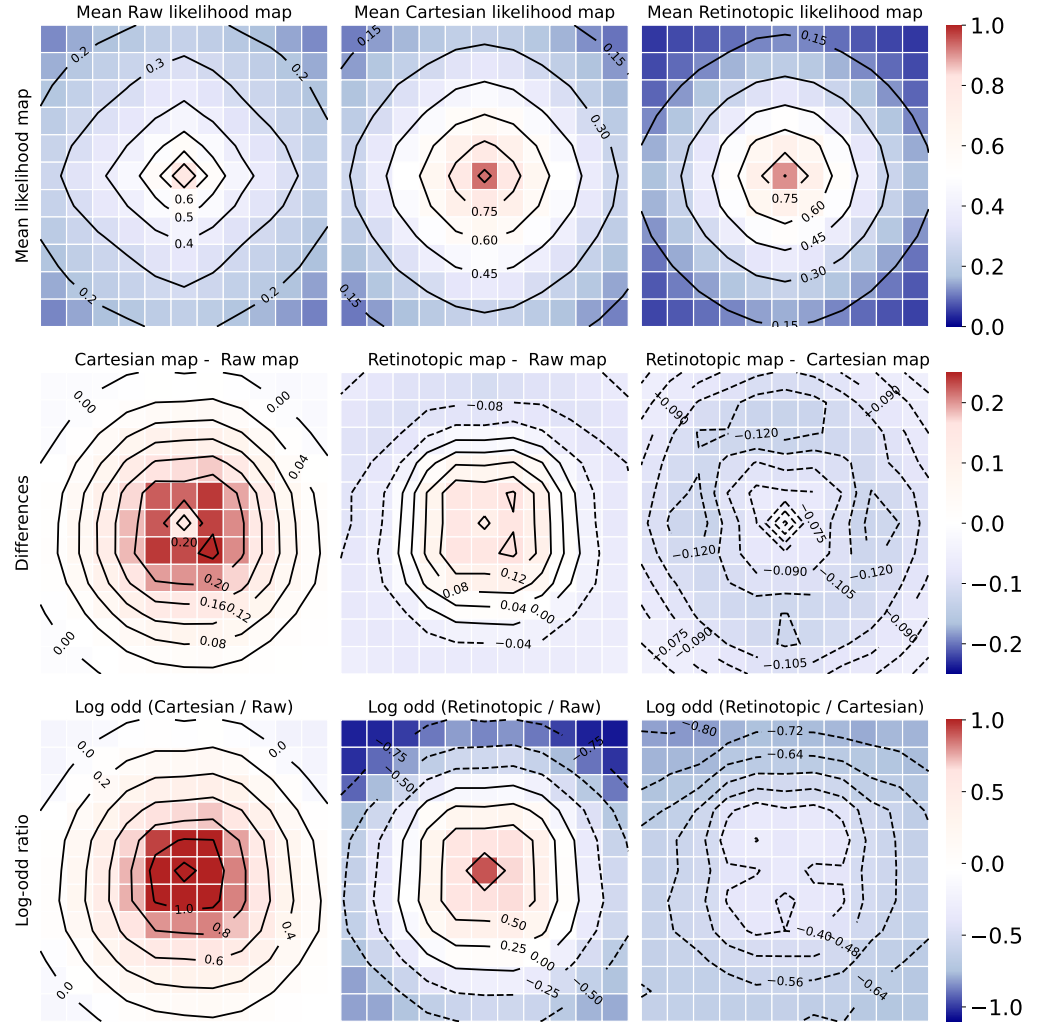
Fig 6. Top row: Mean likelihood on IMAGENET's validation dataset (50000 images). From left to right: "raw" network (no re-training), Cartesian network retrained on the "focus" dataset, and Retinotopic network retrained on the "focus" dataset. **First row:** recentered maps, averaged over the validation dataset. **Middle row:** Difference maps. **Bottom row:** Log-odd ratio maps.
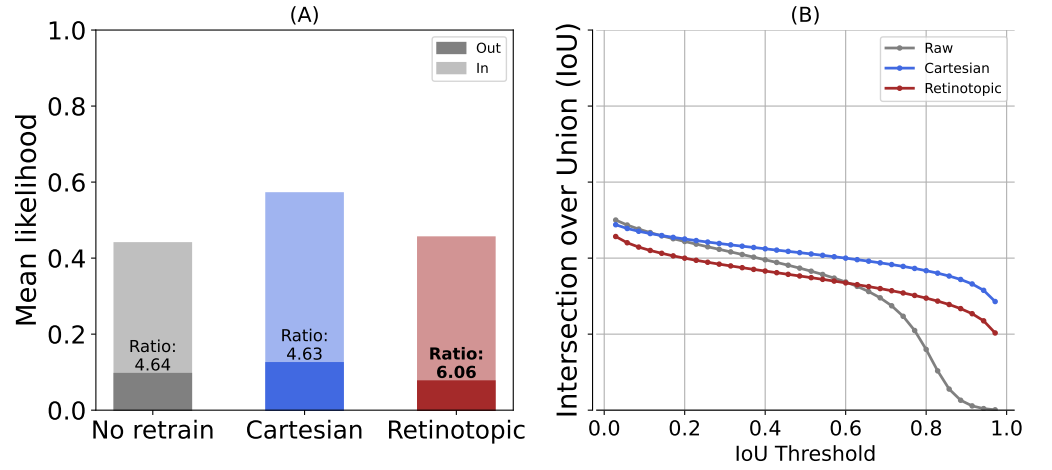
Fig 7. We tested the visual search protocol on the IMAGENET validation dataset (50, 000 images). **(A)** The mean likelihood across the point of fixation inside the bounding box ("In") or the point of fixation outside the bounding box ("Out") and the corresponding ratio of activation. **(B)** The intersection over union as a function of a threshold applied on the likelihood map.

discrimination improvement, and a more reduced one in the retinotopic vs. raw case, reflecting a more local and object-centric spatial discrimination improvement. Moreover, only the retinotopic network (second column) has significant peripheral depletion in comparison with the "raw" network. These observations collectively support the idea that the retinotopic network generates a heat map with enhanced contrast around the area of interest. In addition though, a focal Cartesian network also manifests good object localization capabilities, with high likelihood values around the object of interest, though being less spatially specific.

Comparing likelihood maps and Intersection over Union

To further quantify the difference between the Cartesian and retinotopic cases, we considered the localization information provided by the bounding boxes. A bounding box is defined as the rectangle of minimal surface containing one whole visual object. As such, each bounding box partitions the image into two regions, a region where the object of interest is present and a region where it is absent. Thus, given each likelihood map in the validation set, we compute the mean likelihood for the label under study *within* the bounding boxes (ground truth from the dataset) and the mean likelihood *outside* these boxes. Given the higher classification accuracies of the networks when fine-tuned on bounding boxes only the networks retrained on the "focus" dataset are used for the remainder of the study. The results are shown in Figure 7-A. The figure shows a mean likelihood value obtained inside and outside the bounding boxes for the original RESNET network, our re-trained Cartesian network, and the re-trained retinotopic network. Both networks show significantly higher likelihoods when the fixation point is inside the bounding box than when it is outside the bounding box. This reflects a higher confidence in the label response. At first glance, the average likelihoods in the three conditions seem quite comparable, although the likelihood values appear slightly higher in the re-trained Cartesian case.

For a quantitative comparison, we then calculate the likelihood ratio between the area inside and outside the bounding box (see Figure 7), shown at the top of the bars: the retinotopic network has a higher likelihood ratio than the two Cartesian ones, i.e. 6.1 versus 4.6, providing quantitative evidence for a higher contrast of localization in

favor of the retinotopic networks. This higher contrast is instrumental in localization tasks, as it allows for better identification of the region of interest. This effect can also be deduced from the examples shown in Figure 5: in the retinotopic case, regions of high likelihood are more sparse but still highly contrasted.

In addition, we consider in Figure 7-B the Intersection over Union (IoU) metric to evaluate the agreement between the bounding boxes and the activation maps (see Methods "Localization tools and evaluation"). In contrast with the "raw" network, our two networks fine-tuned on the "focus" datasets show a slower decay rate, reflecting a better fit with the ground truth. As expected from previous remarks, the IoU stays consistently higher for the cartesian network than for the retinotopic one, illustrating a tendancy for the retinotopic networks to "concentrate" likelihoods on smaller portions of the bounding box. The peak IoU is obtained for likelihood thresholds close to 0 for all networks, this could be an artifact of the measurement, since we are using the bounding box as a comparison, which encourages a larger number of positions with increased likelihood values on the map, rather than a sharp contour around the object of interest.

Pointing game

We have shown that the likelihood maps provide us with an indication, for each image, of the best fixation point to identify the object we are looking for. However, this spatial indication (where to place the eye) does not tell us anything about the visual content of this fixation point. To estimate this visual content more precisely, we now consider the "pointing game" metric, that is the rate of successful pointing (i.e. landing *inside* the bounding box) when choosing the highest likelihood position on the likelihood map. The bounding boxes provide us with rather coarse information about the visual content of the image, dividing it into two zones: a box within which the object is present, and a box outline where it is assumed absent. This binary information ('in' or 'out') is, however, sufficient to indicate whether the point of highest likelihood is correctly located on or in the immediate vicinity of the object of interest.

We compute for each network on the validation set the percentage of successful pointings. The results are shown on Figure 8 (green bars, left panel), on the raw, Cartesian and Retinotopic networks. The "raw" network has 71.2% success, the Cartesian has 78.74% success and the Retinotopic 85.14% success. Those rather high pointing game values reflect the fact that, on average, the bounding boxes cover a large portion of the image in the IMAGENET-1K dataset. Nevertheless, the retinotopic network shows pointing scores significantly higher than those of the raw network and the Cartesian network. The position of maximal likelihood appears to be strongly correlated with the presence of the object, and this effect is notably sharper and more
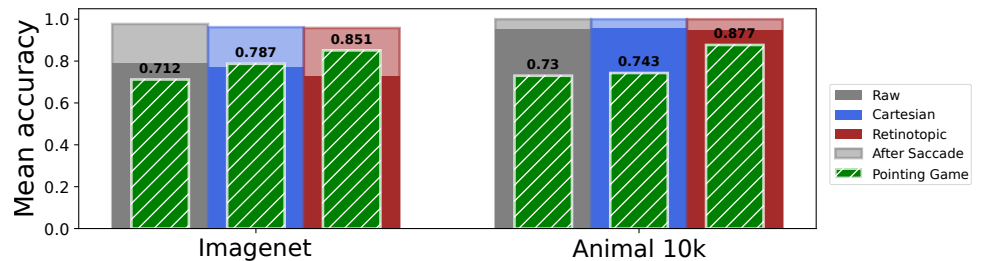

Fig 8. The visual search protocol is tested on the IMAGENET-1K validation dataset (comprising $50,000$ images) (left) and on the ANIMAL 10K dataset (comprising $10,015$ images) (right). The mean accuracy of the networks is displayed, along with the accuracy in the pointing game.

Table 1. Analyses of the key metrics for localisation (in/out likelihood ratio and pointing game) and classification (central fixation and max likelihood ("saccade") fixation) for different network depths.

	RESNET 18		RESNET 50		RESNET 101	
	Cartesian	Retinotopic	Cartesian	Retinotopic	Cartesian	Retinotopic
In/Out Likelihood Ratio	6.50	**9.02**	5.02	**6.40**	4.63	**6.06**
Pointing Game	83.07%	**87.40%**	80.81%	**85.13%**	78.74%	**85.14%**
Central fixation	**0.62**	0.52	**0.76**	0.71	**0.77**	0.73
With saccade	**0.92**	0.86	**0.95**	0.94	**0.96**	0.95
Accuracy increase	+30%	**+34%**	+21%	**+23%**	+19%	**+22%**

precise in the case of the retinotopic network.

We extend this visual search strategy by considering the classification accuracy obtained when placing the eye at the position of highest likelihood on the likelihood map, interpreted as a "saccade" toward the searched object. By doing so, the accuracy of all networks increases to about 96% ("raw" 97.7%, Cartesian 96.2%, retinotopic 95.8%, see Figure 8), providing an estimation of the *improvement* that can be expected, with respect to the baseline, when placing the eye (or camera) at its highest likelihood position (interpreted as the highest *likely* position). Consistently with our previous results, the accuracy improvement is higher in the retinotopic case, with 22.9% improvement, to be compared with 18.7% and 18.9% improvement in the "raw" and Cartesian cases respectively (see Figure 8). This strong response improvement reflects the critical role of the "direction of sight" in image classification, even in the Cartesian case, an issue little evoked in mainstream computer vision. It also suggest that the retinotopic network could be further improved by introducing such saccades during the retraining phase.

Impact of network depth on localization and classification

We extend the analysis by considering the networks depths, ranging from RESNET 18 to RESNET 101. The results are shown on Table 1. We observe the same general trend across different network sizes. Thus, the classification rate improves when the fixation point is placed at the position of highest likelihood, the likelihood values are higher inside the bounding box, and the fixation point is predominantly located within the bounding box. These trends are present in both types of networks (Cartesian or Retinotopic), but they are stronger in the case of the retinotopic one. Network size primarily affects the classification rate, with lower rates observed for smaller networks. Classification rates remain slightly higher for the Cartesian network in all cases, but the improvement in classification rate based on fixation position is consistently stronger for the retinotopic network. Regarding localization (likelihood ratio and pointing game), network size does not appear to be a determining factor, with higher pointing scores and likelihood ratios observed for smaller networks, and once again, a systematic advantage for retinotopic networks. In short, the Cartesian networks keep a slight advantage in categorization, while the retinotopic networks show improved performance in localization. It is noteworthy that smaller networks show improved localization capabilities, challenging the conventional wisdom that accuracy increases with the number of layers.

Beyond IMAGENET, the ANIMAL-10K dataset

IMAGENET-1K provides rich semantic links that allow the construction of task-specific datasets. It has previously been demonstrated that the use of fine-tuning to re-train
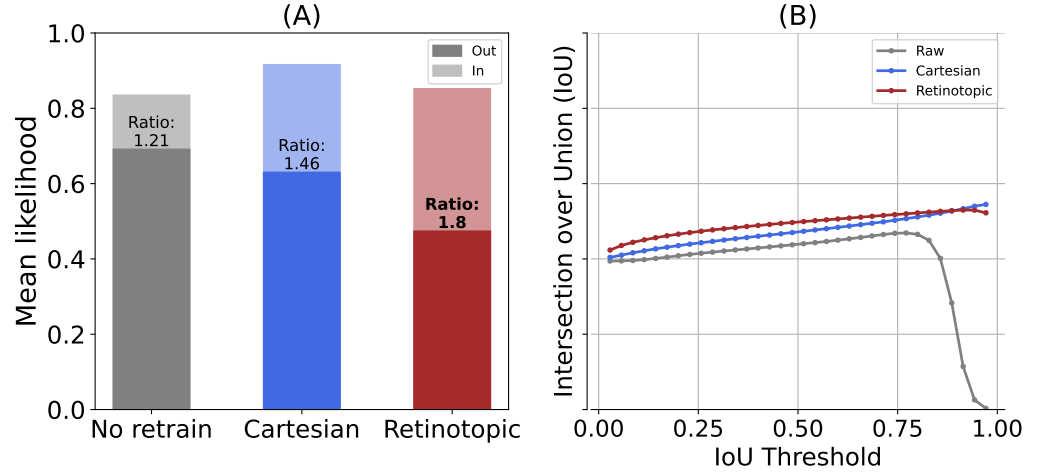
Fig 9. We tested the visual search protocol on the Animal 10k (comprising $10,015$ images). **(A)** The mean likelihood across the point of fixation inside the bounding box ("In") or the point of fixation outside the bounding box ("Out") and the corresponding likelihood ratio. **(B)** The intersection over union as a function of a threshold applied on the likelihood map.

networks such as Vgg16 [47] allows them to be applied to different tasks using the semantic network underlying ImageNet's labels. Furthermore, it has been shown that the probability of a trained network performing a novel task (such as categorizing an animal) can be predicted using this semantic network, which links the outputs of a pre-trained network to a library of labels. This approach has been shown to be an effective method for learning to predict the presence of animals in images [39]. We can then exploit the semantic connection underlying the ImageNet labels by using these networks to perform a categorization task on the second dataset, the Animal 10k dataset, as referenced in Yu et al.'s work [48], with the goal of identifying the animals depicted in the images. This dataset contains $10,015$ images, each depicting an animal. All images was integrated in a validation dataset to test the networks. The output of each networks is interpreted as "animal" if the label represented by the highest likelihood after softmax is indeed an animal. We use this prediction to calculate the accuracy of predicting an animal for all networks. Note that the accuracy is slightly biased due to the absence of distractors, it only represents the probability of finding an animal when there is actually an animal in the input. Considering first the accuracy improvement, both networks are compared on Figure 8 (right panel). The general accuracy is high, around 95% in all cases, because of our simpler binary (animal/non-animal) classification task. Sacade improvement is sensible still, with all networks reaching a maximal accuracy of 100% after a saccade to the maximum likelihood position for the label "animal". The highest improvement is once again found in the retinotopic case ("raw" 4.5%, Cartesian 4.1%, retinotopic 6.3%).

Regarding the localization capability, the In/Out likelihood ratio (Figure 9A) shows clear contrast enhancement, from 1.2 (raw network) toward 1.5 (Cartesian "focus" network) and 1.8 (Retinotopic "focus" network). The general lower contrast, when compared to ImageNet, comes again from the simpler binary (animal/non-animal) classification task, making it possible to guess the label "Animal" outside the bounding box from the background information. Looking now at the IoU metric (Figure 9B)), a different trend is observed, with monotonically increasing IoU with the likelihood map threshold, reflecting a convergent matching of the likelihood map with the animal

Fig 10. The likelihood maps presented here were generated by RESNET101 networks that were trained and tested on retinotopic space using the log-polar grid. **(A)** The original image, it is processed on a single iteration, with the likelihood maps displayed in accordance with the specific combinations of labels of interest. **(B)** Likelihood maps indicating the probability of the presence of the labels "dog", "cat" and "bird" in the image. **(C)** The sample derived from all fixation points that yield the highest likelihood value for a label of interest.

contour. The optimal threshold, close to 1, reflects a clear and sharp transitions in both cases. The retinotopic network provides here slightly higher IoU values, except at the highest threshold case where the Cartesian network prevails. The difference between this and IMAGENET lies in the fact that masks are used in the calculation instead of boxes (see Methods).

Last, the pointing game results (see Figure 8B, green bars) provide a clear assesssment of the cartesian/retinotopic contrast. While both the raw and cartesian network remain quite low (73-74% success), the retinotopic framework shows much higher pointing success (87.7%) despite refined animal contour, once again suggesting a more more reliable animal localization. Despite the increased diversity of poses, environments, and intra-class variability in ANIMAL 10K compared to IMAGENET-1K, our results indicate that the regions containing the target animals are effectively identified.

Multi-label and multi-task extension

The likelihood map can be extended along the dimension of labels. In fact, the networks perform a discriminative categorization among 1000 IMAGENET labels and the likelihood maps generated by the model are expected to vary depending on the label or

class 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 insight 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 "dog", "cat", and "bird". As expected, the different regions are separated accordingly (see Figure 10).

Discussion

Retina-inspired mapping enhances the robustness of CNNs

The first and main result of this study is to demonstrate the excellent ability of standard deep CNNs to deal with foveated retinotopic inputs, even though this mapping enforces a radical transformation of the visual inputs. RESNET networks easily adapt to these inputs and the accuracy rates achieved with retinotopic inputs are equivalent to those of the original models. This is surprising given that the networks used in this re-training process were previously trained on Cartesian images and that images with a log-polar transformation show this high degree of distortion, in particular a high compression of visual information around the fixation point and a degradation of textures in the periphery, see Figure 1. One possible hypothesis is that the degradation of texture during the frame of reference change may cause the network to rely on shape rather than texture [33]. A further study could entail a comparison and replacement of the aforementioned parameters with those of humans, thereby providing insight into the evolutionary trajectory of these parameters.

In addition, the log-polar transformation has the advantage of better invariance to zooms and rotations. This study shows first that for the original RESNET network on "regular" images, the average accuracy with a zoom or a rotation dropped sharply compared to baseline accuracy, confirming that this simple geometrical transformation mislead the networks. For rotations, the decline was steepest around $160°$, and demonstrate limited rotational invariance compared to humans [49, 50]. The robustness differs slightly when studying the impact of zooms, and this may be attributed to the transformation under examination that already provides a zooming-in effect of the fovea which accentuates the information in the area surrounding the point of fixation. To test this hypothesis, we may train a retinotopic version without applying the logarithmic function to the eccentricity axis. Finally, 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 initial consequence of this transformation can be observed in a comparison between the "regular" version of the networks and the "focus" version.

The integration of a retinotopic mapping approach holds great promise for improving the efficiency and accuracy of existing image processing tasks. First, it involves only a simple preprocessing of the input images, transforming images into novel matrices, and is therefore potentially applicable to existing image processing algorithms. The log-polar compression used in our approach allows a seamless extension to larger images without a significant increase in computational cost. On the other hand, this biological inspiration results in a more frugal energy consumption. This is an essential aspect when porting such algorithms to embedded systems under high power constraints. Overall, our results are consistent with features of biological vision. In particular, it is consistent with the robustness observed in physiological data on ultrafast image categorization, which show that human accuracy in recognizing briefly

flashed images of animals is robust to rotation [50, 51].

From Foveation to Pre-attentive Mechanisms

The second result of this study is the emergence of localization properties in networks re-trained with foveated inputs. Thanks to the definition of likelihood maps, that collates the output of the model while scanning the visual scene at a limited number of fixation points, this 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. This is reminiscent of pre-attentive mechanisms which allows biological vision to only process the important zones of the visual space. As a result, it provides a proxy for the measurement of saliency, in particular with respect to a set of cued labels.

One hypothesis is that the foveated retinotopic mapping implemented in the log-polar transform implements an efficient prior for visual object geometry. Indeed, a main source of variability in the view of an object comes from displacements of the observer relative to it, for example as the objects scales in the visual field as one approaches it, or as the object rotates with a rotation of the head. The log-polar mapping allows for a more invariant representation which explictly implements this prior for these displacements. Additionally, imposing translation invariance of the representation implements a prior on the possible representations allowed by the network [44]. However, it was observed that some classes have different statistics, and the relative size of buses are on average bigger than that of cats. Globally, we expect that different foveated retinotopies may emerge in different ecological niches.

In our case, it seems that the foveated retinotopic mapping allows for a more precise localization of the category of interest 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 [52], and be 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 instead of the center of the bounding box defined in the "focus" dataset, we could consider refining the training of the network to our retinotopic mapping in a semi-supervised fashion.

Perspectives and future work

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 saccadic planning 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 strategies for producing saccades, allowing a direct comparison with humans' 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. In particular, this line of research should be particularly well fitted for a dual-pathway model which is well adapted to infer ego-motions [53, 54]. 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.

Materials and Methods

The log-polar transform

In most mammals and amphibians, the arrangement of the (external) visual field is preserved in the early visual pathway, a feature called retinotopy. Retinotopic mapping 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. These mappings differ from species to species, and our study concentrates on foveated vision (as in humans) which gives more resolution to the central field of view. In particular, we implement it by transforming the Cartesian pixel coordinates into log-polar coordinates [27].

This simple parameterized transformation models accurately this biologically inspired retinotopic mapping. Considering arbitrary images (potentially with multiple channels such as RGB), each pixel's position is defined by two real coordinates (x, y) on a Cartesian reference frame. By convention, x and y are here considered belonging to the interval $[-1, 1]$, with $(0, 0)$ being the center of the image. To implement the concentration of pixels near the center of the retina, we need to consider an *irregular* grid in the Cartesian referential that maps to a *regular* grid in the log-polar referential. In the log-polar referential, the location of each pixel has corresponding coordinates (ρ, θ) as defined by previous studies [45] by:

$$\rho = \log_2 \sqrt{(x - x_0)^2 + (y - y_0)^2} \tag{1}$$

$$\theta = \arctan(\frac{y - y_0}{x - x_0}) \tag{2}$$

with (x_0, y_0) defining the "center of fixation" (see Figure 1) [55]. Importantly, ρ and θ are only defined for $(x, y) \neq (x_0, y_0)$. In most of our experiments, we consider $x_0 = (0, 0)$, allowing us to focus on the central part of the image. Each dot (ρ, θ) in the log-polar coordinate system has thus a unique correspondence in the Cartesian coordinate system (and vice versa): For each $(\log \rho, \theta)$ belonging to the grid, the corresponding pixel coordinate is $(x_0 + \rho \cos \theta, y_0 + \rho \sin \theta)$.

In practice, images have a finite resolution and to avoid biases in the evaluation between networks, the number of angles sampled (N_θ) and the number of eccentricities sampled (N_ρ) are set to 224, so that the size of the transformed image match the resolution 224×224 of the input images. Note that native image resolution is generally higher than that used during network processing, with the average resolution in the Imagenet dataset being around 500×500 pixels. This transformation is performed with the PyTorch library [56] through the use of the grid_sample() function, which maps the pixels of an input image to the coordinates of any arbitrary grid, using a linear interpolation to estimate the value of the pixels. This function is used, for instance, in spatial transformer networks [57].

Let's now define each coordinates. All θ values are within a linear distribution in $[0; 2\pi)$, while ρ values are within a logarithmic interval with $r_{\min}$ the minimal radius and $r_{\max}$ be the maximal radius (with N_ρ the radial resolution). In practice, we use a log-polar grid with an outer log-radius of $\log_2 r_{\max} = 0$ ($r_{\max} = 1$, defining a circle

tangent to the image box) and an inner log-radius of $\log_2 r_{\min} = -5$ ($r_{\min} = 2^{-5}$). In summary, the regular grid is the interval
$[\log_2 r_{\min}, ..., \log_2(r_{\min} + i \times \frac{r_{\max} - r_{\min}}{N_\rho - 1}), ..., \log_2 r_{\max}]$, for i in $[0, ..., N_\rho - 1]$, in the $\log_2 \rho$ dimension and $[0, ..., j \times \frac{2\pi}{N_\theta}, ..., 2\pi(1 - \frac{1}{N_\theta})]$, for j in $[0, ..., N_\theta - 1]$ in the θ dimension. Note that some pixels in the log-polar grid may be smaller than the pixels from the Cartesian grid.

Convolutional Neuronal Networks (CNNs)

Convolutional neural networks (CNNs) have become essential tools in image classification, with several pre-trained models available for download. For example, the VGG family, including VGG16 and VGG19, introduced by Simonyan [47], uses deep architectures with 16 or 19 layers, consisting of stacked convolutional layers followed by fully connected layers. RESNET (Residual Networks), introduced by He et al. [36], addresses the vanishing gradient problem in deep networks by incorporating skip connections, allowing for the training of extremely deep networks (e.g. RESNET 50, RESNET 101 with respectively 50 and 101 layers). These models are widely used due to their robustness and scalability. The implementation of these deeper networks have demonstrated that deeper networks display enhanced resilience; however, this improvement is coupled with an overall increase in computational complexity [36]. Therefore, based on these findings, we focus on the deep CNN RESNET (with 18 to 101 layers from the PYTORCH library [1]) on the IMAGENET-1K [35] categorization challenge which consists in classifying natural images into 1000 labels.

Datasets

Typical image classification datasets used in machine learning consist of sets of RGB images of different resolutions, each image being associated with a single label. The classification task involves learning a parametric function that learns to associate a unique label with a set of (high-dimensional) pixels. Example-based learning therefore assumes that there are regions within the representation space (or feature space) that can be learned in order to separate objects into these different classes, regardless of their position in the image, size, orientation, lighting, contrast etc...

Two datasets were used for our study: the first dataset is the one from the IMAGENET [35] challenge, which is the most widely benchmarked due to its huge collection of images and associated labels (the subset of IMAGENET used in this study, i.e. IMAGENET-1K with 1000 labels and about 1000 examples per label). In addition to the classification task, we consider here the localization task, which is prominent in computer vision. It consists in identifying both the label and the position of one or more objects of interest in an image. A distinction can be made between (i) the *visual search* task, where the label is given in advance and the task is simply to find the position of the object in the image, and (ii) the *image labeling* task, which consists in identifying both the objects and their position in the image, in the form of a labeled bounding box (see Figure 11). Despite its advantages, IMAGENET has limitations for localization tasks. For example, the dataset lacks multi-labeling, with only one label per image. It is worth mentioning that IMAGENET has some biases, the objects of interest are generally centered in the images, and the proportion of bounding boxes relative to the image size is relatively large, which may limit the impact of certain analyses.

We consider two dataset configurations for the IMAGENET dataset: In our first configuration, the center of gaze is set to the center of the original image, taking

---

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

advantage of the fact that most IMAGENET images are human-made and that photographers have a bias toward centering the object of interest. This defines our "regular" image dataset. Notwithstanding the above, this a priori assumption of centered position is not sufficient to generate a dataset perfectly suited for retinotopic transformation. In a second setup (called the "focus" dataset), we use the bounding box information provided by the IMAGENET dataset. Consequently, a sample is selected, defined as the smallest square containing a bounding box, assuming a center of gaze now at the center of the bounding box for the label of interest. This novel dataset is used to train a second generation of networks. Again, we use a circular mask for the Cartesian frame. This approach is more robust to the position of the visual object, but requires reliable bounding boxes.

The dataset provides a set of key points for each animal present in an image. 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 decreasing with a standard deviation proportional to object size, thus capturing the true spatial extent and location of the target animal within each image (see Figure 12). This approach allows for more effective localization and analysis of the visual distribution of animals in images.

Datasets transformations and Transfer learning

In this study, three series of transformations were used, depending on whether the network used the Cartesian or the retinotopic reference frame. In the case of the latter, a log-polar grid is used to facilitate the transformation of the image into the retinotopic frame. Due to the intrinsic nature of the transformation, which results in the cropping of a circular sample within the original image, a circular crop is implemented for the Cartesian frame to ensure comparability. To allow for a more straightforward comparison, the "raw" datasets were processed without a circular mask or polar logarithmic transformation. Each set underwent a uniform transformation, including normalization to tensors and resizing to a resolution of 244×244 to match the pre-trained parameters of the model.

To assess the efficacy of our retinotopic mapping, we examine popular off-the-shelf CNNs pre-trained on standard, large image datasets. These networks are re-trained on our datasets, either with (or without) a log-polar retinotopic transformation, using the cross-entropy loss from the PYTORCH library. We use the stochastic gradient descent (SGD) optimizer from the PYTORCH library and validate parameters such as batch size, learning rate, and momentum by performing a sweep of these parameters for each network. During the sweep, we vary 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 choose the parameter values that give the best average accuracy on the validation set: batch size = 80, learning rate = 0.001, momentum = 0.9. We re-trained the networks during 2 epochs of the full training dataset, keeping all learning parameters identical.

Attacking classical CNNs with a geometrical rotation

A common approach to evaluating the robustness of deep learning models is to subject them to adversarial attacks. In this study, we investigate the robustness of the deep learning models to natural image transformations that are easily perceived by humans. In particular, we evaluated the performance of the networks on the IMAGENET dataset when the images were rotated by different angles and averaged the accuracy for each angle. To further evaluate the robustness to rotations, we also designed a "rotation-based attack" scenario. 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, which we denote as $\mathcal{L}(m(I), y)$. This is the loss minimized during gradient descent training. We then systematically rotated the images and tracked the change in model loss. 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 for each image in 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 is to first choose the rotation angle that maximizes the loss, and then infer the most likely 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. We use a similar strategy for other geometric transformations, such as zooms or translations.

Localization tools and evaluation

A widely accepted technique for evaluating the performance of Convolutional Neural Networks (CNNs) in localization tasks is the Class Activation Mapping (CAM) method. CAM works by analyzing the output of the CNN with respect to the target class, assigning weights to activations in each spatial feature map. This process generates a heat map that highlights significant areas of the image based on their contribution to the prediction. Building on the foundation of CAM, several derivative methods have emerged, including Grad-CAM [58], Score-CAM [59], and Opti-CAM [60].

In an effort to fairly quantify the respective contributions of each method, many quantification techniques have been developed. Here we select some of them to compare the models using the retinotopic or Cartesian reference. ENERGY-BASED POINTING GAME: Localization is successful if the peak activation of the heatmap of a given label is 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 activation inside and outside the box; the higher the value, the more efficient the heatmap is at indicating the position for a given label. INTERSECTION OVER UNION (IOU): Ratio of the area of overlap between the heatmap and the ground truth to the area of union between the heatmap and the ground truth. Peak-IoU and Peak-Threshold: For a modulation of a threshold on the heat map, the Peak-IoU is the maximum IoU value reached at the Peak-Threshold.

Visual object localization : Protocol

Deep convolutional networks such as RESNET output a vector of real numbers which predict the logit of the likelihood in label space, and this prediction is optimized 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, and when not specified this resolution is set to 11×11. 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 13). At each viewpoint, the largest possible sample is cropped. Thus a minimum sample with $1 : 10$ ratio at the border and the whole image at the center. From the Cartesian or retinotopic reference frame, this sample is then resized, if necessary, to a 224×224 resolution to match the input size for the CNN before processing or transformed by the retinotopic mapping (also at 224×224 resolution) before being used as input for the corresponding network, see Appendix Figure 13-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.

Acknowledgments

References

1. 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.

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

3. Polyak SL. The retina: the anatomy and the histology of the retina in man, ape, and monkey, including the consideration of visual functions, the history of physiological optics, and the histological laboratory technique. University of Chicago Press: Chicago; 1941.

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

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

6. 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. Journal of Vision;3(10):1. doi:10.1167/3.10.1.

7. Nauhaus I, Nielsen KJ, Callaway EM. Efficient Receptive Field Tiling in Primate V1. Neuron. 2016;91(4):893–904. doi:10.1016/j.neuron.2016.07.015.

8. Golomb JD, Chun MM, Mazer JA. The Native Coordinate System of Spatial Attention Is Retinotopic. The Journal of Neuroscience. 2008;28(42):10654–10662. doi:10.1523/JNEUROSCI.2525-08.2008.

9. Lu Z, Doerig A, Bosch V, Krahmer B, Kaiser D, Cichy RM, et al.. End-to-End Topographic Networks as Models of Cortical Map Formation and Human Visual Behaviour: Moving beyond Convolutions; 2023. Available from: http://arxiv.org/abs/2308.09431.

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

11. Martins DM, Manda JM, Goard MJ, Parker PRL. Building egocentric models of local space from retinal input. Current Biology. 2024;34(23):R1185–R1202. doi:10.1016/j.cub.2024.10.057.

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

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

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

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

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

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

18. Daucé E, Albiges P, Perrinet LU. A Dual Foveal-Peripheral Visual Processing Model Implements Efficient Saccade Selection. Journal of Vision. 2020;20(8):22–22. doi:10.1167/jov.20.8.22.

19. 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; 2020. p. 165–178.

20. 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.

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

22. Araujo H, Dias JM. An Introduction to the Log-Polar Mapping. Proceedings II Workshop on Cybernetic Vision. 1997;(1):139–144. doi:10.1109/CYBVIS.1996.629454.

23. Faramarzi F, Linares-Barranco B, Serrano-Gotarredona T. A 128×128 Electronically Multi-Foveated Dynamic Vision Sensor with Real-Time Resolution Reconfiguration. IEEE Access. 2024; p. 1–1. doi:10.1109/ACCESS.2024.3519035.

24. Sarvaiya JN, Patnaik S, Bombaywala S. Image Registration Using Log-Polar Transform and Phase Correlation. In: IEEE Region 10 Annual International Conference, Proceedings/TENCON; 2009. p. 1–5.

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

26. Palander K, Brandt SS. Epipolar Geometry and Log-Polar Transform in Wide Baseline Stereo Matching. In: 2008 19th International Conference on Pattern Recognition. IEEE;. p. 1–4. Available from: http://ieeexplore.ieee.org/document/4761515/.

27. Javier Traver V, Bernardino A. A Review of Log-Polar Imaging for Visual Perception in Robotics. Robotics and Autonomous Systems;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. Journal of Real-Time Image Processing;10(3):533–550. doi:10.1007/s11554-012-0281-6.

29. Wolfe JM. Guided Search 2.0 a Revised Model of Visual Search. Psychonomic bulletin & review;1:202–238. doi:10.3758/BF03200774.

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

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

32. Cao J, Bao C, Hao Q, Cheng Y, Chen C. LPNet: Retina Inspired Neural Network for Object Detection and Recognition. Electronics;10(22):2883. doi:10.3390/electronics10222883.

33. da Costa D, Kornemann L, Goebel R, Senden M. Convolutional neural networks develop major organizational principles of early visual cortex when enhanced with retinal sampling. Scientific Reports. 2024;14(1):8980. doi:10.1038/s41598-024-59376-x.

34. Lukanov H, König P, Pipa G. Biologically Inspired Deep Learning Model for Efficient Foveal-Peripheral Vision. Frontiers in Computational Neuroscience. 2021;15. doi:10.3389/fncom.2021.746204.

35. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV);115:211–252. doi:10.1007/s11263-015-0816-y.

36. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv:151203385 [csCV]. 2015;doi:10.1109/CVPR.2016.90.

37. Bengio Y, Lee DH, Bornschein J, Mesnard T, Lin Z. Towards Biologically Plausible Deep Learning; 2016.

38. Kubilius J, Schrimpf M, Nayebi A, Bear D, Yamins DLK, DiCarlo JJ. CORnet: Modeling the Neural Mechanisms of Core Object Recognition;. Available from: http://biorxiv.org/lookup/doi/10.1101/408385.

39. Jérémie JN, Perrinet LU. Ultrafast Image Categorization in Biology and Neural Models. Vision. 2023;7(2):29. doi:10.3390/vision7020029.

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

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

42. Bashivan P, Ibrahim A, Dehghani A, Ren Y. Learning Robust Kernel Ensembles with Kernel Average Pooling; 2023.

43. Mineault P, Zanichelli N, Peng JZ, Arkhipov A, Bingham E, Jara-Ettinger J, et al.. NeuroAI for AI Safety; 2024.

44. Lempitsky V, Vedaldi A, Ulyanov D. Deep Image Prior. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2018. p. 9446–9454. Available from: https://ieeexplore.ieee.org/document/8579082/.

45. Hao Q, Tao Y, Cao J, Tang M, Cheng Y, Zhou D, et al. Retina-like Imaging and Its Applications: A Brief Review. Applied Sciences;11(15):7058. doi:10.3390/app11157058.

46. Potter MC. Meaning in Visual Search. Science;187(4180):965–966. doi:10.1126/science.1145183.

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

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

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

50. Guyonneau R, Kirchner H, Thorpe SJ. Animals Roll around the Clock: The Rotation Invariance of Ultrarapid Visual Processing. Journal of Vision;6(10):1. doi:10.1167/6.10.1.

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

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

53. Bakhtiari S, Mineault P, Lillicrap T, Pack CC, Richards BA. The Functional Specialization of Visual Cortex Emerges from Training Parallel Pathways with Self-Supervised Predictive Learning;. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.06.18.448989.

54. Mineault PJ, Bakhtiari S, Richards BA, Pack CC. Your Head Is There to Move You around: Goal-driven Models of the Primate Dorsal Pathway;. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.07.09.451701.

55. Traver VJ, Pla F. Designing the Lattice for Log-Polar Images. Discrete Geometry for Computer Imagery; p. 164–173.

56. 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, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.;. p. 8024–8035.

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

58. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision. 2020;128(2):336–359. doi:10.1007/s11263-019-01228-7.

59. 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.

60. 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.
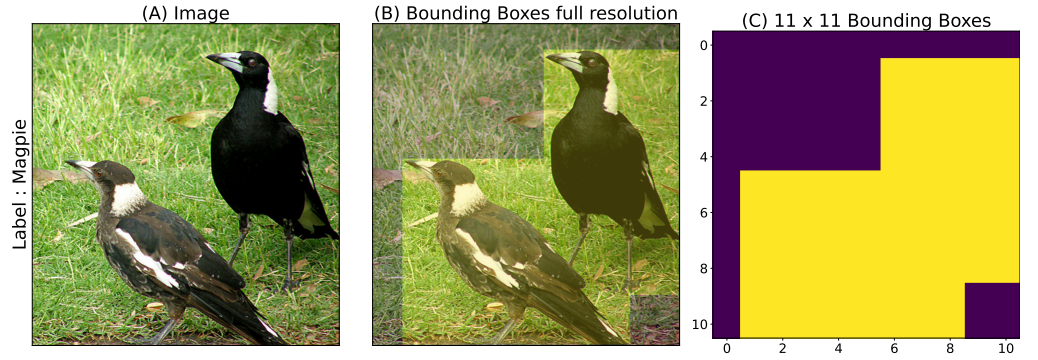
Appendix

Fig 11. (A) An original image from the IMAGENET dataset. **(B)** A heat map constructed by fitting bounding boxes to the image. **(C)** The heat map constructed in **(B)** and reduced to a resolution of 11×11 to be used as ground truth when evaluating the localisation performance of the networks.
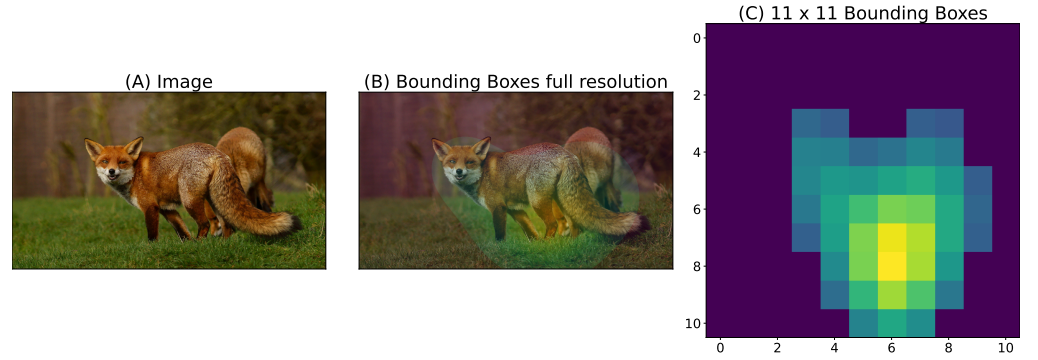

Fig 12. (A) An original image of the ANIMAL 10K dataset. **(B)** A heat map constructed by fitting Gaussians to the key points of the image. **(C)** The heat map constructed in **(B)** is normalized and reduced to a resolution of 11×11 to be used as ground truth when evaluating the heat map. A threshold (0.2) is applied to reduce the heat map field to the assumed contour of the animal.
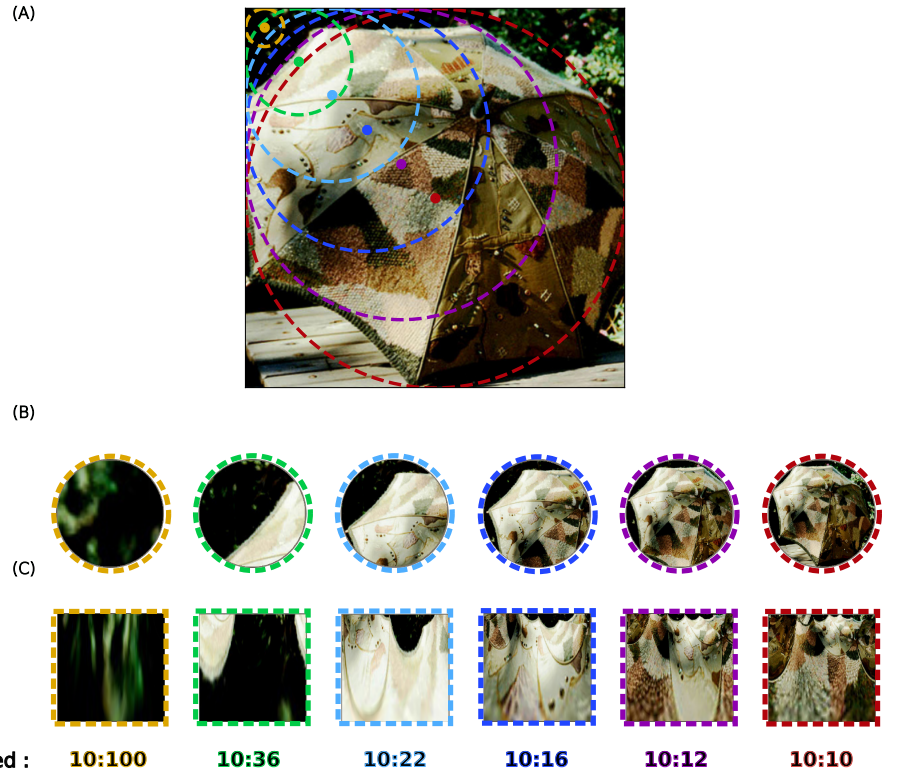
(A)

(B)

(C)

Ratio sampled : **10:100** **10:36** **10:22** **10:16** **10:12** **10:10**

Fig 13. We illustrate the protocol used to obtain the likelihood map with an example on a single image **(B)**. We show one sample for each sample ratio size ($1:10$ to $1:1$, from left to right). **(A)** The samples are cropped on the image in the Cartesian frame, we use circular cropping to match the area covered by the log-polar frame. **(C)** The corresponding samples cropped on the image in the retinotopic frame.