What if Eye...? Computationally Recreating Vision Evolution

Kushagra Tiwary*[1], Aaron Young*[1], Zaid Tasneem[2], Tzofi Klinghoffer[1, 6], Akshat Dave[1], Tomaso Poggio[3], Dan-Eric Nilsson[4], Brian Cheung**[3,5], Ramesh Raskar**[1]

[1]Camera Culture, MIT Media Lab, Cambridge, 02139, MA, USA.
[2]Computational Imaging Lab, Rice University, Houston, 77005, TX, USA.
[3]Center for Brains Minds and Machines, MIT, Cambridge, 02139, MA, USA.
[4]Lund Vision Group, Lund University, Lund, 22100, Sweden.
[5]InfoLab, MIT CSAIL, Cambridge, 02139, MA, USA.
[6]Charles Stark Draper Laboratory, Cambridge, 02139, MA, USA.

Abstract

Vision systems in nature show remarkable diversity, from simple light-sensitive patches to complex camera eyes with lenses [1, 2]. While natural selection has produced these eyes through countless mutations over millions of years, they represent just one set of realized evolutionary paths [3, 4]. Testing hypotheses about how environmental pressures shaped eye evolution remains challenging since we cannot experimentally isolate individual factors [5]. Computational evolution offers a way to systematically explore alternative trajectories [6–10]. Here we show how environmental demands drive three fundamental aspects of visual evolution through an artificial evolution framework that co-evolves both physical eye structure and neural processing in embodied agents. First, we demonstrate computational evidence that task specific selection drives bifurcation in eye evolution - orientation tasks like navigation in a maze leads to distributed compound-type eyes while an object discrimination task leads to the emergence of high-acuity camera-type eyes. Second, we reveal how optical innovations like lenses naturally emerge to resolve fundamental tradeoffs between light collection and spatial precision. Third, we uncover systematic scaling laws between visual acuity and neural processing, showing how task complexity drives coordinated evolution of sensory and computational capabilities. Our work introduces a novel paradigm that illuminates evolutionary principles shaping vision by creating targeted single-player games where embodied agents must simultaneously evolve visual systems and learn complex behaviors. Through our unified genetic encoding framework, these embodied agents serve as next-generation hypothesis testing machines while providing a foundation for designing manufacturable bio-inspired vision systems [11].

Keywords: Embodied Artificial Intelligence, Computer Vision, Evolutionary Biology, Computational Neuroscience

1 Introduction

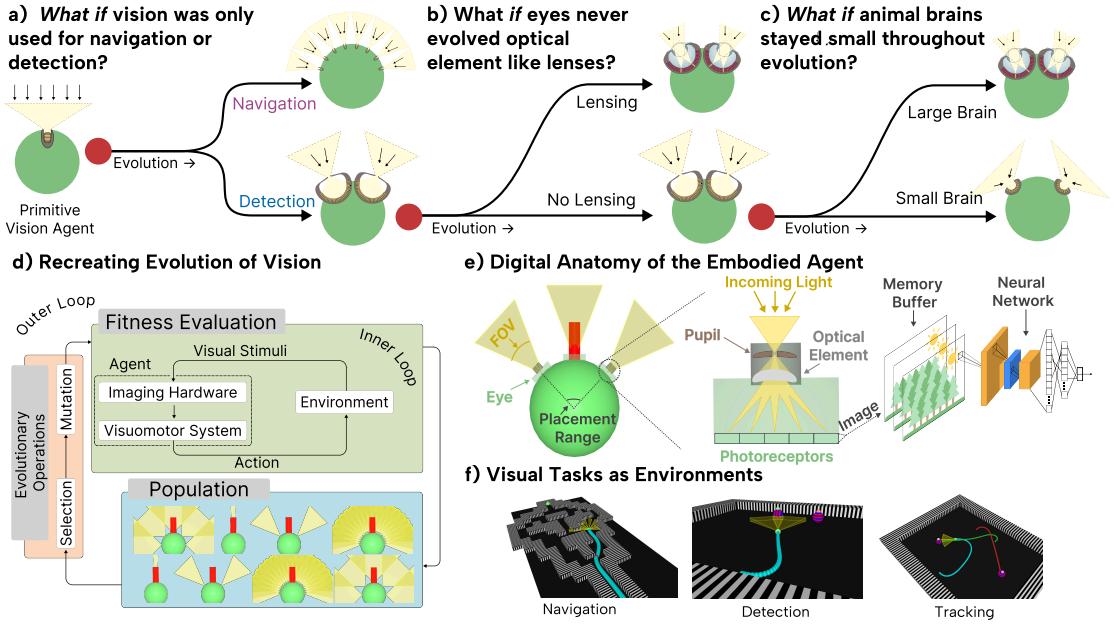

Fig. 1: **Computational evolution of embodied artificial intelligence (AI) agents reveals how environmental pressures shaped natural vision evolution.** We evolve artificial embodied agents to show how three evolutionary branch points shaped vision evolution. **(a)** We demonstrate how environmental specificity led to distinct eye morphologies, **(b)** a physical trade-off of light throughput vs. spatial precision lead to the emergence of optical elements, and **(c)** how visual acuity and neural capacity co-evolve to reveal hardware-software trade-offs and scaling laws. **(d)** Our framework mirrors natural selection: an outer loop governs genetic inheritance and selection over evolutionary timescales, while an inner loop enables agents to learn through sensory feedback (lifetime adaptation). This nested structure reflects the Baldwin effect [12], where lifetime learning can guide and accelerate evolutionary adaptation. **(e)** The agent's digital anatomy parallels biological visual systems: from eye morphology and placement, through optical elements and photoreceptors (mimicking retinal organization), to neural processing (analogous to visual cortex). **(f)** Agents are evolved to solve three distinct visual tasks to probe how environmental pressures shape vision: *(i)* NAVIGATION: orientation and obstacle avoidance through a maze-like environment; *(ii)* DETECTION: object discrimination between a 'good' object (food) and two 'bad' objects (poison); *(iii)* TRACKING: tracking of moving 'food' targets. Our results highlight how embodied agents can serve as scientific instruments to understand biological visual intelligence.

What if vision was only used for navigation or detection? What if eyes never evolved optical elements like lenses? What if animal brains stayed small throughout evolution? Operating over millions of years and culminating in millions of unique perception systems [1], natural evolution has followed specific evolutionary trajectories in its development of vision. What if there was a tool to instead simulate *alternative* paths that evolution didn't take? By computationally recreating the evolutionary dynamics (i.e., mutation, selection, adaptation) which gave rise to the remarkable diversity of eyes we see today, we can explore different evolutionary trajectories and systematically probe the principles that shape visual diversity. This framework would enable

us to test hypotheses about the relationships between eye morphology, neural processing, and environmental pressures, and guide the design of novel vision systems for artificial agents both in nature and engineering.

In this work, we introduce a framework to elucidate how environmental pressures shaped visual system evolution using embodied artificial intelligence (AI). Our approach evolves the eyes and neural circuitry of embodied agents inside physics-based simulation environments to understand what environmental factors drove vision evolution. While comparative biology has revealed remarkable insight into eye evolution [1, 3, 13], empirically testing causal hypotheses about environmental influences on eye evolution remains challenging. Our work builds on two foundational directions. First, pioneered by Grey Walter's machina speculatrix [6, 7], the use of evolutionary robotics to test scientific hypotheses about biological mechanisms and processes [8, 10, 14–16], and advanced through studies about predator-prey dynamics [17], brain-body co-evolution [11], and environmental adaptation [9, 18–21]; however, vision evolution has yet to be studied in this context. Second, the emergence of deep reinforcement learning (DRL) as a powerful tool for discovery in domains that can be formulated as reward-driven games [22–25]. Our work is the first that illuminates evolutionary principles shaping vision by creating single-player games with specific environmental conditions that embodied agents solve by evolving their vision and learning complex behavior simultaneously. We demonstrate that visually-capable embodied agents trained via DRL can serve as next-generation hypothesis testing machines.

We first implement a genetic encoding that unifies physical eye morphology, eye optics, and neural processing (Figure 2), and then computationally mimic the evolutionary process in embodied agents by evolving this genetic encoding to best complete a visual task (Figure 1.d). Our framework is the first to computationally recreate vision system evolution – where complex eyes and behaviors coevolve due to specific environmental pressures. We use this framework to study the emergence of visual capabilities documented across animal phylogeny [3, 4]. Our encoding integrates morphological, optical, and neural components into a unified genome capable of describing over 10^{20} unique configurations, and provides a continuous space for exploring evolutionary pathways (i.e., lens-less cup eyes, camera-type eyes, compound eyes). Subsequently, over generations, agent genes are selected and mutated, leading to the emergence of complex eyes and behaviors for specific visual tasks. This computational survival-of-the-fittest mimics the interplay of variation and selection that shaped biological vision.

Through targeted computational experiments, we establish causal links between specific visual functions [4, 26, 27] and solutions, validate aspects of eye evolution as trade-offs between light collection and spatial precision [4, 5], and study relationships between eyes design, neural processing, and visual tasks. Concretely, our scientific contributions are: (1) By strictly changing the visual task an agent is subject to, orientation (NAVIGATION) task vs. object discrimination (DETECTION) task [4, 27], we observe a bifurcation in our evolved agents between camera-type and compound-type eyes (Figure 1a); (2) We show that the emergence of optical structures, such as focused lensing from primitive simple eyes [5], is a key innovation that addresses the fundamental trade-off between light collection and spatial precision (Figure 1b); (3) We reveal that sensory acuity and neural capacity scales as a power-law, where decreasing task error requires complimentary improvement in both dimensions — consistent with observations made in animal vision and AI [28, 29]. Additionally, our engineering contributions are: (1) A framework that evolves embodied agents through genetic algorithms and deep reinforcement learning in a custom simulation framework; (2) A genetic encoding scheme for vision that describes a diverse set of eyes and cognitive capabilities in addition to being physically-based and realizable.

Due to biological complexity and computational tractability, we scope our work to recreate the system-level process of vision evolution rather than an imitation of its historical timeline. Since our goal is to understand the principles driving vision evolution (not imitate evolution's exact path), we computationally recreate essential elements that shaped natural vision. We model key components universal to biological evolution (Figure 1.d), agent's anatomy (Figure 1.e),

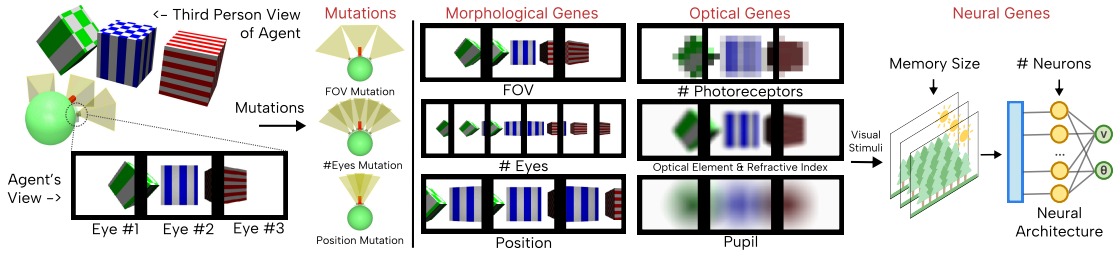

Fig. 2: **Our genetic encoding enables vision to evolve computationally.** Our encoding mirrors the natural separation between sensory and neural development through three gene clusters. *Morphological genes* control eye placement and field of view. *Optical genes* determine visual sensing capabilities (# photoreceptors, optical elements, pupil size). *Neural genes* describe the structure of the underlying processing mechanisms. These independently mutable traits enable the computational exploration of evolutionary pathways that mirror the pathways which shaped biological vision.

and consult biological studies [27, 30] when designing the simulated environments (Figure 1.f). Additionally, when studying optical transitions, we use physics-based approximations that are widely-used (Figure 6), a dynamics engine [31] for interaction, and train our agents only through sensory feedback using reinforcement learning. While our computational framework represents a novel approach to exploring vision evolution, fundamental limitations remain. For example, complex biological phenomena is poorly understood such as eye genomics and development [32, 33], bio-physical models [34], or mechanistic neural circuits [16]. Moreover, modeling evolutionary dynamics is often akin to modeling a chaotic system [35] as eye adaptions are often a result of many interdependent pressures. However, our results highlight how embodied agents trained via deep reinforcement learning can serve as scientific instruments to understand biological visual intelligence.

2 Results

Our computational framework tests hypotheses about how specific environmental pressures shape eye morphologies, neural architectures, and behavior. While previous work has used evolutionary algorithms to independently design optical systems [36, 37], embodied agent morphologies [8, 38, 39], or visually-guided behaviors [40], our approach uniquely evolves embodied agents with both eyes and behaviors *together* in a hierarchical approach. This enables the automatic task-driven discovery of diverse vision-based embodied agents.

2.1 A *What if ...?* World

We model the world in which embodied agents interact as a deep reinforcement learning environment, where agents must evolve appropriate vision capabilities such that they can learn effective behavior. It's infeasible, however, to model *all* factors which contribute to the evolution of vision; thus, we model each environment as corresponding to a specific *function* of vision, such as simple light detection or object discrimination. In this way, each environment represents a single task which models the functional pressures hypothesized to garner the emergence of vision. We focus on modeling three distinct tasks to isolate their effects on vision evolution [2, 27]: Navigation, Detection, and Tracking. We create these tasks in a MuJoCo simulation environment [31, 41] with custom features to support complex imaging models and evolutionary search. Each agent in this environment is modeled as a point mass (the green sphere in Figure 2) with a heading and forward velocity. For more technical details, please see Section 4.

2.2 Genetic Encoding for Vision

Vision in nature has co-evolved as a function of sensing and the underlying neural circuitry; subsequently constraining behavior that an animal learns during its lifetime [2, 4, 42]. At a population scale, this continuous feedback loop between evolution and learning ensures that learned behaviors affect the evolved genetic traits of future generations. Similarly, we create a genotype that directly encodes both the physical (morphological and optical) and the neurological components of an agent's vision. Rather than incorporating neural network weights directly in the genomic encoding [43–45], we train the agent each agent from scratch. The genetic encoding is discussed in more detail in Section 4.

Similar to nature, our genetic encoding scheme needs to be general enough to describe a diverse set of eyes and cognitive capabilities while being physically realizable. Therefore, we conceptualize an agent's eye as a physical sensor that converts photons into neural impulses. These impulses are then followed by cognition, which enables agent's to interact within the environment. We categorize the genotype into three broad subspaces (Figure 1e,f): morphological, optical, and neural. Each subspace describes a subset of an agent's visual system, where each gene is mutated through specific evolutionary operators (Figure 2). We provide more details in Section 4.

2.3 Co-Evolution of Vision and Behavior

Our approach computationally mimics nature's co-evolution approach to vision innovation: changes in sensory capabilities directly influence behavioral performance, which in turn guides the evolution of future eye and network designs. We implement co-evolution through two nested loops that mirror the interplay between evolutionary timescales and lifelong adaptation. Over generations, the outer evolutionary loop utilizes the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) [46, 47] to enable efficient selection and mutation of populations of agents. Within each generation, the inner learning loop, an agent with the selected genotype is instantiated and trained to solve a visual task through reinforcement learning via Proximal Policy Optimization (PPO) [48]. During the agent's lifetime, the agent's performance is evaluated within the same visual task. The fitness of each agent is then used in the following generation to selectively adapt the populations genotypes. We discuss the evolution and learning loops in more detail in Section 4.

2.4 What if vision was only needed to discern food from poison?

Understanding how specific visual tasks shaped the evolution of eyes remains a major challenge because animals are required to solve multiple visual tasks simultaneously. For instance, honeybees have evolved compound eyes with around 5,000 individual receptors, balancing trade-offs between extracting optic flow to maintain equidistance from obstacles and regulate flight speed, and sufficient spatial resolution to discern body movements of other bees in their colony [30]. This coupling of tasks in nature makes it difficult to understand how individual visual demands influence eye evolution. For instance, while dragonflies primarily use compound eyes for navigation, they have also evolved high-resolution regions for prey detection, making it challenging to identify which visual adaptation was a result of which environmental pressure. Thus, would evolution converge on similar eye morphologies as found in nature if we could isolate individual visual tasks? To address this, we create two distinct environments that isolate specific visual demands, allowing us to observe how eye morphologies evolve when optimizing for a single task.

Our computational framework tests vision evolution through two distinct tasks that isolate different environmental pressures. The NAVIGATION task is a goal-less orientation task ([27]) where agents are incentivized to traverse a maze environment as fast as possible while avoiding collisions with walls and forward barriers which are alternating with white/black striped patterns of different frequencies (similar to navigational setups that test navigational abilities within

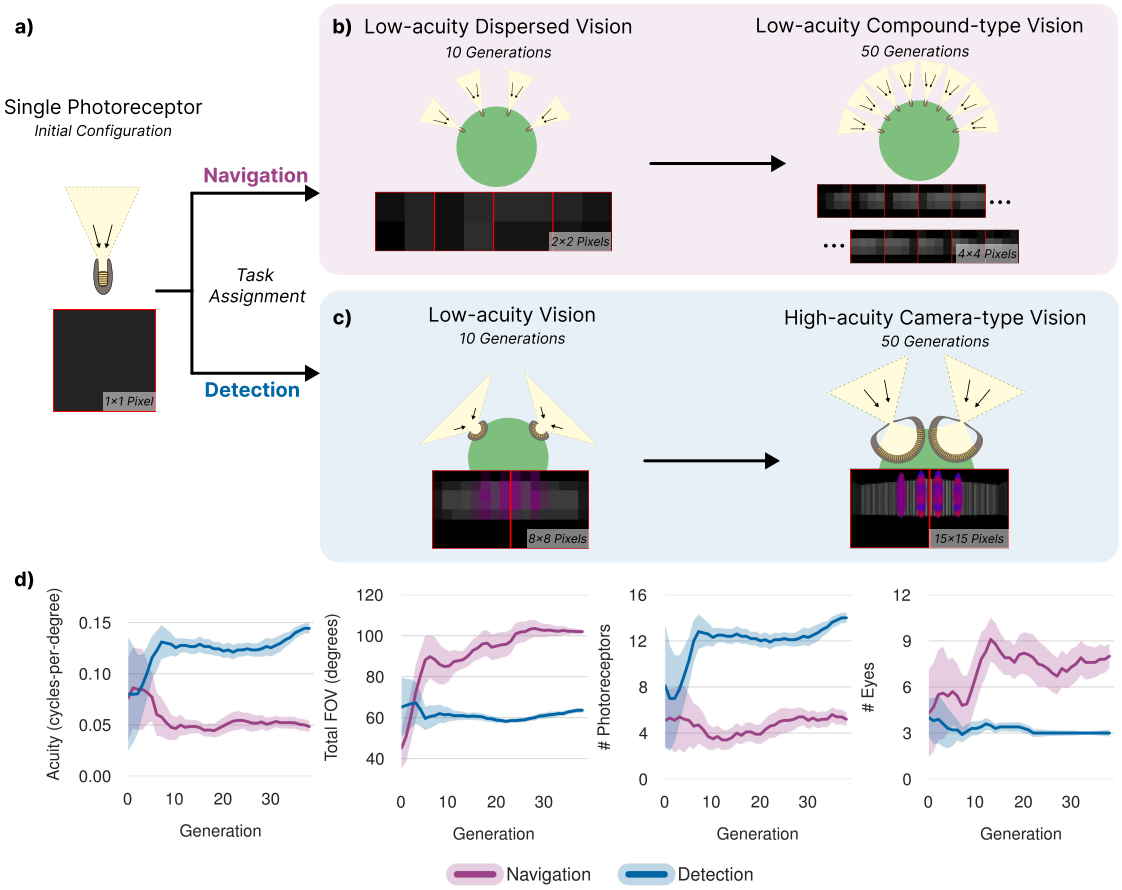

Fig. 3: **Low- and high-acuity spatial tasks lead to compound and camera eyes, respectively. (a)** We initialize a population of agents for two visual tasks (DETECTION and NAVIGATION) with a single eye with one photoreceptor. We then evolve a population of agents subject to morphological mutations: add photoreceptor, add eye, and adjust placement. In the NAVIGATION task, we first observe an emergence of dispersed vision, where many eyes are employed. By 50 generations, a compound-type eye emerges; that is, a vision system consisting of 10 individual eyes, each with four photoreceptors, distributed over the entire diameter of the agent. **(c)** In the DETECTION task, we initially observe the emergence low-resolution vision. After 50 generations, the population has converged on a morphology consisting of two forward facing, high-resolution camera-type eyes. **(d)** Configuration vs generation plots are shown, depicting the evolutionary progression of the mean agent in the population and the task dependence on evolutionary adaptation. The plots show the mean and 95% bootstrapped confidence interval, respectively.

honeybees [30]). Conversely, the DETECTION task is an object discrimination task where agents choose the goal sphere between three visually similar spherical objects in an open environment (this can be conceptualized as identifying food from poison with the only difference being the rotation of a high frequency spherical pattern on the sphere). In both environments, agents control only their forward speed and heading. Both tasks start are initialized by generating a population of agents via randomly mutating the genotype of a primitive agent (a single eye with one photoreceptor and a field-of-view of 100°). Over the course of evolution, agents mutate by using only morphological constraints: adding or remove photoreceptors, adding or removing eyes, and adjusting the eye's placement. Agent fitness in NAVIGATION scales with completion speed and collision avoidance, while DETECTION fitness rewards quick, accurate food selection

with evolutionary termination for selecting poison (the exact reward functions are described in Section 4)

From an initial configuration of one eye with a single photoreceptor, agents evolve distinctly different morphologies under each task. To enforce realistic physical constraints, we limit the allowed configurations to prevent overlap based on photoreceptor size and eye radius. Similar to flying insects that navigate complex environments at high speeds [30], our navigation-specialized agents evolved compound-type eyes with 8 widely distributed visual units, each containing 5 photoreceptors (Figure 3b). This configuration optimizes full-body coverage with a 270° total field-of-view, enabling rapid environmental sampling during high-speed maze traversal. In contrast, detection-specialized agents develop two forward-facing eyes with 15 photoreceptors each (Figure 3c), concentrating their visual resources frontally. This evolutionary divergence reflects task-specific optimization: navigation agents maximize spatial awareness through distributed low-resolution sensing, while detection agents sacrifice peripheral vision for enhanced frontal acuity, enabling object discrimination at greater distances within their fixed time constraint.

Our agent's morphological divergence directly influences the topology of their neural network used to learn task behavior. A compound eye configuration in our genetic scheme enables parallel processing of visual information — each additional eye gets its own visual processing unit (MLPs) that handles processing of a specific part of the visual field before the low dimensional features from each eye are concatenated. Conversely, in the DETECTION task, agents evolve camera-type vision (2-3 forward facing eyes) with with larger input arrays (15x15x3) per eyes.

Our computational isolation of visual tasks reveals fundamental patterns that parallel natural evolution [1, 46, 49–51]. The emergence of distinct morphologies from identical starting conditions demonstrates how environmental demands can drive visual specialization. Like bees using wide-field motion detection [30], our navigation agents evolved distributed sensing for efficient environmental sampling. When quantifying visual capabilities using cycles-per-degree measurements [29, 52, 53], we found a clear trade-off between spatial coverage and resolution that mirrors natural systems (Figure 3d). This trade-off manifests across species [13], where animals evolve either enhanced acuity or broader fields of view based on their ecological needs. Our results provide computational evidence that task-specific selection pressures can drive the emergence of these distinct visual and neural processing strategies.

2.5 What if eyes never evolved optical element like lenses?

Early visual systems faced a fundamental trade-off between light collection and acuity, progressing from simple light-sensitive patches to cup-shaped eyes with smaller apertures [54, 55]. While decreasing the size of the aperture and creating pinhole-like designs is a straightforward way to improve image formation, they severely limit light collection. This trade-off creates a performance ceiling, where further improvements in spatial resolution through pinhole designs are limited by the lack of light. This inherent limitation ultimately restricts the visual capabilities of such systems, causing a saturation in performance. We see this manifested in our results where agents with pinhole eyes plateau in fitness and are not able to achieve the performance benefits of improved spatial resolution. What if we introduced optical elements or lenses into our agents? Lenses emerged as a innovation in biological evolution and we investigate the impact of enabling lensing in our framework.

We investigate here whether artificial evolution would replicate these major transitions in eye morphology, tracking the emergence of optical structures from simple light-sensitive patches to complex lens-based systems. To isolate optical evolution, we restrict mutations to the optical subspace: pupil size, optical element, and refractive index. We fixed the remaining morphological genes to the parameters evolved in the DETECTION task in Figure 3. Pupil size controls the signal-to-noise (SNR) ratio by controlling the total light throughput on the retina; since the the noise in the environment is fixed, SNR decreases as pupil size decreases. The optical element is represented as a 2D array that can be programmed into lenses of different shapes (modeled as a

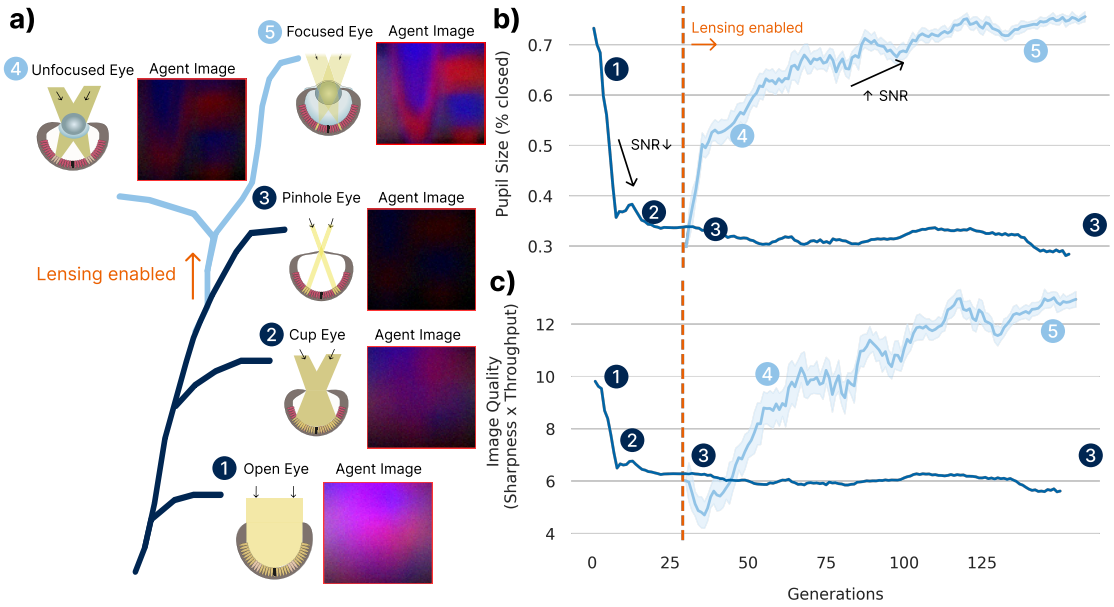

Fig. 4: **Evolution of eye morphology reveals how lensing resolves a fundamental trade-off in vision.** We demonstrate that to achieve maximum fitness in the DETECTION task, evolution learns to evolve optical structures against two competing objectives: achieving high spatial precision and maximizing light collection. **(a)** The evolutionary sequence shows five key stages of eye development: (1) open pupil with maximum light collection but poor spatial precision, (2) cup eye and (3) pinhole eye that achieve better spatial precision by reducing pupil size at the cost of light collection, followed by the emergence of (4) unfocused and (5) focused lens-based eyes that maintain spatial precision by evolving optical structures while allowing larger pupils for more light collection. Agent images show the scene as perceived at each stage. **(b)** Without optics (dark blue), pupil size decreases to improve precision, sacrificing the signal-to-noise ratio (SNR). When lensing is enabled (orange line, generation 30), larger pupils emerge as lenses are evolved maintain precision while increasing light throughput. **(c)** The Image Quality metric (image sharpness × light throughput) quantifies this trade-off resolution: pinhole eyes (3) plateau at low values due to limited light collection, while lens-based eyes (4,5) achieve higher quality by combining good spatial precision with larger pupils. This mirrors the evolutionary pressure that drove the emergence of biological lenses, which enabled enhanced vision across lighting conditions.

diffractive optical element (DOE)) [56–58]. Refractive index controls the bending of light within the optical element. These three parameters are general enough to be represent a large number of different lens shapes. We discuss these parameters, the physics-based rendering model, and their relation to real eyes in Section 4.

We conduct two large-scale evolution experiments within the DETECTION task: Phase I, a baseline evolution with only pupil size mutations; Phase II, a counterfactual study where we enable optical element mutations and refractive index mutations after 30 generations of Phase I. After the experiment concludes, we perform an analysis of the evolved agents' vision systems with the "Image Quality" metric, which is defined as the product of spatial precision and maximum light throughput. Spatial precision is determined using the Modulation Transfer Function (MTF) of the eye's point spread function (PSF). We refer to Appendix Section A for Fitness and MTF plots, PSNR and SSIM graphs (measures signal-to-noise ratio and structural information between agent and reference image), and 3D plots of evolved lenses.

In Phase I of the experiment, where we only enable pupil size mutations, we observe a clear evolutionary progression from open apertures (Figure 4.1) to cup eyes (Figure 4.2) and finally to pinhole eyes (Figure 4.3). As shown in the agent's image, the blur is significant for the open eye, and meaning the agent cannot discern between the objects reliably. To evolve cup eyes, agents with decreased pupil size are selected for. A smaller pupil restricts light throughput and FOV but reduces the overall blur of the image at the cost of decreased SNR (See Section A for PSNR/SSIM analysis). A pinhole eye is formed at (Figure 4.3), where the agent fitness saturates at around 30% of the maximum pupil size; the converged pupil size is dependent on the total ambient light in the environment, as in less light in the environment results in a evolved larger pupil.

In Phase II of the experiment, we enable DOE and refractive index mutations while maintaining pupil size mutations at generation 30. Initially, the image quality decreases as the random DOE shapes perform worse than pinhole eyes. Between generation 30 to 50, we observe a crucial transition: DOEs slowly evolve from a diffusing lens (surfaces that scatter light everywhere) to developing convex shapes that mark evolution's first step toward lens-like structures. These early convex surfaces can focus some light but remain unoptimized, creating unfocused eyes with larger pupils that collect more light than pinhole eyes (Figure 4.4). This intermediate stage demonstrates evolution discovering the basic principle of light focusing before convergence. As evolution progresses beyond generation 50, the DOE is refined into focused lenses and symmetric PSFs while maintaining large pupils (Figure 4.5), achieving both high spatial precision and improved light collection necessary for DETECTION. This optical improvement is reflected in multiple metrics: maximum agent fitness increases substantially from 15 to 25, while PSNR improves from 7.5 to 8.6 dB and SSIM rises from 0.15 to 0.275 (Section A). Figure A1 shows high-performing agents ($F>24.4$) evolved singular, smooth optical responses, while poor performers ($F<3.4$) developed fragmented, multi-peaked patterns.

Our results demonstrate a critical sequence that illuminates why lenses emerged over the course of vision evolution. While our agents were simply tasked with discriminating between similar-looking objects, it required the populations to evolve effective eyes subject to the fundamental trade-off in vision evolution: balancing spatial precision (needed to discriminate similar objects) with light collection (needed for reliable vision). Our Image Quality metric, which combines MTF-derived spatial precision with light throughput, quantifies this trade-off directly. In Phase I, where only pupil size could vary, spatial precision saturates as pinhole eyes sacrifice light collection for acuity, resulting in dark, noisy vision. Phase II reveals how lens evolution resolves this constraint: DOEs evolve into lenses that maintain spatial precision while allowing larger pupils, eliminating the precision vs. light throughput trade-off. Our counterfactual experiment suggests that without this innovation, accurate vision would have been restricted to high-light conditions; for example, while pinhole eyes can technically produce sharp images, they are rarely found in nature due to the decrease in light throughput. This aligns with theoretical predictions which suggest that lens development coincided with an increase in eye size [54, 55]. The lens thus represents a fundamental innovation in the evolutionary solution space, discovered by our agents not through direct optimization of optical properties, but through the demands of achieving accurate perception on a specific behavioral task.

2.6 What if animal brains stayed small throughout evolution?

Biological Visual intelligence emerges from the interplay and scaling between sensory hardware, morphology, and neural processing [59, 60]. While artificial intelligence relies on fixed sensors (RGB cameras) and scales with the number of parameters, nature has evolved diverse eye-brain systems that scale in complexity together to solve intelligent tasks. By varying eye acuity (cycles-per-degree) [61], neural network size, and temporal processing [62], we investigate how these resources shape the evolution of visual intelligence and task-specific performance in embodied agents.

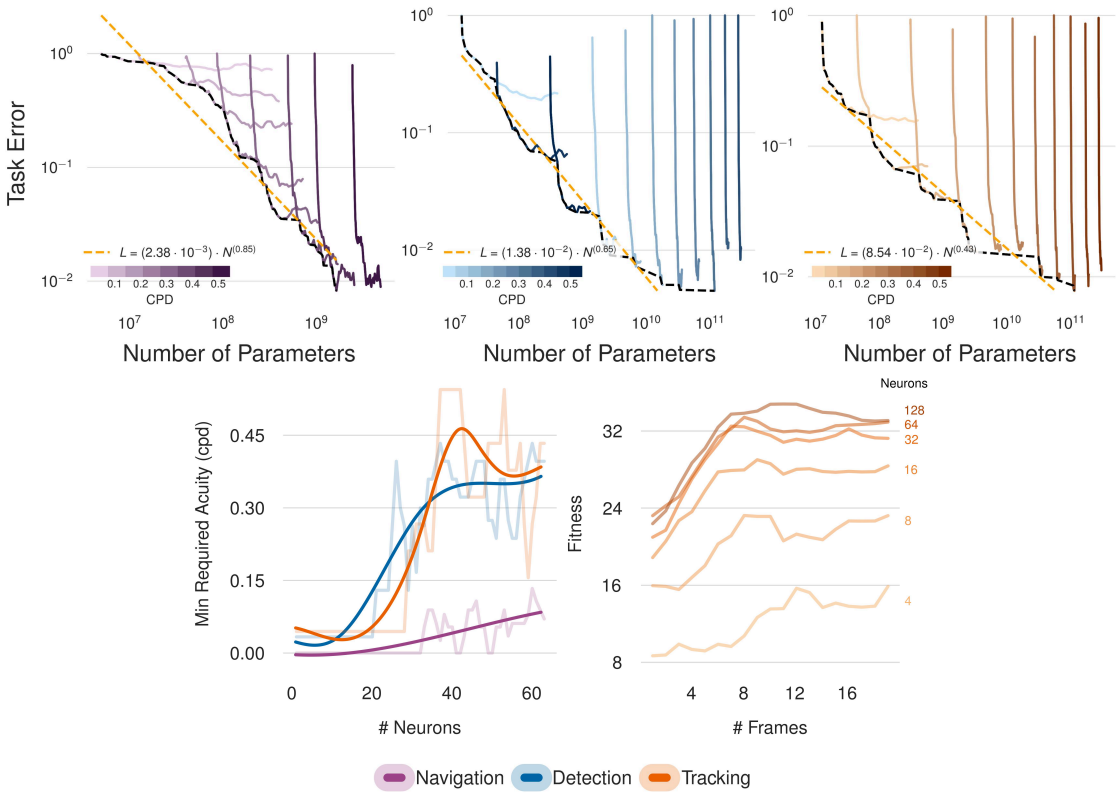

Fig. 5: **Task-dependent scaling laws reveal how sensory acuity bounds performance and how temporal memory compensates for neural capacity (a-c)** Our experiments reveal visual task-dependent power-law scaling between number of parameters and sensory acuity (CPD). This demonstrates that scaling in sensory input is required for embodied tasks to avoid a bottleneck that cannot be overcome by neural scaling alone. **(d)** Minimum required visual acuity versus number of parameters for different embodied tasks suggest a hierarchy in the task emergence. textbf(e) Temporal processing shows complementary scaling with neural capacity, where increased temporal memory (number of frames) can compensate for reduced neural processing, particularly evident in tasks with larger networks (128-32 neurons). Together, these results quantify how visual intelligence emerges from the interplay between sensory, neural, and temporal capabilities.

Our analysis reveals distinct power-law scaling between task performance and neural capacity across navigation, detection, and tracking tasks. Performance across network sizes (Figure 5.a-c) follows characteristic power laws ($L = (2.38 \cdot 10^{-3}) \cdot N^{0.85}$ for navigation, $L = (1.38 \cdot 10^{-2}) \cdot N^{0.85}$ for detection, and $L = (8.54 \cdot 10^{-2}) \cdot N^{0.43}$ for tracking). This power law defines a predictable improvement in task error as a function of increasing network size. But this trend is only persists when another quantity, the level of visual acuity, is also able to improve. Each acuity level bounds the maximum achievable task performance (minimum task error). Low acuity models hit performance ceilings, demonstrating that poor visual acuity creates a fundamental bottleneck that cannot be overcome by simply scaling neural capacity. This resource limitation mirrors the power laws for scaling seen in artificial intelligence systems, where performance is bounded by the interplay of computational resources, data availability, and model size [28].

These scaling relationships reveal how evolving morphological constraints like eye structure and number of parameters (brain size) act as fundamental resources that can affect scaling

in embodied agents performance. Sensory acuity (measured by CPD) is also a resource which limits the throughput of information from the scene to agent based on fundamental limitations of light transport. Our results demonstrate that power-law scaling only holds when both acuity and parameter resources scale appropriately, with performance saturating when either becomes a bottleneck (i.e. in the relationship between visual acuity and number of parameters increasing model size cannot overcome fundamental sensing limitations). However, biological evolution has repeatedly overcome such constraints through scaling across different genetic traits [59, 60] - suggesting parallel opportunities for artificial systems where scaling of data and parameters alone may be insufficient without corresponding scaling in sensory capabilities.

Critically, we also identify transition points that reveal fundamental limits in visual processing (Figure 5.d). For a fixed neural network size, increasing sensory acuity (CPD) beyond certain thresholds yields diminishing returns, with each task showing distinct saturation points. For example, the detection task requires higher minimum acuity to achieve comparable performance, suggesting a hierarchy in the task emergence and visual processing demands of different behaviors.

For time-oriented tasks like tracking, we uncover a compensatory relationship between neural processing and temporal memory (Figure 5.e), where increased temporal information can offset reduced neural processing capacity and vice versa. This demonstrates that equivalent task performance can be achieved either through sophisticated processing of individual frames or through extended temporal integration of simpler visual features. The shallower scaling exponent (0.43) for tracking compared to navigation and detection (0.85) suggests that temporal integration may provide an alternative pathway to improved performance beyond pure neural scaling.

3 Discussion

Similar to natural evolution, we follow a *function over form* approach, where we code the desired function through fitness and let evolutionary search discover a variety of forms that are optimal for the fitness. This results in our agent's form (eye design and learned behavior) to emerge solely from functional pressures of orientation and navigation or object discrimination. The emergent features resemble principles of real biological evolution. These results affirm our central claim that embodied agents trained via DRL can serve as hypothesis-testing machines for vision and vision evolution. The evolutionary outcomes we present are a result of the co-evolution of vision-hardware (physical eye morphology and structure) and software (learned behavior of the agent). Lastly, in our current approach, we evolve our agents under isolated environmental pressures i.e. cases where agents are heavily biased to evolve to solve a single task. However, in the natural world animals have evolved to jointly solve diverse tasks found in their ecological niches. While our framework can be easily extended for diverse visual tasks, isolated scenarios help us understand the extreme cases.

Since our work is the first in this space, our results point towards open technical challenges that will enable a wider variety of hypothesis to be tested. For instance, future research can be extended to incorporate explore multi-agent interactions where multiple species evolve in shared environments, applying gradient-based methods, or incorporating richer light properties like spectral, polarization, or temporal sampling. Additionally, future work could include incorporating bio-physical models of vision [34] or replace neural networks with mechanistic circuits derived from fly connectomes [16, 63].

Our framework provides a discovery tool by enabling large scale computational evolution of vision in embodied artificial agents. For biologists and cognitive scientists, this approach allows systematic manipulation of key variables to test alternative hypothesis or counterfactuals - such as isolating the effects of optical elements from neural processing, or testing how specific environmental pressures drive eye morphology. Much like natural evolution [64], our framework demonstrates remarkable creativity in discovering solutions - for instance, it independently evolved compound-eye architectures without being explicitly designed to do so. For engineers, these evolutionary simulations reveal design principles for artificial vision systems, particularly

valuable when optimizing for practical constraints like energy efficiency and manufacturability [11, 65].

4 Methods

Learning loop. The learning loop is the mechanism for which we score each agent. Via reinforcement learning, we train the brain of the agent (i.e., neural network parameters). Reinforcement learning serves as a mechanism for learning representations of the environment through interactions with it. The subsequent score, or fitness, of the agent is determined from the average reward it receives over six evaluation episodes after training. We utilize an open-source implementation of the Proximal Policy Optimization (PPO) algorithm [48, 66]. Hyperparameter values can be found in the Supplementary Material. Each agent has 1 million total training steps, though training may be terminated early if no improvement is found after five evaluations (which takes place every 50,000 training steps).

Reinforcement learning algorithms have been shown to have a strong dependence on the random seed used to initialize the environment [67]. Thus, during the evolution loop, we allow the same agent genotype to be sampled multiple times. Additionally, each agent's training loop is initialized with a unique random seed such that configurations sampled with the same genotype are not subject to the same seed. This allows for a more robust evaluation of the agent's performance.

Observations. An agent interacts with its environment through actions based on its observations. The observations are created by compositing the images captured from each eye. For example, if an agent's vision system consists of 5 eyes with four photoreceptors in each eye, the resulting observation by the full animal "eye" at each time step will be a tensor of size 5 x 4 x 1 x 3. Furthermore, for each eye, the previous observations are stacked in a memory buffer. If the memory buffer is of size 10, then a tensor if size 10 x 5 x 4 x 1 x 3 is provided to the underlying agent network. In addition to the visual stimuli, we provide a single boolean that describes whether an agent is in physical contact with an object at the current time step and the previous action that was taken. Although not needed for the agent to solve the tasks, we have found that providing contact information and previous action as observations led to convergence nearly twice as fast; in an evolutionary search context, this speed-up significantly improves the overall optimization time.

Reward function. The reward function is used in RL to drive policy optimization towards some desired observation and action mapping. In our case, each task has a unique reward function:

$$\begin{aligned}
\textsc{Navigation} \quad & R_{\textsc{Navigation}} = \quad \lambda\left(\|\mathbf{x}_t - \mathbf{x}_0\| - \|\mathbf{x}_{t-1} - \mathbf{x}_0\|\right) + w_g + w_c \\
\textsc{Detection} \quad & R_{\textsc{Detection}} = -\lambda\left(\|\mathbf{x}_t - \mathbf{x}_f\| - \|\mathbf{x}_{t-1} - \mathbf{x}_f\|\right) + w_g + w_a + w_c \\
\textsc{Tracking} \quad & R_{\textsc{Tracking}} = -\lambda\left(\|\mathbf{x}_t - \mathbf{x}_f\| - \|\mathbf{x}_{t-1} - \mathbf{x}_f\|\right) + w_g + w_a + w_c
\end{aligned} \tag{1}$$

where R_X is the reward at time t for each task, λ is a scaling factor, x_t and x_{t-1} is the position of the agent at time t and $t-1$ respectively, x_0 is the initial position of the agent, and x_f is the position of the goal (i.e., end of maze for Navigation, goal object in Detection). The w variables are non-zero when certain conditions are met. w_g and w_a indicates the reward/penalty given for reaching the goal and adversary, respectively. w_c is the penalty for contacting a wall. In essence, in the Navigation task, the agent is incentivized to move from it's initial position as fast as possible. In the Detection and Tracking tasks, the agent is incentivized to navigate to the goal as quickly as it can. During training, $\lambda = 0.25$, $w_g = 1$, $w_a = -1$, and $w_c = -1$. Additionally, when an agent reaches the goal or adversary, the episode terminates.

Fitness function. As compared to the reward function, the fitness function is used to evaluate the current performance of an agent. Where the reward function is used to inform the RL

algorithm for it's weight optimization, the fitness function informs the evolutionary search algorithm for further selection and mutation. For each task, the fitness function F_X in generation g is identical to R_X, except with different weights to emphasize the relative performance difference between agents. During fitness evaluation, $\lambda = 1.5$, $w_g = 10$, $w_a = -10$, and $w_c = -2$. Instead of terminating when the goal or adversary is reached, in evaluation, the object is respawned and the agent continues to solve the task.

Evolution loop. Evolving the physical and neural representation, each with $>\sim 10^{20}$ parameters, of hundreds of agents efficiently necessitates "intelligent" optimization. The vast size of the search space alone means all combinations cannot be tested in a timely fashion. Strategically selecting morphologies that simultaneously explore new configurations and exploit previously gained knowledge is imperative to not waste resources on suboptimal solutions. In ACI, we accomplish this intelligent search mechanism through the integration of evolutionary strategies (ES) [44]. ES is a broad optimization technique that is inspired by natural evolution and operates by iteratively refining a population of candidate solutions through processes such as mutation, selection, and adaptation. Unlike traditional genetic algorithms, ES emphasizes mutation over crossover and is particularly well-suited for optimizing continuous, high-dimensional spaces.

We use a population size of 16 agents, and evolve for 50 to 100 generations depending on the experiment. The specific ES algorithm we use is the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [68]. CMA-ES is a variant of ES that adapts the mutation distribution based on the covariance matrix of the population. This adaptation allows for faster convergence and better exploration of the search space. We use the open-source implementation of CMA-ES provided by nevergrad [47]. Hyperparameters for CMA-ES can be found in the Supplementary Material.

Agent's genotype. An agent's genotype encodes the language that is used to create the agent's vision. The overarching genotype compromises of a physical genotype (encodes the agent's eye properties) and a neural genotype (encodes the neural topology). The properties within each genotype are further divided into subspaces that can be mutated independently, and incorporate both continuous and discrete parameters. The subgenes are: morphological, optical, and neural. Rather than modeling complex biological mechanisms like photoreceptor dynamics, we implement these traits using a physics-based rendering model that captures the essential functional properties of vision while remaining computationally tractable. Our encoding scheme is capable of representing approximately 10^20 unique agent morphologies.

Morphological subspace. The morphological subspace defines properties used to spatially sample the environment, such as the number of eyes, their position/orientation, and field of view (FOV). We model the agent as a sphere of fixed radius with eyes distributed uniformly along its equator. Thus, the position and orientation of each eye is dependent on both the number of eyes, and a placement range (i.e., the maximum angle from latitude 0°) that eyes are uniformly distributed within. For instance, if an agent has 3 eyes and the placement range is 90°, the eyes are placed at -45°, 0°, and 45°. We assume bilateral symmetry, consistent with the observation that the overwhelming majority of animals have bilateral symmetry [69]. The FOV is a continuous decimal value which can be between 1° and 100°.

Optical subspace. The optical subspace describes how each eye interacts with incoming light in a physically plausible way. It encompasses a programmable Diffractive Optical Element, or a phase mask, that modulates the phase of the incoming light (represented as a 4×4 array with $\in [0, 1]$), pupil radius (scaled dynamically with sensor size as $r = a \times L$, where $a \in [0, 1]$ and L is the sensor size), and refractive index ($\eta \in [1.0, 2.0]$). We use continuous parameters for phase mask, pupil radius, and refractive index and then upsample the phase mask to a size of $51, 51$ for sharper PSFs. The optical subspace can also be disabled in which case a rasterization-based

imaging model is used to create visual stimuli for the agent; this model is analogous to an eye with a perfect lens, i.e no blur.

Neural subspace. The neural subspace defines the properties of the neural network, such as memory size, number of neurons, and the neural network architecture. The memory size represents the number of historical frames relative to the current frame the agent has access to; for instance, a memory size of five means the agent's visual stimuli is a flattened vector composed of the current frame and the previous five frames. The underlying neural network is a fully connected feedforward network with two identically sized hidden layers. The number of neurons in the hidden layers is an integer mutation parameter that can be between 1 and 512. Although we don't present experiments in this work evolving the neural network architecture, it is possible to mutate the underlying architecture (i.e., add layers, change activation functions, etc.) in our framework.

Mutation operators. The agent's genotype is designed with specific mutation operators for each parameter type. For continuous parameters (such as FOV, phase mask values, pupil radius scaling factor, and refractive index), mutations occur through Gaussian perturbation within their defined ranges. For discrete integer parameters (such as number of eyes and number of neurons), mutations increment or decrement the current value while respecting the parameter bounds. Bilateral symmetry is preserved during all morphological mutations.

Experimental Control. In each experiment, we enable specific mutations to isolate and study specific aspects of vision evolution. For example, the number of eyes in the morphological subspace is represented as an integer parameter that can be mutated. This controlled approach allows us to systematically investigate the evolution of different visual traits.

Agent phenotype. An agent's phenotype is the physical manifestation of its genotype. The phenotype is the realized form that interacts within the environment that acquires and acts on observed stimuli. An agent in this work is represented as a fixed radius sphere with eyes distributed uniformly along its equator. The output of the underlying policy is direction and speed, which are used to actuate the joints to move the agent in the simulation environment. In the case of the TRACKING task, we assign computed action profiles to the goal and adversary to move to random locations within the environment.

Imaging model. Our model consists of a lens described by a phase mask, refractive index, and amplitude modulating (aperture) elements. We model a retina emulated by a discretized pixel at the sensor plane at a focal length distance away from the pupil (Figure 6).

All imaging systems capture the scene as an optically encoded image on to the sensor plane. These optical encodings are commonly referred to as the blur or point spread function (PSF), and are dependent on the phase and amplitude of the pupil function along with wavelength and depth of the scene point. We follow the wave propagation model described in [71–73] to estimate the depth in-dependent PSF.

Given a point light source at a distance z and the pupil function $P(x, y) = A\exp(i\phi)$ (Figure 6) the response of the agent's eye can be measure by the PSF. The PSF at the sensor plane s distance away from the pupil plane is described as:

$$PSF_{\lambda,z}(x', y') = \left|\mathcal{F}^{-1}\left\{\mathcal{F}\left\{P(x, y)U_{in}(x, y)\right\} H_s(f_x, f_y)\right\}\right|^2, \tag{2}$$

where $H_s(\cdot)$ represents the field propagation transfer function [74] for distance s with (f_x, f_y) as the spatial frequencies given as

$$H_s(f_x, f_y) = \exp\left[iks\sqrt{1 - (\lambda f_x)^2 - (\lambda f_y)^2}\right]; \tag{3}$$

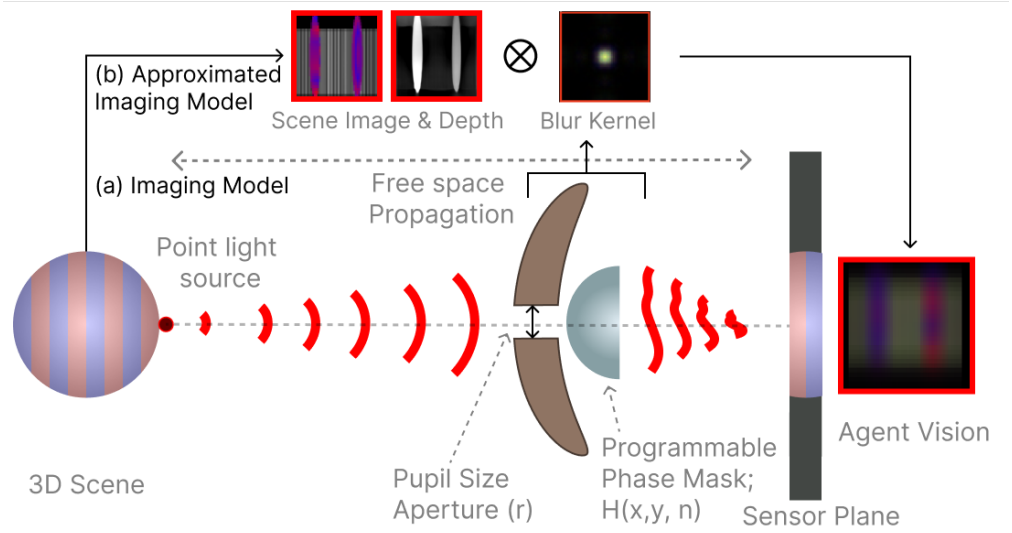

Fig. 6: **Imaging Model.** Our simulation implements both wave and geometric optics using a OpenGL [70] **(a)** Our scene imaging model shows how a depth dependent blur kernel is derived: light from a point source in the 3D scene propagates through free space, passing through a pupil plane which is composed of (1) an aperture with a variable aperture radius (r) and (2) a programmable phase mask with height map $H(x, y)$ and refractive index n, before forming an image on the sensor plane. **(b)** The approximated model uses a 2D convolution between the scene, a depth map, and a single blur kernel using the far-field approximation (i.e., the Point Spread Function) for computational efficiency.

where $k = \frac{2\pi}{\lambda}$ is the wavenumber; $U_{in}(x, y)$ denotes the complex-valued wave field immediately before the lens which for a point light source is given as

$$U_{in}(x, y) = \exp\left(ik\sqrt{x^2 + y^2 + z^2}\right); \tag{4}$$

$\mathcal{F}\{\cdot\}$ is the 2D Fourier transform; (x', y') are the spatial coordinates on the camera plane, and (x, y) are the coordinates on the lens plane.

The phase modulation function $t_\phi(x, y) = e^{i\frac{2\pi}{\lambda}\phi(x,y)}$ in Equation (2) is generated by the lens surface profile $\phi(x, y)$ which in our case is a square 2D phase-mask array of size 25 pixels that is mutated by the outer evolution loop, where $\phi(x, y) \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. These values are scaled appropriately based on the agent eye's refractive index.

Finally, our agent's image formation follows a shift-invariant convolution of the image and the depth-independent PSF to yield the final image, I_ℓ, perceived by the agent.

$$I_\ell = \mathcal{S}_\ell(H_\ell * X_\ell) + N_\ell, \tag{5}$$

where the sub-index ℓ denotes the color channel; $X_\ell \in \mathbb{R}_+^{w\times h}$ represents the underlying scene with $w \times h$ pixels; H_ℓ represents the discretized version of the PSF in Equation (2); $N_\ell \in \mathbb{R}^{w\times h}$ denotes the Gaussian noise in the sensor; $\mathcal{S}_\ell(\cdot) : \mathbb{R}^{w\times h} \to \mathbb{R}^{w\times h}$ is the camera response function, modeled as a linear operator; and $*$ denotes the 2D convolution operation.

In practice, the discretized version of the PSF is of size $(H+1, W+1)$ where (H, W) is the resolution of the agent's eye I_ℓ as chosen by the evolutionary search. This is an explicit choice to make the PSF larger than the image to enable a full blur on the eye when the aperture is fully open.

The scene image X_ℓ is rendered by padding I_ℓ of size (H, W) to $\big(H + (2 * \frac{H+1}{2}), W + (2 * \frac{W+1}{2})\big)$. This enables the corner pixels to accumulate light from areas directly due to the aperture size which is more physically-based. This also means that closing the aperture also helps with reducing the total effective field of view of the agent's eye, which is how the agent controls the blur in Phase I experiment of Section 2.5. For a pinhole eye, the field-of-view becomes equivalent to the encoded fov in the agent's morphological gene.

Quantifying vision capabilities. We quantify an agent's morphological phenotype and optical phenotype in cycles per degree (cpd) and the Modulation Transfer Function, respectively. The cycles per degree is a measure of the spatial frequency observable to the imaging system; a higher cpd value corresponds to a better ability to resolve fine spatial details and distinguish closely spaced features in the visual scene [29]. CPD is also a commonly used metric to measure visual capabilities in real-life animals [29]. The Modulation Transfer Function (MTF) quantifies how well different spatial frequencies are preserved by an optical system. Unlike CPD, which represents a limit on resolvable detail, MTF describes the gradual degradation of contrast across spatial frequencies in a given optical system. We analyze agents' performance relative to its MTF in more detail in the Appendix.

Simulated environment. This framework is built on top of a the MuJoCo physics engine [31] and within a gymnasium-style [41] setup. The underlying dynamics of each agent is governed by the MuJoCo physics and images are rendered via a rasterization pipeline using the builtin OpenGL renderer.

In this work, we differentiate between a visual task and an environment. A visual task is the specific goal an agent is trying to achieve, such as light seeking or object tracking. The environment is the physical space in which the agent is placed, and so can contain multiple tasks. For instance, the same environment is used to train agents on both the detection and tracking tasks; the only difference is the reward function and the movement of the goal/adversary. In addition to the physical positioning of objects in an environment, the textures, light, colors, etc. can be modified to create a diverse set of various environments.

Each environment in the experiments presented here have walls, which can are organized as boundary or in a maze-like configuration. These walls are rigid and contactable, and provide barriers where the agent cannot move escape.

In any one environment, there is only ever one agent which is considered trainable. Trainable agents have observations and an underlying neural network which is optimized during training. Conversely, non-trainable agents are static or privileged (privileged in that they can access all information about an environment at any point in time), and have a fixed action policy. An example policy for a non-trainable agent is the goal object in the TRACKING task; it's policy is to continuously move to a random location within the environment.

Appendix A Lensing Analysis

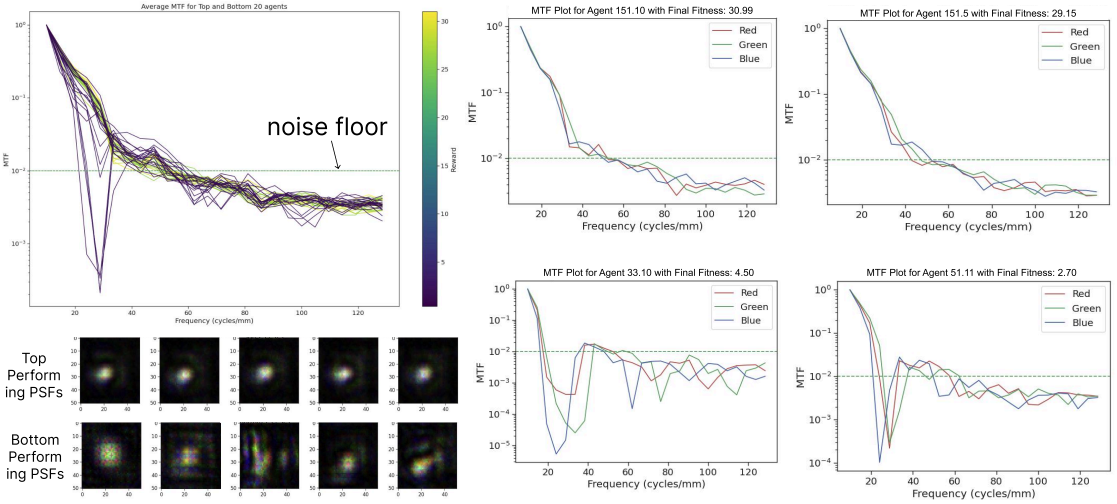

Fig. A1: **Modulation Transfer Function (MTF) analysis of evolved vision systems:** We analyze the spatial frequency response of evolved vision systems using MTF curves, which quantify how well different spatial frequencies are preserved. Left: Average MTF curves for top 25 and bottom 25 performing agents, colored by reward. The horizontal dashed line indicates the noise floor, below which spatial information becomes unreliable. Bottom: Point Spread Functions (PSFs) for the best and worst performing agents, showing the characteristic light distribution patterns. Top Performing Agents develop compact and symmetric PSFs even though we don't enforce any symmetry in our setup. Right: Individual MTF curves for agents at different evolutionary stages (Generation 33, 51, and 151) and performance levels. Early generations (Gen 32, 51) show erratic frequency responses with significant dips, indicating poor optical performance. By Generation 151, high-performing agents develop smooth MTF curves that maintain good contrast above the noise floor up to 40 cycles/mm, demonstrating evolution of effective lens-based vision systems. The RGB channels show similar responses, suggesting achromatic optimization of the optical system.

Our analysis quantifies the optical performance of evolved vision systems using three metrics. The Point Spread Function (PSF) represents the system's response to a point source of light - how a perfect point gets "spread out" by the optical system. In Figure A3, high-performing agents develop compact, symmetric PSFs indicating precise light focusing, while poor performers show diffuse, irregular patterns suggesting inefficient light management. A perfect PSF would appear as an infinitesimally small point, while real optical systems produce some degree of spread due to diffraction and optical imperfections.

The Modulation Transfer Function (MTF), mathematically derived as the Fourier transform of the PSF, quantifies how well different spatial frequencies are preserved by the optical system. On the MTF plots in Figure A3, the y-axis represents contrast preservation (from 0 to 1) while the x-axis shows spatial frequency in cycles/mm. The area above the noise floor (10^{-2}) represents useful spatial information - frequencies where the signal can be reliably distinguished from noise. Early-generation agents show erratic MTF curves with sharp dips below this noise floor, while later-generation agents maintain smooth curves above the noise floor up to 40 cycles/mm.

The PSNR and SSIM curves in Figure A2 reveal limitations of these conventional metrics. Initially, with fully open apertures, both metrics show high values because extreme blur acts as a noise-reducing low-pass filter. However, this blurred vision makes the detection task impossible,

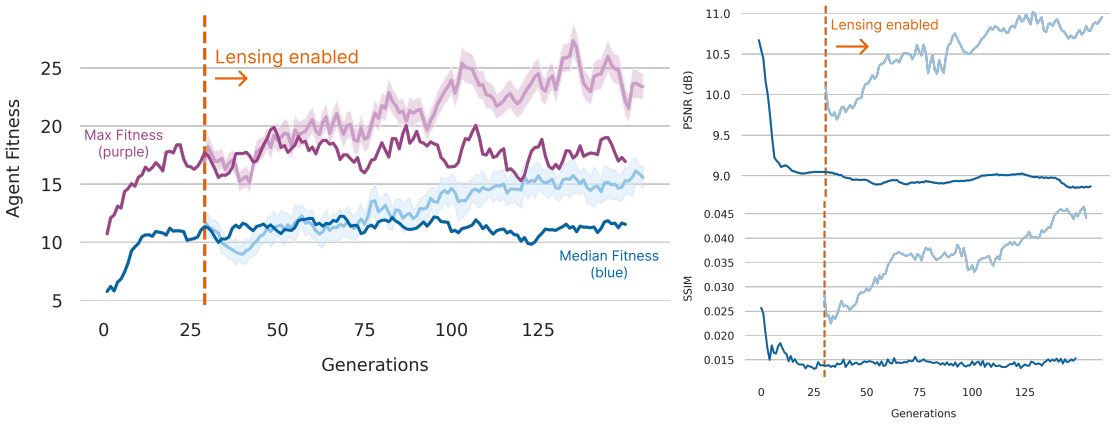

Fig. A2: **Agent Fitness, PSNR (dB) and SSIM over generations:** We show additional graphs from Section 2.4 that show steady increase in agent fitness. The main plot tracks both maximum (purple) and median (blue) agent fitness, which initially plateau around generation 25 due to pinhole eye limitations. After enabling lensing, maximum fitness shows steady improvement, reaching significantly higher levels compared to the pre-lensing phase. To compute SSIM and PSNR we render a rastered version of the image using the pinhole camera model as the reference image. For PSNR we show a substantial increase from the pinhole eyes. This shows that lensing significantly improves the signal-to-noise tradeoff that we discussed in Section 2.4. We can also compare the SSIM between Phase I and II. Unlike PSNR, SSIM is a perception-based model that considers changes in structural information between reference and target images. The SSIM increases are little as it relies on pixel wise calculations but the trend demonstrates that lens-based eyes better preserve image structure while maintaining higher light collection compared to pinhole eyes, enabling more reliable discrimination between similar visual features.

resulting in low agent fitness. As apertures begin to close (forming pinhole eyes), PSNR and SSIM decrease as the system preserves more high-frequency information but with increased noise due to limited light collection. When lensing is enabled at generation 30, we observe steady improvement in both metrics as the system evolves the ability to maintain high-frequency detail while collecting sufficient light.

These observations led us to develop an Image Quality metric that multiplies two factors: (1) the area under the MTF curve above the noise floor (marked in Figure A3 by the dashed noise floor line), representing spatial precision, and (2) light throughput, which decreases quadratically with smaller apertures. This metric captures the trade-off between spatial precision and light collection. While pinhole eyes can achieve good MTF performance, their limited light throughput constrains their overall image quality. Lens-based eyes resolve this trade-off by maintaining strong MTF performance while allowing larger apertures for better light collection.

Notably, some agents with excellent optical properties (high MTF and light throughput) show slightly lower rewards due to the stochastic nature of reinforcement learning. This variance in reward despite similar optical quality suggests that the relationship between optical performance and task success is not purely deterministic - better vision enables but does not guarantee better task performance.

Appendix B Sampling from the Design Space

The genetic encoding for vision can be fundamentally understood as operating on the plenoptic function, which describes the complete flow of light in a scene [75]. While our current implementation demonstrates proof-of-concept by allowing evolution to explore a subset of plenoptic

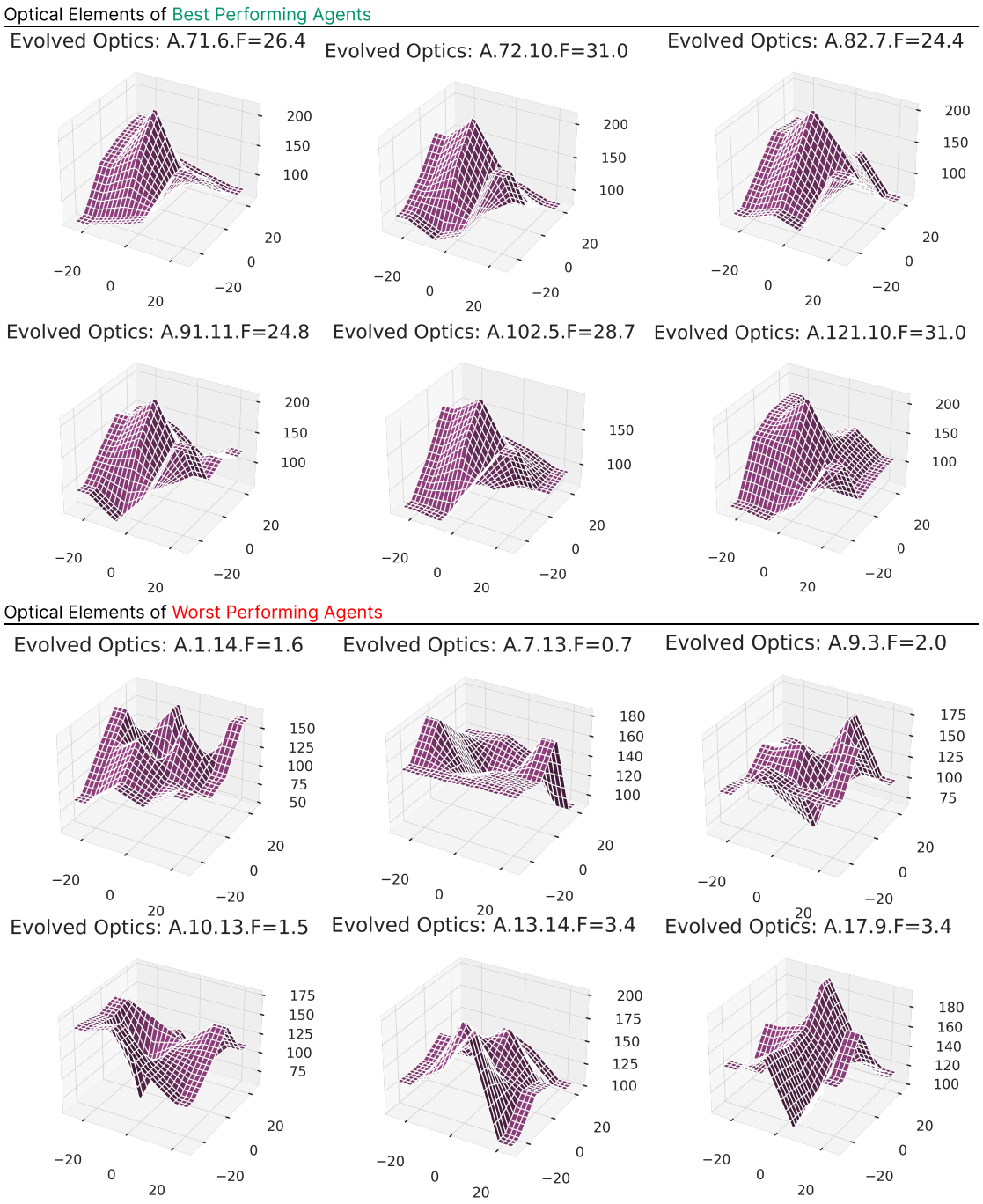

Fig. A3: **Comparison of evolved optical elements between best and worst performing agents.** Three-dimensional surface plots showing the optical response patterns for the top six (upper panel) and bottom six (lower panel) performing agents. Best performing agents (F=24.4-31.0) exhibit well-defined, singular peak formations with smooth gradients, while worst performing agents (F=0.7-3.4) display irregular, multi-peaked patterns with abrupt transitions. Each plot represents a unique evolved optical configuration denoted by its agent identifier (A) and corresponding fitness score (F).

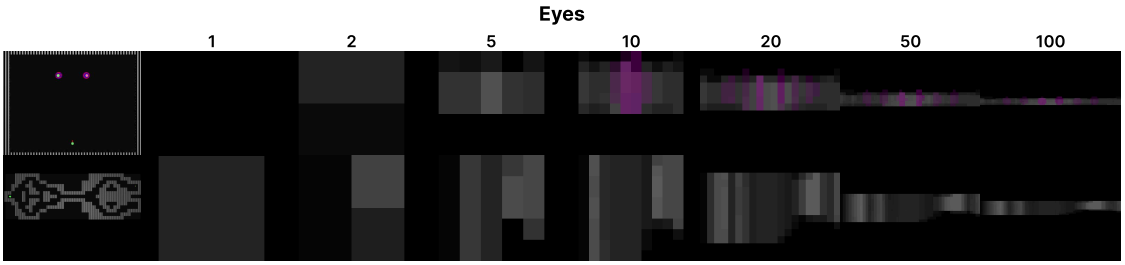

Fig. B4: **Sampling greater number of eyes by modifications to the Morphological Gene:** We display our agent's vision captured with progressively more number of eyes. Top row shows progressively increasing eyes allows for larger FOV of the scene and creates multiple copies of the spheres from slightly different perspectives in each eye. The bottom row shows that for the navigation task number of eyes allows the agent to see different parts of the wall which it uses to orient itself against wall collisions.

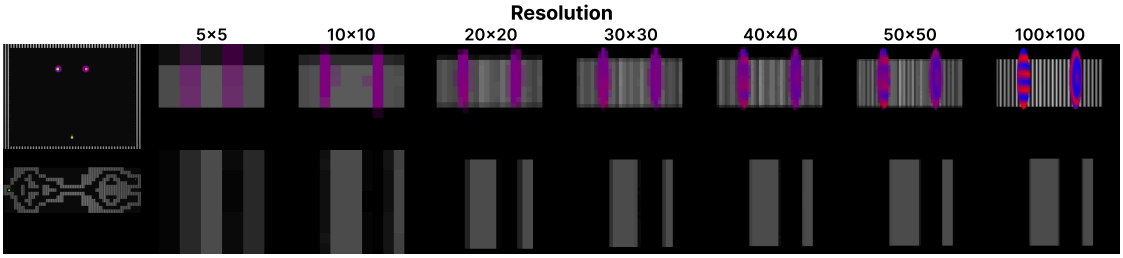

Fig. B5: **Sampling larger resolutions by modifications to the Optical Gene:** We display our agent's vision captured with progressively larger resolutions. Top row shows progressively increasing resolution resolved the difference between food and poison which can be differentiated with the orientation of the stripes. The bottom row shows that for the navigation task resolution helps resolve the stripes on the wall.

dimensions such as placement, optics constraints, movement of the agent etc. , the framework can naturally extend to encompass the full plenoptic representation of light - including spectral sensitivity, polarization detection, and varied spatiotemporal resolutions. Just as our computational experiments have shown evolution discovering diverse and creative solutions within a limited set of visual parameters, expanding the genetic language to sample from the complete plenoptic dimensions would enable the discovery of even more sophisticated visual systems, analogous to those found in nature. For instance, the mantis shrimp (stomatopods) evolved 16 different photoreceptor types that can detect both linear and circular polarized light [76], while jumping spiders (Salticidae: Dendryphantinae) developed a unique combination that provide both high acuity and wide-field motion detection [77]. Our language enables co-evolution of eyes, neural circuitry and subsequent behavior (learnt through reinforcement learning) and provides a unified way to think about vision evolution as a creative optimization process operating directly on the fundamental properties of light. As we expand the available plenoptic dimensions in our genetic language, we expect to see the emergence of increasingly sophisticated and novel visual systems that may parallel, or even exceed, the remarkable diversity found in biological evolution.

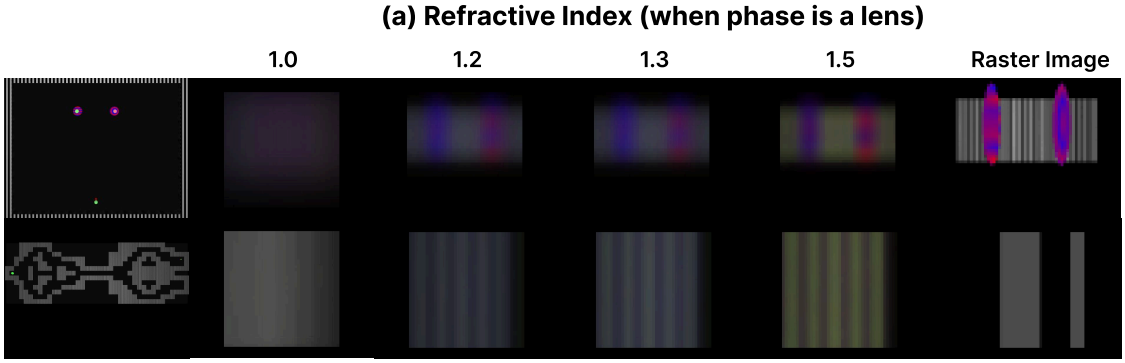

Fig. B6: **Sampling refractive indices in the Optical Gene:** We illustrate examples of sampled refractive indices within the Optical Gene for Detection (top row) and Navigation(bottom row) tasks. For a fixed phase mask (a perfect lens) increases in refractive index causes increase sharper images.

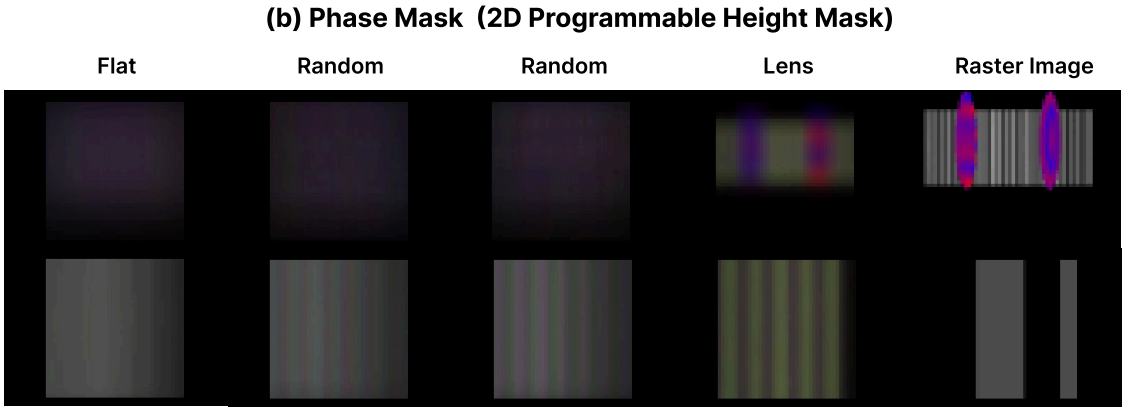

Fig. B7: **Sampling optical elements in the Optical Gene:** We illustrate examples of sampled optical elements (phase masks) within the Optical Gene for Detection (top row) and Navigation(bottom row) tasks. The figure shows Flat, and 2 Randomly samples phase masks which shows the complexity of the design space. These visualizations also demonstrate that while using a lens is a major innovation in eye design, creating a focused lens is a hard problem that evolution solved really well.

Appendix C Acuity-Neural Processing Trade-offs and Task-Specific Scaling

In our framework, we systematically explore how visual task performance emerges from the interplay of three key components. The first component is the eye's physical characteristics, measured in cycles-per-degree (CPD, ranging from 0.0056 to 0.5444), which determines the ability to resolve spatial detail. The second is neural capacity, where we vary the number of parameters in the vision-processing layers (ranging from 800 to 98,000 parameters). Our parameter sweep reveals emergent power-law scaling relationships between sensory acuity and neural capacity Figure C9. The relative fitness plots (top row) demonstrate that navigation achieves high performance (>0.8) at lower CPDs (0.05) with modest neural capacity (8000 parameters). Detection and tracking tasks show a distinct scaling pattern, requiring both higher CPDs (>0.3) and larger networks (>40,000 parameters) for comparable fitness levels. The error plots (bottom row) reveal

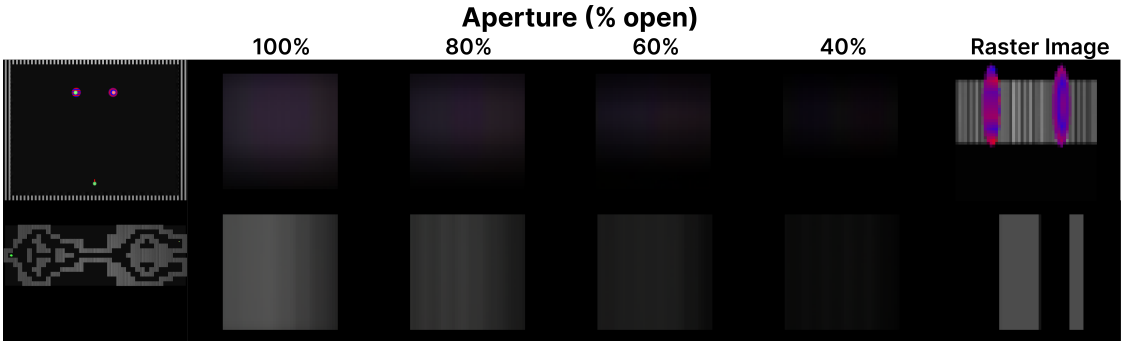

Fig. B8: **Sampling Different Aperture by modifications to the Optical Gene:** We display our agent's vision captured with progressively smaller apertures, demonstrating how reducing the aperture size leads to increased image sharpness. However, as the aperture closes, the signal strength decreases quadratically with its radius, leading to higher noise levels. The balance between sharpness and noise is a critical factor for agents to successfully complete their visuomotor tasks.

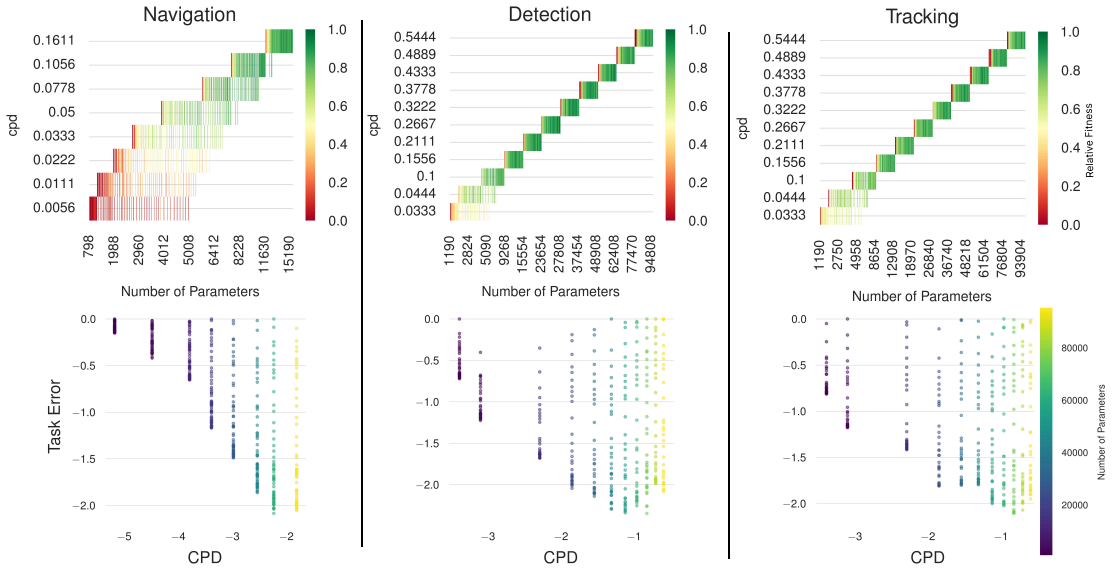

Fig. C9: **Dense parameter analysis revealing task-specific relationships between sensory acuity and neural processing. Top:** Performance visualization showing individual trials (vertical lines) across CPD values and network sizes for navigation (left), detection (middle), and tracking (right). **Bottom:** Corresponding scatter plots with log-scaled axes demonstrate how error rates vary with CPD for different network sizes (indicated by color intensity). The distinct patterns across tasks support our findings about task-dependent scaling relationships between sensory and neural resources.

fundamental constraints in how these capabilities emerge. At fixed CPD values, increasing neural capacity follows characteristic power-law improvements until hitting task-specific performance ceilings. These ceilings are particularly evident in the scattered error distributions, where higher CPDs enable lower minimum error rates across all tasks. This demonstrates that poor visual acuity creates a fundamental bottleneck that cannot be overcome by simply scaling neural capacity. Notably, detection and tracking display continuous improvements in error rates as both CPD and

network size increase, suggesting these tasks benefit from simultaneous scaling of both sensory and neural resources. These computational scaling relationships emerged spontaneously through evolution in our framework, revealing how physical constraints in sensory acuity interact with neural processing capacity to shape task performance. The distinct scaling patterns across tasks, particularly the earlier saturation in navigation compared to detection and tracking, suggest a natural hierarchy in the visual processing demands of different behaviors [27]. This emergent relationship between sensory hardware and neural processing mirrors both biological evolution [59, 60] and contemporary artificial intelligence scaling laws [28, 78, 79].

For temporal performance, we find that performance saturates beyond 10 frames across all configurations. This is particularly evident in tracking tasks, where agents are incentivized to complete objectives quickly, typically achieving success in under 10 frames. This reveals an optimal balance between temporal information and computational efficiency that varies by task complexity.

References

[1] Land, M., Nilsson, D.-E.: Animal Eyes. Oxford University Press, United Kingdom (2002)

[2] Nilsson, D.-E.: The diversity of eyes and vision. Annual Review of Vision Science **7**(Volume 7, 2021), 19–41 (2021) https://doi.org/10.1146/annurev-vision-121820-074736

[3] Fernald, R.D.: Casting a genetic light on the evolution of eyes. Science **313**(5795), 1914–1918 (2006)

[4] Nilsson, D.-E.: The evolution of eyes and visually guided behaviour. Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1531), 2833–2847 (2009)

[5] Nilsson, D.-E., Pelger, S.: A pessimistic estimate of the time required for an eye to evolve. Proceedings of the Royal Society of London. Series B: Biological Sciences **256**(1345), 53–58 (1994)

[6] Walter, W.G.: A machine that learns. Scientific American **185**(2), 60–64 (1951)

[7] Holland, O.: The first biologically inspired robots. Robotica **21**(4), 351–363 (2003)

[8] Nolfi, S., Bongard, J., Husbands, P., Floreano, D.: Evolutionary Robotics. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32552-1_76 . https://doi.org/10.1007/978-3-319-32552-1_76

[9] Winfield, A.F.: Evolutionary robotics as a modelling tool in evolutionary biology. Frontiers in Robotics and AI **11**, 1278983 (2024)

[10] Krause, J., Winfield, A.F., Deneubourg, J.-L.: Interactive robots in experimental biology. Trends in ecology & evolution **26**(7), 369–375 (2011)

[11] Lipson, H., Pollack, J.B.: Automatic design and manufacture of robotic lifeforms. Nature **406**(6799), 974–978 (2000)

[12] Baldwin, J.M.: A new factor in evolution. The American Naturalist **30**(354), 441–451 (1896)

[13] Cronin, T.W., Johnsen, S., Marshall, N.J., Warrant, E.J.: Visual Ecology. Princeton University Press, Princeton, NJ (2014)

[14] Bongard, J.C.: Evolutionary robotics. Commun. ACM **56**(8), 74–83 (2013) https://doi.org/

10.1145/2493883

[15] Trianni, V.: Evolutionary robotics: model or design? Frontiers in Robotics and AI **1**, 13 (2014)

[16] Lappalainen, J.K., Tschopp, F.D., Prakhya, S., McGill, M., Nern, A., Shinomiya, K., Takemura, S.-y., Gruntman, E., Macke, J.H., Turaga, S.C.: Connectome-constrained networks predict neural activity across the fly visual system. Nature, 1–9 (2024)

[17] Floreano, D., Nolfi, S.: Adaptive behavior in competing co-evolving species. In: 4th European Conference on Artificial Life, pp. 378–387 (1997)

[18] Miras, K., Ferrante, E., Eiben, A.E.: Environmental influences on evolvable robots. PloS one **15**(5), 0233848 (2020)

[19] Ferrante, E., Turgut, A.E., Duéñez-Guzmán, E., Dorigo, M., Wenseleers, T.: Evolution of self-organized task specialization in robot swarms. PLoS computational biology **11**(8), 1004273 (2015)

[20] Waibel, M., Floreano, D., Keller, L.: A quantitative test of hamilton's rule for the evolution of altruism. PLoS biology **9**(5), 1000615 (2011)

[21] Waibel, M., Keller, L., Floreano, D.: Genetic team composition and level of selection in the evolution of cooperation. IEEE transactions on Evolutionary Computation **13**(3), 648–660 (2009)

[22] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., Hassabis, D.: Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm (2017). https://arxiv.org/abs/1712.01815

[23] Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz, F.J., Schrittwieser, J., Swirszcz, G., *et al.*: Discovering faster matrix multiplication algorithms with reinforcement learning. Nature **610**(7930), 47–53 (2022)

[24] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., *et al.*: A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science **362**(6419), 1140–1144 (2018)

[25] Mankowitz, D.J., Michi, A., Zhernov, A., Gelmi, M., Selvi, M., Paduraru, C., Leurent, E., Iqbal, S., Lespiau, J.-B., Ahern, A., *et al.*: Faster sorting algorithms discovered using deep reinforcement learning. Nature **618**(7964), 257–263 (2023)

[26] Warrant, E., Nilsson, D.-E.: Invertebrate Vision. Cambridge University Press, ??? (2006)

[27] Nilsson, D.-E.: The evolution of visual roles – ancient vision versus object vision. Frontiers in Neuroanatomy **16** (2022) https://doi.org/10.3389/fnana.2022.789375

[28] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)

[29] Caves, E.M., Brandley, N.C., Johnsen, S.: Visual acuity and the evolution of signals. Trends in ecology & evolution **33**(5), 358–372 (2018)

[30] Srinivasan, M.V., Zhang, S., Lehrer, M., Collett, T.: Honeybee navigation en route to the goal: visual flight control and odometry. Journal of Experimental Biology **199**(1), 237–244 (1996)

[31] Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: IROS, pp. 5026–5033. IEEE, ??? (2012)

[32] Gehring, W.J.: The evolution of vision. WIREs Developmental Biology **3**(1), 1–40 (2014) https://doi.org/10.1002/wdev.96 https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wdev.96

[33] Gehring, W.J., Ikeo, K.: Pax 6: mastering eye morphogenesis and eye evolution. Trends in genetics **15**(9), 371–377 (1999)

[34] Aguirre, G.K.: A model of the entrance pupil of the human eye. Scientific reports **9**(1), 9360 (2019)

[35] Doebeli, M., Ispolatov, I.: Chaos and unpredictability in evolution. Evolution **68**(5), 1365–1373 (2014)

[36] O'Shea, D.C.: Monochromatic quartet: a search for the global optimum. In: Lawrence, G.N. (ed.) 1990 Intl Lens Design Conf, vol. 1354, pp. 548–554. SPIE, ??? (1991). https://doi.org/10.1117/12.47896 . International Society for Optics and Photonics. https://doi.org/10.1117/12.47896

[37] Gagné, C., Beaulieu, J., Parizeau, M., Thibault, S.: Human-competitive lens system design with evolution strategies. Applied Soft Computing **8**(4), 1439–1452 (2008) https://doi.org/10.1016/j.asoc.2007.10.018 . Soft Computing for Dynamic Data Mining

[38] Gupta, A., Savarese, S., Ganguli, S., Fei-Fei, L.: Embodied intelligence via learning and evolution. Nature communications **12**(1), 5721 (2021)

[39] Sims, K.: Artificial evolution for computer graphics. In: Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '91, pp. 319–328. Association for Computing Machinery, New York, NY, USA (1991). https://doi.org/10.1145/122718.122752 . https://doi.org/10.1145/122718.122752

[40] Floreano, D., Kato, T., Marocco, D., Sauser, E.: Coevolution of active vision and feature selection. Biological cybernetics **90**, 218–228 (2004)

[41] Towers, M., Kwiatkowski, A., Terry, J., Balis, J.U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al.: Gymnasium: A standard interface for reinforcement learning environments. arXiv preprint arXiv:2407.17032 (2024)

[42] Baden, T.: From water to land: Evolution of photoreceptor circuits for vision in air. PLOS Biology **22**(1), 3002422 (2024) https://doi.org/10.1371/journal.pbio.3002422

[43] Stanley, K.O., Miikkulainen, R.: Evolving neural networks through augmenting topologies. Evolutionary Computation **10**(2), 99–127 (2002) https://doi.org/10.1162/106365602320169811

[44] Salimans, T., Ho, J., Chen, X., Sidor, S., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864 (2017)

[45] Conti, E., Madhavan, V., Such, F.P., Lehman, J., Stanley, K.O., Clune, J.: Improving Exploration in Evolution Strategies for Deep Reinforcement Learning via a Population of Novelty-Seeking Agents (2018). https://arxiv.org/abs/1712.06560

[46] Hansen, N., Ostermeier, A.: Adapting arbitrary normal utation. Proceedings of the IEEE International Conference on Evolutionary Computation, 312–317 (1996)

[47] Rapin, J., Teytaud, O.: Nevergrad - A gradient-free optimization platform. GitHub (2018)

[48] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. CoRR **abs/1707.06347** (2017) 1707.06347

[49] Wehner, R.: Desert ant navigation: how miniature brains solve complex tasks. Journal of Comparative Physiology A **189**, 579–588 (2003)

[50] Jeffery, W.R.: Regressive evolution in astyanax cavefish. Annual review of genetics **43**(1), 25–47 (2009)

[51] Hogg, C., Neveu, M., Stokkan, K., Folkow, L., Cottrill, P., Douglas, R., Hunt, D., Jeffery, G.: Arctic reindeer extend their visual range into the ultraviolet. Journal of Experimental Biology **214**, 2014–2019 (2011) https://doi.org/10.1242/jeb.053553

[52] Burton, R.F.: The scaling of eye size in adult birds: relationship to brain, head and body sizes. Vision research **48**(22), 2345–2351 (2008)

[53] Brooke, M.d.L., Hanley, S., Laughlin, S.: The scaling of eye size with body mass in birds. Proceedings of the Royal Society of London. Series B: Biological Sciences **266**(1417), 405–412 (1999)

[54] Schwab, I.R.: The evolution of eyes: major steps. the keeler lecture 2017: centenary of keeler ltd. Eye **32**(2), 302–313 (2018)

[55] Nilsson, D.-E.: Eye evolution and its functional basis. Visual neuroscience **30**(1-2), 5–20 (2013)

[56] Martel, J.N.P., Müller, L.K., Carey, S.J., Dudek, P., Wetzstein, G.: Neural sensors: Learning pixel exposures for hdr imaging and video compressive sensing with programmable sensors. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(7), 1642–1653 (2020) https://doi.org/10.1109/TPAMI.2020.2986944

[57] Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., Veeraraghavan, A.: Phasecam3d—learning phase masks for passive single view depth estimation. In: 2019 IEEE International Conference on Computational Photography (ICCP), pp. 1–12 (2019). IEEE

[58] Sitzmann, V., Diamond, S., Peng, Y., Dun, X., Boyd, S., Heidrich, W., Heide, F., Wetzstein, G.: End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. ACM Trans. Graph. **37**(4) (2018) https://doi.org/10.1145/3197517.3201333

[59] Allometry: The study of biological scaling. (2018). https://api.semanticscholar.org/CorpusID:199531454

[60] Venditti, C., Baker, J., Barton, R.A.: Co-evolutionary dynamics of mammalian brain and body size. Nature Ecology & Evolution **8**(8), 1534–1542 (2024)

[61] Caves, E.M., Brandley, N.C., Johnsen, S.: Visual acuity and the evolution of signals. Trends in Ecology & Evolution **33**(5), 358–372 (2018) https://doi.org/10.1016/j.tree.2018.03.001

[62] Abrams, R.A., Weidler, B.J.: Trade-offs in visual processing for stimuli near the hands. Attention, Perception, & Psychophysics **76**, 383–390 (2014)

[63] Shinomiya, K., Nern, A., Meinertzhagen, I.A., Plaza, S.M., Reiser, M.B.: Neuronal circuits integrating visual motion information in drosophila melanogaster. Current Biology **32**(16), 3529–3544 (2022)

[64] Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P.J., Bernard, S., Beslon, G., Bryson, D.M., Chrabaszcz, P., Cheney, N., Cully, A., Doncieux, S., Dyer, F.C., Ellefsen, K.O., Feldt, R., Fischer, S., Forrest, S., Frénoy, A., Gagné, C., Goff, L.L., Grabowski, L.M., Hodjat, B., Hutter, F., Keller, L., Knibbe, C., Krcah, P., Lenski, R.E., Lipson, H., MacCurdy, R., Maestre, C., Miikkulainen, R., Mitri, S., Moriarty, D.E., Mouret, J.-B., Nguyen, A., Ofria, C., Parizeau, M., Parsons, D., Pennock, R.T., Punch, W.F., Ray, T.S., Schoenauer, M., Shulte, E., Sims, K., Stanley, K.O., Taddei, F., Tarapore, D., Thibault, S., Weimer, W., Watson, R., Yosinski, J.: The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities (2019). https://arxiv.org/abs/1803.03453

[65] Klinghoffer, T., Tiwary, K., Behari, N., Agrawalla, B., Raskar, R.: Diser: Designing imaging systems with reinforcement learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23632–23642 (2023)

[66] Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N.: Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research **22**(268), 1–8 (2021)

[67] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D.: Deep reinforcement learning that matters. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[68] Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. Evolutionary computation **9**(2), 159–195 (2001)

[69] Holló, G., Novák, M.: The manoeuvrability hypothesis to explain the maintenance of bilateral symmetry in animal evolution. Biology Direct **7**, 1–7 (2012)

[70] Woo, M., Neider, J., Davis, T., Shreiner, D.: OpenGL Programming Guide: the Official Guide to Learning OpenGL, Version 1.2. Addison-Wesley Longman Publishing Co., Inc., ??? (1999)

[71] Tasneem, Z., Milione, G., Tsai, Y.-H., Yu, X., Veeraraghavan, A., Chandraker, M., Pittaluga, F.: Learning phase mask for privacy-preserving passive depth estimation. In: European Conference on Computer Vision, pp. 504–521 (2022). Springer

[72] Chang, J., Wetzstein, G.: Deep optics for monocular depth estimation and 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10193–10202 (2019)

[73] Chang, J., Wetzstein, G.: Deep Optics for Monocular Depth Estimation and 3D Object Detection (2019). https://arxiv.org/abs/1904.08601

[74] Goodman, J.W.: Introduction to Fourier Optics. Roberts and Company publishers, ??? (2005)

[75] Adelson, E., Bergen, J.: The plenoptic function and the elements of early vision (1997)

[76] Marshall, N.J.: A unique colour and polarization vision system in mantis shrimps. Nature **333**(6173), 557–560 (1988)

[77] Land, M.: Structure of the retinae of the principal eyes of jumping spiders (salticidae: Dendryphantinae) in relation to visual optics. Journal of experimental biology **51**(2), 443–470 (1958)

[78] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L.: Training Compute-Optimal Large Language Models (2022). https://arxiv.org/abs/2203.15556

[79] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., Zhou, Y.: Deep Learning Scaling is Predictable, Empirically (2017). https://arxiv.org/abs/1712.00409