Paleoinspired Vision: From Exploring Colour Vision Evolution to Inspiring Camera Design

Yijie Lu[1] **Zhimin Zong**[2] **Junjie Zhang**[3]
Shenghan Su[4] **Zeyun Zhao**[5] **Ziteng Cui**[4,6] **Yan Pu**[7]
Jing Lu[7] **Daisuke Kojima**[6] **Lin Gu**[8*] **Ruogu Fang**[5]

[1]Peking University [2]University of Florida [3]The University of Hong Kong
[4]The Hong Kong University of Science and Technology (Guangzhou)
[5]Vanderbilt University [6]The University of Tokyo
[7]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences
[8]Tohoku University

Abstract

The evolution of colour vision is captivating, as it reveals the adaptive strategies of extinct species while simultaneously inspiring innovations in modern imaging technology. In this study, we present a simplified model of visual transduction in the retina, introducing a novel opsin layer. We quantify evolutionary pressures by measuring machine vision recognition accuracy on colour images shaped by specific opsins. Building on this, we develop an evolutionary conservation optimisation algorithm to reconstruct the spectral sensitivity of opsins, enabling mutation-driven adaptations to more effectively spot fruits or predators. This model condenses millions of years of evolution into seconds using GPU, providing an experimental framework to test long-standing hypotheses in evolutionary biology, such as vision of early mammals, primate trichromacy from gene duplication, retention of colour blindness, blue-shift of fish rod and multiple rod opsins with bioluminescence. Beyond validating existing paleontological hypotheses, our model further enables speculative explorations of hypothetical species, such as organisms with visual systems adapted to Martian conditions. These constitute our contributions to AI for life science and neuroscience. Besides, the principles derived from these biological studies can translate into practical engineering solutions. In practice, our findings suggest a minimalist yet effective approach to task-specific camera filter design, optimising the spectral response function to meet application-driven demands. Code repository: `https://github.com/luyijie-research/PaleVision`.

1 Introduction

To suppose that the eye with all its inimitable contrivances, could have been formed by natural selection, seems, I freely confess, absurd in the highest possible degree.

- Charles Darwin "The Origin of Species"

Darwin argued that natural selection producing the eye 'can hardly be considered real.' However, modern biology has revealed eyes evolved not just once, but at least 40 different times [78]. Colour vision, in particular, reflects a tightly coupled interaction between environmental selection pressures and biological constraints, and continues to underpin modern camera design. As shown in Fig.1, early vertebrates exhibited tetrachromatic vision (4 colour vision) around 518 million years ago

*Corresponding author: `lin@tohoku.ac.jp`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

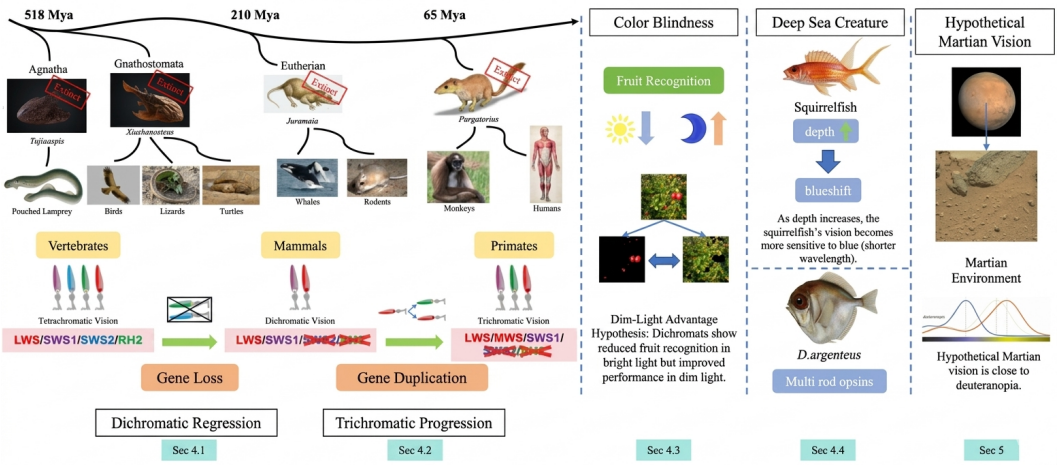


Figure 1: Our framework reconstructs the 518-million-year evolutionary trajectory of color vision across the vertebrate-to-primate lineage and evaluates corresponding explanatory hypotheses. This includes the transition from ancestral tetrachromacy through mammalian gene loss (dichromatic regression) to primate gene duplication (trichromatic progression). Furthermore, the framework quantifies the dim-light advantage hypothesis, deep-sea opsin blueshifts, and predictive visual adaptations and sensor designs for novel environments, such as the Martian landscape.

[102, 20]. Early mammals like Juramaia [54], which are only active in evening to avoid predation, then lost two colour in 210 million years ago, widely attributed attributed to this nocturnal lifestyles. Then, after an asteroid struct earth and wiped out dinosaurs approximately 65 million years ago, our primates ancestors returned to day life and regained the third colour vision[20, 40]. While prior palaeontological studies link evolutionary transitions to selection pressures, direct evidence is limited, and the timescales involved make these processes difficult to reproduce or quantitatively validate in controlled settings. Motivated by recent advances in reconstructing evolutionary trajectories of both extant and extinct specie through controlled robotic experimental platforms, we propose a computational framework called *evolutionary conservation optimization* to quantitatively analyse enduring palaeontological hypotheses and translate the resulting principles into practical camera design solutions.

Our framework, shown in Fig. 2, models visual evolution as a regularised optimisation problem over spectral representations. In Fig. 2.a, we represent visual input as wavelength-dependent light signals and introduce an opsin-inspired layer to simulate colour perception that performs spectral filtering through learnable sensitivity functions. Each opsin is parameterized as a smooth, Gaussian-like response with a constrained peak wavelength λ_{max}, ensuring biological plausibility[28, 53, 76]. The filtered responses are processed by a vision backbone (MiT-Encoder) to perform downstream tasks such as semantic segmentation, which we use as a proxy for evolutionary pressure, capturing survival-relevant behaviours such as foraging and predator avoidance.

We simulate evolution by optimizing the spectral parameters based on task performance. Small changes in λ_{max} represent mutation-like updates (Fig. 2.b), while the task objective defines the fitness landscape. To keep the evolution realistic, we apply conservative regularisation that encourages smooth and interpretable spectral changes (Fig. 2.c). This allows us to efficiently study how visual systems adapt to different environments in history. Beyond palaeontological analysis, we use the same framework to design cameras by changing the task objective in Fig. 2.a. By optimizing spectral parameters for application-specific environments and tasks, we obtain task-adaptive sensitivity functions that can be directly translated into physical optical filters.

Using this framework, we reconstruct key transitions in the evolution of colour vision, including the regression from tetrachromacy to dichromacy in early mammals and the re-emergence of trichromacy in primates. We further quantitatively analyse benefits of color deficiency under low-light conditions,

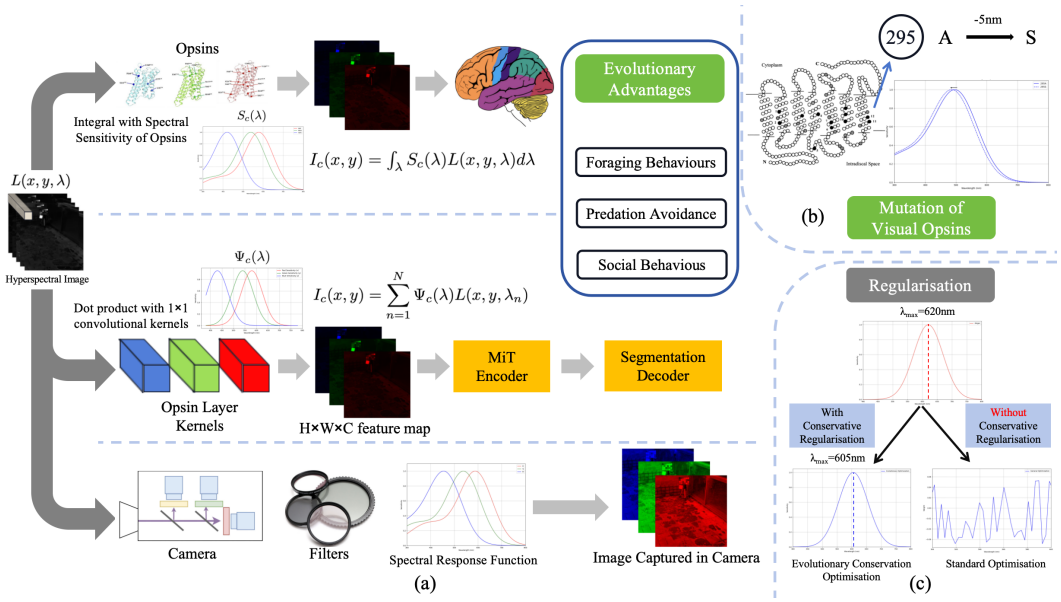


Figure 2: Computational framework for Evolutionary Conservation Optimization. The pipeline translates ecological pressures into semantic segmentation tasks, utilizing an opsin layer to bridge low-level spectral perception with high-level decision-making. By incorporating conservative regularization—parameterizing kernels as Gaussian functions with constrained λ_{max} shifts, the framework simulates the evolution and derive task-adaptive spectral response functions for specialized camera designs in Martian and medical applications.

adaptations in aquatic environments, such as the blue-shift of rod opsins with increasing depth and the emergence of multiple rod opsins in bioluminescent conditions. Together, these represent our contributions to the intersection of artificial intelligence with life science and neuroscience.

Beyond palaeontological analysis, the optimized spectral responses can be directly mapped to physical optical filters, enabling the design of task-adaptive camera system by optimizing spectral parameters for application-specific environments and tasks. We demonstrate this with proof-of-concept designs for 1) **Mars exploration**: By improving the segmentation of shadows and rocks on Mars, the designed filters show potential for cameras used in autonomous vehicle navigation in the Martian environment. 2) **Cancer detection**: Our design also demonstrates potential for medical-specific cameras capable of detecting abnormal tissue patterns associated with cancer. The evolutionarily grounded design paradigm is manufacturing-friendly, requiring only minor, easily implementable adjustments per iteration for rapid updates [88]. This aligns with minimalist camera design principles [44].

In summary, we present three main contributions:

- A computational framework combining opsin layers with evolutionary conservation optimization to model opsin evolution trajectories.
- An experimental platform for reconstructing and quantifying colour vision transitions in mammals, primates and deep-sea species.
- A minimalist, task-adaptive camera design methodology optimizing spectral response functions for Martian exploration and cancer detection applications.

2 Related Works

2.1 Relationship between Evolution Strategies and Gradient Descent

Recent research increasingly utilizes deep learning to simulate biological evolution [5]. Specifically, researchers have explored the intrinsic relationship between Evolution Strategies (ES) and gradient-based optimization [73]. While these paradigms are often considered fundamentally different, recent analytical work has demonstrated a formal correspondence between them; for instance, training

neural networks via conditioned stochastic mutation or neuroevolution is equivalent, in the limit of small mutations, to gradient descent on the loss function in the presence of Gaussian white noise [98]. This mathematical bridge suggests that gradient descent can be viewed as a simulator of mutation and selection pressure. Such a perspective aligns with findings that ES derive learning signals by stochastically perturbing parameters to perform randomized finite-difference estimation [74, 109]. To further integrate these fields, frameworks like Evolutionary Stochastic Gradient Descent (ESGD) combine SGD with gradient-free evolutionary steps to improve population fitness [18]. Similarly, the Gradient Evolution (GE) algorithm incorporates the Newton-Raphson method into evolutionary operators to maintain diversity and avoid local optima [45]. These hybrid approaches complement theoretical investigations suggesting that natural selection may implicitly approximate natural gradient descent on statistical manifolds [16]. Beyond weight optimization, evolutionary principles can synthesize entire network architectures, mimicking heredity to produce highly efficient models [79]. Collectively, these works reinforce the hypothesis that gradient descent is not merely an artificial optimizer but a process that can naturally emerge from evolutionary dynamics [80], justifying its use for simulating complex evolution via backpropagation in our framework.

2.2 Camera Design

Traditional RGB cameras mimic human colour perception [39] through multiple processing steps [41], but their spectral response functions are not optimized for specific imaging tasks [67]. Empirical studies have modeled camera color pipelines to link output to scene attributes [11], others have explored the potential of standard sensors [68] and joint sampling-reconstruction designs—using panchromatic-dominant patterns to maximize light intake and sharpness while suppressing aliasing [10]. Deep optics, which integrates optical design with deep learning [97], has been utilized to create higher-quality imaging systems [17, 52, 87, 92, 91, 83, 12, 89]. Pioneered by FlexISP [33], this paradigm enables end-to-end optimization of the entire imaging chain [93, 84]. Pre-sensor computing further accelerates these computations via optical neural networks (ONNs). Shi et al. [82] introduced a miniaturized lensless ONN, though it was limited by shallow layers. To overcome this, a multilayer ONN (MONN) was proposed [34], incorporating quantum dot films for non-linear activation [61] to enhance machine vision tasks. Klotz and Nayar introduced minimalist vision with freeform pixels [44], utilizing arbitrarily shaped pixels to capture only task-essential information for privacy-preserving, self-sustaining sensing. Complementing these architectural shifts, the Tianmouc vision chip [103] demonstrates the power of co-designing optics and computation, achieving 10,000 fps and 90% bandwidth reduction through dual-path processing for robust perception.

3 Proposed Approach

3.1 Overview

Fig.2 illustrates colour vision in vertebrates is mediated by opsins in the retina [21], each with a distinct spectral sensitivity function, $S_c(\lambda)$, quantifying light detection efficiency across wavelengths (λ) for each specific opsin channel c. Given the spectral signal at position (x, y), represented as $L(x, y, \lambda)$, the perceived intensity for each opsin is the integral of its spectral sensitivity over wavelength range, as shown by equation in Fig.2, generating a distinct signal channel for brain. We simplify this visual transduction using an *opsin layer* with c convolutional kernels, where each kernel's weight $\Psi_c(\lambda)$ represents the spectral sensitivity function. The opsin layer performs a 1×1 convolution, which is equivalent to integration above[67], transforming the $H \times W \times N$ hyperspectral images (HSI) into a size $H \times W \times C$ feature map (where $C = 3$ for trichromatic vision). HSI, collected by hyperspectral cameras[49, 7, 3], capture the spectral per pixel, where N represents the number of spectral bands (e.g. 400nm). The C channel maps are then passed to Mix Transformers (MiT) encoder for the recognition task.

3.2 Opsin Layer

The spectral sensitivity of opsins is commonly approximated using Gaussian functions, with a maximum sensitive wavelength defining their response to light [28, 53, 76]. Therefore, we proposed a simple yet efficient computational model, namely opsin layer to simulate the spectral sensitivity function shown in Fig.2. Since the integration of spectral and spectral sensitivity function of opsins equvalents to dot product between entries of the convolution kernel and the hyperspectral data,

we introduce a specialized 1×1 convolutional layer with c convolution kernels whose weight is parameterized by a Gaussian function, which acts like the spectral sensitivity of visual opsins.

$$\psi_c \in \left\{ \Psi \mid \Psi = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\lambda_i - \lambda_{max,c})^2}{2\sigma^2}}, i \in \{0, 1, \cdots, N\} \right\}, \quad c \in \{0, 1, \cdots, C\} \tag{1}$$

As shown in Eq.1, ψ_c denotes the c-th 1×1 convolution kernel weight and λ_i represents the wavelength corresponding to the i-th channel of HSI. $\lambda_{max,c}$ is the maximum sensitive wavelength and σ is the standard deviation of the Gaussian function.

3.3 Evolutionary Conservation Optimization

3.3.1 MiT Encoder and Segmentation Decoder

While any encoder architecture could process the opsin layer output [25, 26], we employ Mix Transformers [100] for optimal segmentation performance. The opsin output is tokenized into patches via linear projection. As Eq. 2 shows, the encoder E generates multi-resolution features Φ, with resolutions decreasing while channel dimensions increase.

$$E(Input_i) = \Phi = \{\Phi_1, \Phi_2, \Phi_3, \Phi_4\} \tag{2}$$

Unlike ViT's single-scale output [25, 14], MiT's hierarchical structure produces CNN-like multi-scale features [100], providing flexible representations for downstream tasks. This accommodates varying resolution requirements through progressive downsampling ($\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where $i \in \{1, 2, 3, 4\}$) and channel expansion ($C_{i+1} > C_i$). And we employ a lightweight all-MLP decoder[100] that utilises all multi-resolution features $\{\Phi_1, \Phi_2, \Phi_3, \Phi_4\}$ from the encoder.

3.3.2 Conservative Regularization

As illustrated in Fig. 2.c, direct optimisation of the opsin layer tends to yield unconstrained spectral sensitivities that diverge from biological reality. This discrepancy arises from two fundamental biological constraints: first, opsin sensitivities are typically characterized by Gaussian-like functions centered at the peak absorption wavelength, λ_{max} [53, 76, 28]; second, natural mutations at specific amino acid residues—known as spectral tuning sites (Fig. 2.b)—primarily induce shifts in λ_{max} ranging from 5 to 25 nm rather than altering the overall functional form [104, 77, 22].To bridge this gap, we introduce a conservative regularisation framework to govern the evolution of the opsin layer. Specifically, we parameterize the convolution kernel weights within the opsin layer as a Gaussian function, thereby restricting optimization exclusively to the adjustment of λ_{max}. To further ensure the optimization trajectory aligns with evolutionary dynamics, we limit the λ_{max} shift to a maximum of 0.5 nm per epoch through precise learning rate control, reflecting the incremental nature of natural mutational effects [104, 22].

Fig.2 shows evolutionary pressures and selective forces shaping vision arise from factors like food detection, predation avoidance, and social behaviour [9]. Furthermore, modern neuroscience [69] suggests that semantic segmentation embodies the core challenges of biological vision: bridging low-level perception (textures, edges, and spectral signatures) with high-level semantic reasoning (context, categories, and spatial relationships). Therefore, we employ the segmentation task by utilizing cross-entropy loss on relative datasets. Under our evolutionary conservation optimization, we concurrently train the opsin layer, transformer head, and segmentation decoder on HSI datasets.

4 Experimental Platform for Evolution

The number of visual opsins in vertebrate species has undergone two major evolutionary transitions: from four opsins in early vertebrates to two in the mammalian lineage, and from two to three opsins in the primate lineage [6]. Interestingly, colour blindness persists in certain primate populations, raising questions about its evolutionary significance [85]. In contrast, deep-sea creatures adapted to low light often possess only rod opsins [55], with many relying on a single rod opsin [63]. Shozo *et al.* [105] proposed the maximum sensitive wavelengths λ_{max} of rod opsin in squirrelfishes exhibit a

blue-shift trend as living depth increases, as shown in Fig.5.a. Notably, Musilova *et al.* [63] found that certain deep-sea teleost lineages have independently developed multiple rod opsins.

Our analysis begins by reconstructing opsin transitions in mammalian (Sec. 4.1) and primate lineages (Sec. 4.2). We then conduct three key quantitative investigations: 1) the advantages of colour blindness over normal vision (Sec. 4.3), 2) the depth-dependent blue-shift phenomenon in squirrelfish opsins and development of multiple rod opsins in deep-sea species (Sec. 4.4), and 3) predictions of colour vision for Martian-adapted creatures (Sec. 4.5). Complete experimental details, including datasets (Sec. A.1.3.1, Sec. A.1.3.3) and noise modeling (Sec. A.1.3.4), are provided in the appendix.

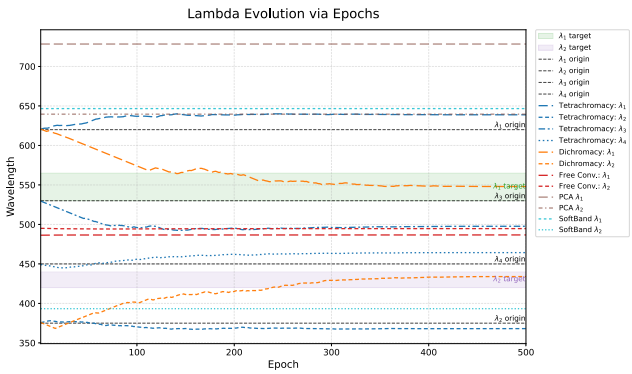


Figure 3: Simulated evolution trajectories of wavelength peak shifts for Tetrachromacy, Dichromacy, and other methods.

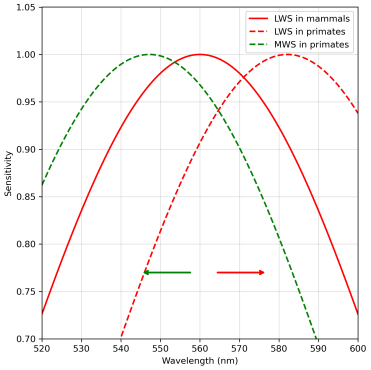


Figure 4: Emulation of the gene duplication process in primates.

4.1 Evolutionary Transition from Tetrachromacy to Dichromacy

During mammalian evolution, ancestral species experienced a reduction from four to two visual opsin genes [2], consistent with the nocturnal bottleneck hypothesis [20]. Our experiments model this transition using the four ancestral teleost opsins: LWS (620 nm, λ_1 origin), SWS1 (375 nm, λ_2 origin), MWS (530 nm, λ_3 origin) and SWS2 (450 nm, λ_4 origin) [6]. This framework specifically examines the loss of MWS and SWS2 opsins in the marsupial and eutherian lineages that preceded primates [104]. Using the LIB-HSI dataset [30] under dim-light simulation, we modelled dichromatic vision with LWS and SWS1 opsins. As shown in Fig.3 and Tab. 1, our model produced spectral sensitivities consistent with evolutionary predictions and the physiology of extant nocturnal mammals. Through conservation optimization, our in silico evolution achieved a 70nm blue-shift (decrease in wavelength) of LWS and a 60nm red-shift (increase in wavelength) of SWS1 within only 500 epochs, reaching the target wavelength ranges (the $\lambda 1$ and $\lambda 2$ target intervals in Fig.3) of the two opsins in eutherians. We further compared Tetrachromacy versus Dichromacy. Under both fixed (Simulation of the time point of gene loss from Tetrachromacy to Dichromacy) and learnable (Simulation of the respective evolution of Tetrachromacy and Dichromacy) opsin layer conditions, dichromatic vision consistently outperformed tetrachromatic vision in segmentation tasks. Finally, we compared our model with several pure machine learning and statistical methods (see Sec.A.1.4). Although those methods achieved performance comparable to our simulated dichromatic evolution, their corresponding peak spectral responses deviated markedly from biological observations, thereby highlighting the biological plausibility of the opsin layer model and the conservation optimization.

4.2 Evolutionary Transition from Dichromacy to Trichromacy

The evolution of trichromacy in early primates (65 MYA) via LWS/MWS opsin gene duplication [20, 40] enabled spectral discrimination critical for fruit/foliage identification [19] and camouflage detection [46]. This adaptation provided selective advantages for diurnal foraging, establishing the neural basis for red-green colour vision in primates. Accordingly, we performed segmentation tasks using the MinneApple dataset [35] (Fig. 1). Following Lamdouar et al. [48], we quantify fruit recognition using a camouflage score (Sec.A.1.2). This metric evaluates how effectively foliage obscures red fruits in segmentation tasks. Lower values indicate better discriminability between leaves and fruits. We simulate trichromatic vision using an opsin layer with λ_{max} at 580/540/425

Table 1: Tetrachromacy-to-Dichromacy Evolution.

Methods	**LWS** λ_1 (nm)	**SWS1** λ_2 (nm)	**MWS** λ_3 (nm)	**SWS2** λ_4 (nm)	**mIoU** (%)
Ours					
Dichromacy (Fixed)	620.00	375.00	–	–	26.95
Dichromacy (Evolve)	548.35	433.25	–	–	**30.49**
Tetrachromacy (Fixed)	620.00	530.00	450.00	375.00	19.93
Tetrachromacy (Evolve)	638.85	368.05	497.60	464.30	26.33
Baselines					
Free Conv.	486.71	494.72	–	–	29.03
PCA	728.40	639.60	–	–	26.86
SoftBand	646.67	393.33	–	–	27.49

Table 2: Dichromacy-to-Trichromacy Evolution via Gene Duplication.

Methods	λ_1 (nm)	λ_2 (nm)	λ_3 (nm)	**mIoU** (%)	S_R
Ours					
Dichromacy	425.00	560.00	–	46.29	0.334
Trichromacy	425.00	540.00	580.00	48.45	**0.208**
Dichromacy (Evolve)	440.65	547.34	582.10	47.99	0.239
Baselines					
Free Conv.	510.20	530.42	570.22	**49.44**	0.224
PCA	525.10	546.25	681.43	44.79	0.312
SoftBand	400.39	460.75	610.38	48.67	0.250

nm, and dichromatic vision by deactivating one channel. Comparative camouflage scores (Tab. 2) demonstrate trichromacy's superior fruit recognition, quantitatively validating its evolutionary advantage in foliage foraging.

We modeled the dichromatic-to-trichromatic transition via simulated opsin gene duplication [36], generating 540 nm and 580 nm variants from a single ancestral gene [40]. This reconstruction used the LIB-HSI dataset [30] to replicate ancestral primate diurnal conditions. Our model initialises opsin layers with λ_{1max}, λ_{2max}, and λ_{3max} at 560, 560, and 425 (fixed) nm. This constrained optimisation simulates both the initial gene duplication and the subsequent spectral divergence. Fig. 4 demonstrates our model's successful reconstruction of the dichromatic-to-trichromatic transition, providing computational validation of gene duplication's role in colour vision evolution. Moreover, the marked reduction in camouflage scores (Tab. 2) further supports the role of gene duplication and its evolutionary process in colour vision evolution. Additionally, we compared our model with other machine learning methods (see Sec. A.1.4). The results show that although those methods achieve comparable performance, they cannot align with biological facts. This study specifically models the human trichromatic pathway, not to compare tetrachromatic and trichromatic vision systems. The lack of tetrachromacy re-emergence in our simulations may reflect neutral processes like genetic drift [56], rather than trichromatic superiority.

4.3 Retention of Colour Blindness

Table 3: Fruit recognition (S_R) across lighting and colour vision types. Lower S_R indicates better camouflage-breaking ability.

Lighting	**Normal**	**G.-blind**	**R.-blind**	**G.-weak**	**R.-weak**
Bright	**0.208**	0.316	0.333	0.254	0.237
Dark (0.10)	0.309	0.391	0.365	0.293	**0.278**
Dark (0.05)	0.355	0.420	0.397	0.324	**0.324**

The persistence of primates colour blindness remains puzzling. While trichromats have advantages for frugivory in daylight, dichromats may gain compensatory benefits in dark. The leading dim-light advantage hypothesis holds that colour-deficient animals recognise fruit more accurately in near darkness [19] and and show a 0.44 log-unit lower light perception threshold after dark adaptation [95]. However, some researchers found no clear differences from colour-deficient subjects in visual tests [86]. Using the MinneApple dataset [35] and camouflage score, we compared fruit recognition between normal trichromatic vision and simulated colour blindness (via opsin channel modulation). Under dim-light conditions (simulated via noise layer in Sec. A.1.3.4]), colour-weak vision outperformed normal trichromacy in fruit detection (Tab. 4.3). Our experimental findings provide substantive support for the leading dim-light advantage hypothesis.

4.4 Deep Sea Creatures

It is worth mentioning that in this section we simulate underwater data by pre-processing the LIB-HSI dataset [30][50]. Specifically, we utilise the spectrum of ambient light E in Eq.3 to reduce the intensity distribution of 204 channels based on different wavelengths. Therefore, we simulate the underwater spectrum and intensity of ambient at different depths.

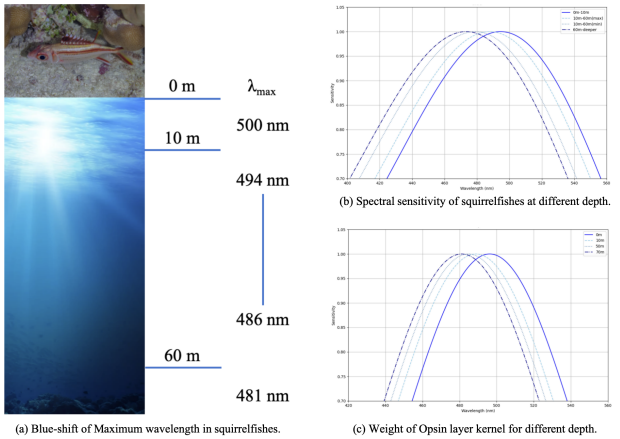
$$E(d, \lambda) = E(0, \lambda)e^{-K_d(\lambda)d} \tag{3}$$

Here, d represents the depth, λ is the wavelength, $E(d, \lambda)$ is the ambient light spectrum at depth d, and $E(0, \lambda)$ represents the spectrum at the sea surface. $K_d(\lambda)$ is diffuse downwelling attenuation [1]. Additionally, we apply the noise layer (Sec.A.1.3.4) for the dim-light conditions in the deep sea.

4.4.1 Blue-shift in Rod Opsin of Squirrelfishes

Fig. 5.a–b demonstrates the depth-dependent spectral tuning of squirrelfish rod opsins [105]. The nine studied species exhibit a progressive blue-shift in λ_{max} (481–502 nm) relative to their ancestral opsin (λ_{max} = 493 nm) as depth increases.

We focus on rod opsin (RH1) in deep-sea adaptation studies for two key reasons. First, deep-sea monochromatic light condition falls below cone detection thresholds, making rod opsin the sole functional photoreceptor [64]. Second, unlike most teleosts with multiple opsin genes, squirrelfish possess only RH1 [64, 105], providing a genetically simple system for investigating spectral tuning under extreme selection pressures. We simulated deep-sea spectral adaptation using the LIB-HSI dataset [30], processed via Eq.3 for depths (0-70m). A single-opsin layer (initial λ_{max} = 493nm, ancestral squirrelfish RH1 [105]) performed semantic segmentation, modeling rod opsin functionality across photic zones. Our quantitative analysis (Tab. 4) reveals a clear depth-dependent reduction in optimal λ_{max} for segmentation mIoU. Fig. 5.c reconstructs this rod opsin blue-shift in deep-sea conditions, demonstrating that light diffusion drives the spectral tuning observed in squirrelfishes.



(a) Blue-shift of Maximum wavelength in squirrelfishes.

(b) Spectral sensitivity of squirrelfishes at different depth.

(c) Weight of Opsin layer kernel for different depth.

Figure 5: Depth-dependent blue-shift in squirrelfish rod opsins.

4.4.2 Multi Rod Opsins: Result of Bioluminescence

While many deep-sea species retain a single rod opsin, certain teleost lineages have independently evolved multiple rod opsins [63]. This adaptation likely reflects specialized bioluminescent conditions in their ecological niches [63]. Our bioluminescent deep-sea model (Sec.A.2.1) shows that: without bioluminescence, rod opsins exhibit minimal spectral separation; with bioluminescence, opsins diverge by a maximum 5 nm across 5 wavelengths. However, this separation remains smaller than natural variation, suggesting additional complexity in multi-opsin development. Future studies with more precise anatomical models could clarify these mechanisms.

Table 4: Blue-shift of $\lambda_{\max}$ vs. depth.

Depth	$\lambda_{\max}$ (nm)	**mIoU** (%)
0 m	496.19	30.38
10 m	488.73	25.88
50 m	484.69	16.97
70 m	481.10	**14.98**

4.5 Colour Vision for Hypothetical Martian Life

We train the opsin layer with different number of kernels to explore the colour vision of hypothetical creatures on Mars, which is close to deuteranopia. We place the experimental results in Sec.A.2.2.

5 Camera Design

Building on our Visual Evolutionary Analysis and Prediction (Sec. 4), we demonstrate how evolutionary principles can guide task-specific camera design. Using standard Sony camera filters (600/540/460 nm) as initial values and leveraging the mathematical equivalence between color imaging and convolutional operations [67], we optimized spectral response functions (Fig. 2) for cancer detection and Martian exploration across three datasets (Sec. A.1.3.2).

Camera for Cancer Detection Building on Varghese et al.'s [94] demonstration of AI's potential to transform surgical care, we adapt our evolutionary optimisation framework to develop specialised cancer-detection cameras using the MHSI Choledoch [108] and MODID datasets [13]. This approach extends AI's medical applications from surgical enhancement to diagnostic imaging optimisation. Our optimised spectral filters (Tab. 5) enhance performance across distinct cancer detection tasks, enabling specialized medical imaging for abnormal tissue identification. Sec. A.3.2 presents scalable designs with varying filter counts for different clinical applications.

Camera for Martian Environment Using the Mars-Seg dataset [51], we optimized spectral filter responses via opsin-layer training. In segmentation tasks, our optimized filter designs (Tab. 6) exhibit superior performance, with extended multi-filter configurations detailed in Sec. A.3.1. These findings suggest that such optimized vision systems could substantially enhance the perceptual capabilities of Mars rovers and exploration vehicles.

In Tab. 5 and 6, we first benchmark our approach against established baselines for Hyperspectral Image processing. A longitudinal review of baseline methods from 2015 to 2025 reveals that annual performance gains have averaged only 1%; in contrast, our approach first achieves state-of-the-art performance based on the MiT-Transformer and further yields an improvement of over 2% mIoU after learning through the opsin layer. We observe that while conventional methods—which focus primarily on iterative improvements to high-level perceptual algorithms—show diminishing returns and slow performance gains, our perspective of low-level camera design and simple hardware modifications yields a substantial performance leap. This aligns with previous findings [33, 93] and offers a new perspective on task-oriented camera design. Furthermore, we evaluate our framework across three configurations: General (non-optimized filters), Free (unconstrained filter optimization), and our proposed Design (progressive, conservative optimization via the Opsin Layer). The results demonstrate that the Design method not only achieves peak performance but also maintains spectral parameters (λ) that remain close to original hardware specifications. This "conservative optimization" is of significant practical value, as it ensures that the resulting designs are biologically grounded and feasible for iterative manufacturing and real-world camera deployment.

Table 5: Camera design for bile duct and oral cancer detection (MHSI and MODID datasets).

Dataset	Methods	Param (M)	λ_1 (nm)	λ_2 (nm)	λ_3 (nm)	mIoU (%)
Baselines (MHSI / MODID)						
MHSI/MODID	UNet (2015)[72]	31.07	–	–	–	51.04 / 54.02
	SwinUNet (2022)[8]	41.48	–	–	–	47.84 / 51.12
	ResUNet (2020)[23]	13.11	–	–	–	49.52 / 52.76
	Segformer (2021)[101]	64.47	–	–	–	50.27 / 52.98
	EMCAD (2024)[71]	84.53	–	–	–	50.07 / 52.67
	nnUNet (2021)[37]	19.68	–	–	–	49.99 / 52.57
	MADGNet (2024)[66]	31.39	–	–	–	51.68 / 54.45
	H-VMUNet (2025)[99]	117.49	–	–	–	51.30 / 52.71
	UDMamba (2024)[111]	19.21	–	–	–	50.89 / 53.81
	FDNet (2024)[110]	96.26	–	–	–	53.29 / 56.21
	Spectr (2024)[106]	3.94	–	–	–	52.19 / 55.03
	UFPF (2025)[70]	201.35	–	–	–	54.99 / 56.98
Ours						
MHSI	General	15.10	600.00	540.00	460.00	54.35
	Free	15.10	689.25	716.45	691.44	54.73
	Design	15.10	595.82	716.45	476.60	**56.36**
MODID	General	15.10	600.00	540.00	460.00	56.42
	Free	15.10	542.58	559.25	558.86	57.69
	Design	15.10	608.73	552.93	449.54	**58.67**

Table 6: Camera design for Martian environment (Mars-Seg).

Methods	Input	Param (M)	λ_1 (nm)	λ_2 (nm)	λ_3 (nm)	mIoU (%)
RGB Baselines						
UNet (2015)[72]	RGB	31.07	–	–	–	50.65
SwinUNet (2022)[8]	RGB	41.48	–	–	–	64.74
Segformer (2021)[101]	RGB	64.47	–	–	–	62.36
HSI Baselines						
UNet (2015)[72]	HSI	31.07	–	–	–	35.03
SwinUNet (2022)[8]	HSI	41.48	–	–	–	36.44
ResUNet (2020)[23]	HSI	13.11	–	–	–	35.51
Segformer (2021)[101]	HSI	64.47	–	–	–	36.93
UDMamba (2024)[111]	HSI	19.21	–	–	–	38.55
Spectr (2024)[106]	HSI	3.94	–	–	–	39.17
UFPF (2025)[70]	HSI	201.35	–	–	–	40.14
Ours						
General	HSI	15.10	600.00	540.00	460.00	39.22
Free	HSI	15.10	724.03	712.25	423.43	40.18
Design	HSI	15.10	623.35	580.41	489.13	**41.23**

6 Summary

In this paper, we present a paleoinspired vision framework that models retinal visual transduction and simulates evolutionary transitions of visual opsins. Our approach enables the quantitative analysis of opsin evolution with high computational efficiency, addressing a core challenge in evolutionary biology and neuroscience. While methods like forward genetics experiments often suffer from resource-intensive stochastic searches for difficult-to-quantify traits [31], our framework provides algorithmic tools to streamline biological research. By implementing processes like evolutionary pressure analysis, we can quantify the direction and magnitude of trait changes in successive generations and simulate the specific environmental conditions of extinct species within an optimization-based paradigm. Beyond theoretical insights, our framework establishes practical paradigms for high-performance camera design. The evolutionary conservatism embedded in our optimization supports manufacturing-friendly filter designs [88], while our minimalist approach, maximizing task-specific adaptability, aligns with emerging trends in computational imaging [44].

Acknowledgments and Disclosure of Funding

This work was partly supported by JSPS KAKENHI Grant Number JP26K21296 (to Z. C.), 23K23924 and 24K21963 (to D.K.), Cooperative Research Project of the Research Institute of Electrical Communication, Tohoku University (to R. F.), and the National Key R&D Program of China Grant Number 2025YFF0811700 (to J. L.).

References

[1] D. Akkaynak and T. Treibitz. Sea-thru: A method for removing water from underwater images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1682–1691, 2019.

[2] R. Albalat and C. Cañestro. Evolution by gene loss. *Nature Reviews.Genetics*, 17(7):379–391, 07 2016. Copyright - Copyright Nature Publishing Group Jul 2016; Last updated - 2023-11-30.

[3] L. Bian, Z. Wang, Y. Zhang, L. Li, Y. Zhang, C. Yang, W. Fang, J. Zhao, C. Zhu, Q. Meng, X. Peng, and J. Zhang. A broadband hyperspectral image sensor with high spatio-temporal resolution. *Nature*, 635(8037):73–81, Nov 2024.

[4] H. R. Blackwell. Contrast thresholds of the human eye. *J. Opt. Soc. Am.*, 36(11):624–643, Nov 1946.

[5] M. L. Borowiec, R. B. Dikow, P. B. Frandsen, A. McKeeken, G. Valentini, and A. E. White. Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 13(8):1640–1660, 2022.

[6] J. K. Bowmaker. Evolution of colour vision in vertebrates. *Eye*, 12(3):541–547, May 1998.

[7] Y. Cai, J. Lin, X. Hu, H. Wang, X. Yuan, Y. Zhang, R. Timofte, and L. Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17481–17490, 2022.

[8] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*, 2022.

[9] L. S. Carvalho, D. M. Pessoa, J. K. Mountford, W. I. Davies, and D. M. Hunt. The genetic and evolutionary drives behind primate color vision. *Frontiers in ecology and Evolution*, 5:34, 2017.

[10] A. Chakrabarti, W. T. Freeman, and T. Zickler. Rethinking color cameras. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, 2014.

[11] A. Chakrabarti, D. Scharstein, and T. Zickler. An empirical camera model for internet color vision. 01 2009.

[12] P. Chakravarthula, J. Sun, X. Li, C. Lei, G. Chou, M. Bijelic, J. Froesch, A. Majumdar, and F. Heide. Thin on-sensor nanophotonic array cameras. *ACM Transactions on Graphics (TOG)*, 42(6):1–18, 2023.

[13] S. Chand, K. Namasivayam, J. Dave, S. P. Preejith, S. Jayachandran, and M. Sivaprakasam. In-vivo non-contact multispectral oral disease image dataset with segmentation. *Scientific Data*, 11, 2024.

[14] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao. Vision transformer adapter for dense predictions, 2023.

[15] D. Coldewey. Neuralink's 'breakthrough device' clearance from fda does not mean it has cured blindness. *Tech Crunch*, 2024.

[16] L. D. Costa, T. Parr, B. Sengupta, and K. J. Friston. Natural selection finds natural gradient. *arXiv: Neurons and Cognition*, 2020.

[17] G. Côté, F. Mannan, S. Thibault, J.-F. Lalonde, and F. Heide. The differentiable lens: Compound lens search over glass surfaces and materials for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[18] X. Cui, W. Zhang, Z. Tüske, and M. Picheny. Evolutionary stochastic gradient descent for optimization of deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[19] M. V. D. Osorio. Colour vision as an adaptation to frugivory in primates. *Royal Society*, 263:593–599, May 1996.

[20] W. I. L. DAVIES, S. P. COLLIN, and D. M. HUNT. Molecular ecology and adaptation of visual photopigments in craniates. *Molecular Ecology*, 21(13):3121–3158, 2012.

[21] H. Davson and Perkins. Human eye, 2024.

[22] S. Deeb. The molecular basis of variation in human color vision. *Clinical Genetics*, 67(5):369–377, 2005.

[23] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94 – 114, 2020.

[24] K. Donner. Noise and the absolute thresholds of cone and rod vision. *Vision Research*, 32(5):853–866, 1992.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[26] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers, 2021.

[27] J. Fong, H. K. Doyle, C. Wang, A. E. Boehm, S. R. Herbeck, V. P. Pandiyan, B. P. Schmidt, P. Tiruveedhula, J. E. Vanston, W. S. Tuten, R. Sabesan, A. Roorda, and R. Ng. Novel color via stimulation of individual photoreceptors at population scale. *Science Advances*, 11(16):eadu1052, 2025.

[28] V. I. GOVARDOVSKII, N. FYHRQUIST, T. REUTER, D. G. KUZMIN, and K. DONNER. In search of the visual pigment template. *Visual Neuroscience*, 17(4):509–528, 2000.

[29] N. Gravish and G. V. Lauder. Robotics-inspired biology. *Journal of Experimental Biology*, 221(7):jeb138438, 2018.

[30] N. Habili, E. Kwan, W. Li, C. Webers, J. Oorloff, M. A. Armin, and L. Petersson. A hyperspectral and rgb dataset for building façade segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 258–267. Springer, 2023.

[31] M. B. Hawkins, K. Henke, and M. P. Harris. Latent developmental potential to form limb-like skeletal structures in zebrafish. *Cell*, 184(4):899–911.e13, Feb 2021.

[32] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pająk, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian, J. Kautz, and K. Pulli. Flexisp: a flexible camera image processing framework. *ACM Trans. Graph.*, 33(6), Nov. 2014.

[34] Z. Huang, W. Shi, S. Wu, Y. Wang, S. Yang, and H. Chen. Pre-sensor computing with compact multilayer optical neural network. *Science Advances*, 10(30):eado8516, 2024.

[35] N. Häni, P. Roy, and V. Isler. Minneapple: A benchmark dataset for apple detection and segmentation, 2019.

[36] H. Innan and F. Kondrashov. The evolution of gene duplications: Classifying and distinguishing between models. *Nature reviews. Genetics*, 11:97–108, 02 2010.

[37] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 02 2021.

[38] M. Ishida, F. Berio, V. D. Santo, N. H. Shubin, and F. Iida. Paleoinspired robotics as an experimental approach to the history of life. *Science Robotics*, 9(95):eadn1125, 2024.

[39] J. Jiang, D. Liu, J. Gu, and S. Süsstrunk. What is the space of spectral sensitivity functions for digital color cameras? In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 168–179, 2013.

[40] J. D. M. Kanwaljit S. Dulai, Miranda von Dornum and D. M. Hunt. The evolution of trichromatic color vision by opsin gene duplication in new world and old world primates. *Genome Res*, 9(7):629–638, 1999.

[41] H. C. Karaimer and M. S. Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision (ECCV)*, 2016.

[42] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[43] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[44] J. Klotz and S. K. Nayar. Minimalist vision with freeform pixels. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, editors, *Computer Vision – ECCV 2024*, pages 329–346, Cham, 2025. Springer Nature Switzerland.

[45] R. Kuo and F. E. Zulvia. The gradient evolution algorithm: A new metaheuristic. *Information Sciences*, 316:246–265, 2015. Nature-Inspired Algorithms for Large Scale Global Optimization.

[46] X. Lai, Z. Yang, J. Hu, S. Zhang, L. Cao, G. Jiang, Z. Wang, S. Zhang, and R. Ji. Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection, 2024.

[47] T. D. Lamb. Sources of noise in photoreceptor transduction. *J. Opt. Soc. Am. A*, 4(12):2295–2300, Dec 1987.

[48] H. Lamdouar, W. Xie, and A. Zisserman. The making and breaking of camouflage, 2023.

[49] D. Li, J. Wu, J. Zhao, H. Xu, and L. Bian. Spectratrack: megapixel, hundred-fps, and thousand-channel hyperspectral imaging. *Nature Communications*, 15(1):9459, Nov 2024.

[50] H. Li, W. Song, T. Xu, A. Elsig, and J. Kulhanek. Watersplatting: Fast underwater 3d scene reconstruction using gaussian splatting, 2024.

[51] J. Li, S. Zi, R. Song, Y. Li, Y. Hu, and Q. Du. A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[52] Z. Li, Q. Hou, Z. Wang, F. Tan, J. Liu, and W. Zhang. End-to-end learned single lens design using fast differentiable ray tracing. *Opt. Lett.*, 46(21):5453–5456, Nov 2021.

[53] E. R. Loew. A third, ultraviolet-sensitive, visual pigment in the tokay gecko (gekko gekko). *Vision Research*, 34(11):1427–1431, 1994. The Biology of Ultraviolet Reception.

[54] Z.-X. Luo, C.-X. Yuan, Q.-J. Meng, and Q. Ji. A jurassic eutherian mammal and divergence of marsupials and placentals. *Nature*, 476(7361):442–445, Aug 2011.

[55] N. Lupše, F. Cortesi, M. Freese, L. Marohn, J.-D. Pohlmann, K. Wysujack, R. Hanel, and Z. Musilova. Visual Gene Expression Reveals a cone-to-rod Developmental Progression in Deep-Sea Fishes. *Molecular Biology and Evolution*, 38(12):5664–5677, 09 2021.

[56] M. Lynch, M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster. Genetic drift, selection and the evolution of the mutation rate. *Nature reviews. Genetics*, 17(11):704—714, October 2016.

[57] T. Matthews, D. Osorio, A. Cavallaro, and L. Chittka. The importance of spatial visual scene parameters in predicting optimal cone sensitivities in routinely trichromatic frugivorous old-world primates. *Frontiers in Computational Neuroscience*, 12:15, 2018.

[58] C. Mauer and D. Wueller. Measuring the spectral response with a set of interference filters. In *Digital photography V*, volume 7250, pages 235–244. SPIE, 2009.

[59] A. D. Melin, K. L. Chiou, E. R. Walco, M. L. Bergstrom, S. Kawamura, and L. M. Fedigan. Trichromacy increases fruit intake rates of wild capuchins (Cebus capucinus imitator). *Proceedings of the National Academy of Sciences*, 114(39):10402–10407, 2017.

[60] A. D. Melin, O. Nevo, M. Shirasu, R. E. Williamson, E. C. Garrett, M. Endo, K. Sakurai, Y. Matsushita, K. Touhara, and S. Kawamura. Fruit scent and observer colour vision shape food-selection strategies in wild capuchin monkeys. *Nature Communications*, 10:2407, 2019.

[61] M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger. All-optical nonlinear activation function for photonic neural networks. *Optical Materials Express*, 8(12):3851–3863, 2018.

[62] Y. Morgenstern, M. Rostami, and D. Purves. Properties of artificial networks evolved to contend with natural spectra. *Proceedings of the National Academy of Sciences*, 111(Supplement 3):10868–10872, 2014.

[63] Z. Musilova, F. Cortesi, M. Matschiner, W. I. L. Davies, J. S. Patel, S. M. Stieb, F. de Busserolles, M. Malmstrøm, O. K. Tørresen, C. J. Brown, J. K. Mountford, R. Hanel, D. L. Stenkamp, K. S. Jakobsen, K. L. Carleton, S. Jentoft, J. Marshall, and W. Salzburger. Vision using multiple distinct rod opsins in deep-sea fishes. *Science*, 364(6440):588–592, 2019.

[64] Z. Musilova, F. Cortesi, M. Matschiner, W. I. L. Davies, J. S. Patel, S. M. Stieb, F. de Busserolles, M. Malmstrøm, O. K. Tørresen, C. J. Brown, J. K. Mountford, R. Hanel, D. L. Stenkamp, K. S. Jakobsen, K. L. Carleton, S. Jentoft, J. Marshall, and W. Salzburger. Vision using multiple distinct rod opsins in deep-sea fishes. *Science*, 364(6440):588–592, 2019.

[65] E. Musk. An integrated brain-machine interface platform with thousands of channels. *J Med Internet Res*, 21(10):e16194, Oct 2019.

[66] J.-H. Nam, N. S. Syazwany, S. J. Kim, and S.-C. Lee. Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11480–11491, 2024.

[67] S. Nie, L. Gu, Y. Zheng, A. Lam, N. Ono, and I. Sato. Deeply learned filter response functions for hyperspectral reconstruction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4767–4776, 2018.

[68] K. Nishino, A. Subpa-asa, Y. Asano, M. Shimano, and I. Sato. Variable ring light imaging: Capturing transient subsurface scattering with an ordinary camera. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, page 624–639, Berlin, Heidelberg, 2018. Springer-Verlag.

[69] A. Pasupathy. The neural basis of image segmentation in the primate brain. *Neuroscience*, 296:101–109, 2015.

[70] G. Qin, H. Liu, W. Li, X. Zhang, Y. Guo, and X.-G. Xia. Ufpf: A universal feature perception framework for microscopic hyperspectral images. *IEEE Transactions on Image Processing*, 34:5513–5526, 2025.

[71] M. M. Rahman, M. Munir, and R. Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11769–11779, 2024.

[72] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[73] T. Salimans, J. Ho, X. Chen, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. 03 2017.

[74] R. Salomon. Evolutionary algorithms and gradient search: similarities and differences. *IEEE Transactions on Evolutionary Computation*, 2(2):45–55, 1998.

[75] D. SCHNEEWEIS and J. SCHNAPF. Noise and light adaptation in rods of the macaque monkey. *Visual Neuroscience*, 17(5):659–666, 2000.

[76] R. K. Schott, R. C. Bell, E. R. Loew, K. N. Thomas, D. J. Gower, J. W. Streicher, and M. K. Fujita. Transcriptomic evidence for visual adaptation during the aquatic to terrestrial metamorphosis in leopard frogs. *BMC Biology*, 20(1):138, Jun 2022.

[77] R. K. Schott, M. K. Fujita, J. W. Streicher, D. J. Gower, K. N. Thomas, E. R. Loew, A. G. Bamba Kaya, G. B. Bittencourt-Silva, C. Guillherme Becker, D. Cisneros-Heredia, S. Clulow, M. Davila, J. Firneno, Thomas J, C. F. B. Haddad, S. Janssenswillen, J. Labisko, S. T. Maddock, M. Mahony, R. A. Martins, C. J. Michaels, N. J. Mitchell, D. M. Portik, I. Prates, K. Roelants, C. Roelke, E. Tobi, M. Woolfolk, and R. C. Bell. Diversity and Evolution of Frog Visual Opsins: Spectral Tuning and Adaptation to Distinct Light Environments. *Molecular Biology and Evolution*, 41(4):msae049, 04 2024.

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

[79] M. J. Shafiee, A. Mishra, and A. Wong. Deep learning with darwin: Evolutionary synthesis of deep neural networks. *Neural Processing Letters*, 48(1):603–613, Aug 2018.

[80] A. Shen. Gradient descent emerges via natural selection. In *2025 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2025.

[81] Q. Shen, Y. Li, J. Jin, and B. Liu. Q-net: Query-informed few-shot medical image segmentation, 2023.

[82] W. Shi, Z. Huang, H. Huang, C. Hu, M. Chen, S. Yang, and H. Chen. Loen: Lensless opto-electronic neural network empowered machine vision. *Light: Science & Applications*, 11(1):121, 2022.

[83] Z. Shi, I. Chugunov, M. Bijelic, G. Côté, J. Yeom, Q. Fu, H. Amata, W. Heidrich, and F. Heide. Split-aperture 2-in-1 computational cameras. *ACM Trans. Graph.*, 43(4), jul 2024.

[84] Z. Shi, I. Chugunov, M. Bijelic, G. Côté, J. Yeom, Q. Fu, H. Amata, W. Heidrich, and F. Heide. Split-aperture 2-in-1 computational cameras. *ACM Trans. Graph.*, 43(4), July 2024.

[85] T. Simonite. Colour blindness may have hidden advantages. *Nature*, Dec 2005.

[86] M. Simunovic, B. Regan, and J. Mollon. Is color vision deficiency an advantage under scotopic conditions? *Investigative ophthalmology & visual science*, 42:3357–64, 01 2002.

[87] Q. Sun, C. Wang, F. Qiang, D. Xiong, and H. Wolfgang. End-to-end complex lens design with differentiable ray tracing. *ACM Transactions on Graphics (TOG)*, 40(4), 2021.

[88] S. Thomke. Enlightened experimentation: The new imperative for innovation, 2001.

[89] Q. Tian, H. Liao, X. Huang, B. Yang, D. Lei, S. Ourselin, and H. Liu. Endomamba: An efficient foundation model for endoscopic videos, 2025.

[90] K. Tiwary, A. Young, Z. Tasneem, T. Klinghoffer, A. Dave, T. Poggio, D.-E. Nilsson, B. Cheung, and R. Raskar. What if eye...? computationally recreating vision evolution. *Science Advances*, 11(51):eady2888, 2025.

[91] E. Tseng, S. Colburn, J. Whitehead, L. Huang, S.-H. Baek, A. Majumdar, and F. Heide. Neural nano-optics for high-quality thin lens imaging. *Nature Communications*, 12(1):6493, Nov 2021.

[92] E. Tseng, A. Mosleh, F. Mannan, K. St-Arnaud, A. Sharma, Y. Peng, A. Braun, D. Nowrouzezahrai, J.-F. Lalonde, and F. Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Transactions on Graphics (TOG)*, 40(2), 2021.

[93] E. Tseng, A. Mosleh, F. Mannan, K. St-Arnaud, A. Sharma, Y. Peng, A. Braun, D. Nowrouzezahrai, J.-F. Lalonde, and F. Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Trans. Graph.*, 40(2), June 2021.

[94] C. Varghese, E. M. Harrison, G. O'Grady, and E. J. Topol. Artificial intelligence in surgery. *Nature Medicine*, 30(5):1257–1268, May 2024.

[95] S. Verhulst and F. Maes. Scotopic vision in colour-blinds. *Vision Research*, 38(21):3387–3390, 1998.

[96] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. Von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[97] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, and D. Psaltis. Inference in artificial intelligence with deep optics and photonics. *Nature*, 588(7836):39—47, December 2020.

[98] S. Whitelam, V. Selin, S.-W. Park, and I. Tamblyn. Correspondence between neuroevolution and gradient descent. *Nature Communications*, 12(1):6317, Nov 2021.

[99] R. Wu, Y. Liu, P. Liang, and Q. Chang. H-vmunet: High-order vision mamba unet for medical image segmentation. *Neurocomputing*, page 129447, 2025.

[100] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021.

[101] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[102] C. Yang, X.-H. Li, M. Zhu, D. J. Condon, and J. Chen. Geochronological constraint on the cambrian chengjiang biota, south china. *Journal of the Geological Society*, 175(4):659–666, 2018.

[103] Z. Yang, T. Wang, Y. Lin, Y. Chen, H. Zeng, J. Pei, J. Wang, X. Liu, Y. Zhou, J. Zhang, X. Wang, X. Lv, R. Zhao, and L. Shi. A vision chip with complementary pathways for open-world sensing. *Nature*, 629:1027–1033, 05 2024.

[104] S. Yokoyama. Molecular evolution of vertebrate visual pigments. *Progress in Retinal and Eye Research*, 19(4):385–419, 2000.

[105] S. Yokoyama and N. Takenaka. The Molecular Basis of Adaptive Evolution of Squirrelfish Rhodopsins. *Molecular Biology and Evolution*, 21(11):2071–2078, 07 2004.

[106] B. Yun, B. Lei, J. Chen, H. Wang, S. Qiu, W. Shen, Q. Li, and Y. Wang. Spectr: Spectral transformer for microscopic hyperspectral pathology image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4610–4624, 2024.

[107] B. Yun, Q. Li, L. Mitrofanova, C. Zhou, and Y. Wang. Factor space and spectrum for medical hyperspectral image segmentation. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 152–162, Cham, 2023. Springer Nature Switzerland.

[108] Q. Zhang, Q. Li, G. Yu, L. Sun, M. Zhou, and J. Chu. A multidimensional choledoch database and benchmarks for cholangiocarcinoma diagnosis. *IEEE Access*, 7:149414–149421, 2019.

[109] X. Zhang, J. Clune, and K. O. Stanley. On the relationship between the openai evolution strategy and stochastic gradient descent. *ArXiv*, abs/1712.06564, 2017.

[110] X. Zhang, Q. Li, W. Li, Y. Guo, J. Zhang, C. Guo, K. Chang, and N. H. Lovell. Fd-net: Feature distillation network for oral squamous cell carcinoma lymph node segmentation in hyperspectral imagery. *IEEE Journal of Biomedical and Health Informatics*, 28(3):1552–1563, 2024.

[111] W. zhao, F. Wang, Y. Xie, Q. Wu, and Y. Zhou. UD-mamba: A pixel-level uncertainty-driven mamba model for medical image segmentation, 2024.

A Technical Appendices and Supplementary Material

A.1 Implementation Details

A.1.1 Layers after Opsin Layer

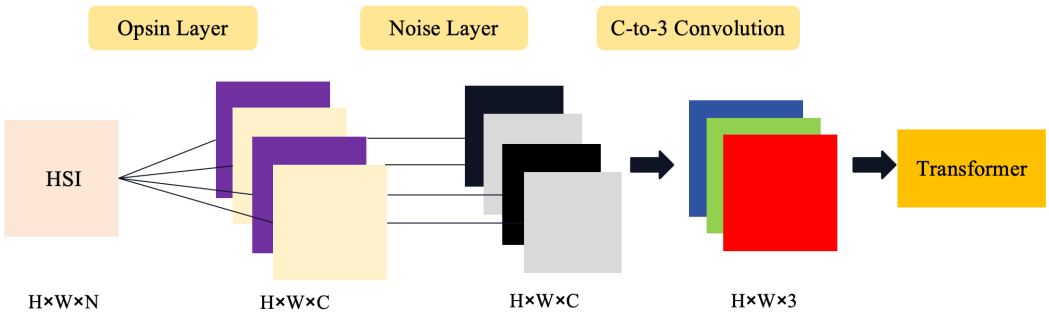


Figure 6: Architecture of the network.

Noise Layer The noise layer is applied after each channel of the opsin layer's output to incorporate the noise present in opsins. Let the output of the opsin layer for channel c be I_c; the output of the noise layer is expressed as:

$$I_c^{noise} = I_c + Gaussian(\mu = 0,\ \sigma^2 = \tau \times I_c) \tag{4}$$

where $Gaussian(\mu, \sigma^2)$ represents Gaussian noise with a mean μ and variance σ^2, and τ is a scaling factor that determines the noise intensity as illustrated in Sec.A.1.3.4.

C-to-3 Convolution Layer. The MiT encoder used in our work requires input data of size $H \times W \times 3$, corresponding to a 3-channel image. However, the output of our opsin layer is of size $H \times W \times C$, where C represents the number of convolutional kernels in the opsin layer, corresponding to the simulated number of opsins. To ensure compatibility with the MiT encoder, we introduce a 1×1 convolutional layer to transform the feature map from $H \times W \times C$ to $H \times W \times 3$. As shown in Fig.6, this 1×1 convolution layer is applied after the opsin layer and the noise layer.

Training Stage. In the training stage, we only train the opsin layer and C-to-3 convolution layer (if exist) simultaneously with the encoder and decoder, while fixing the noise layer (if exist).

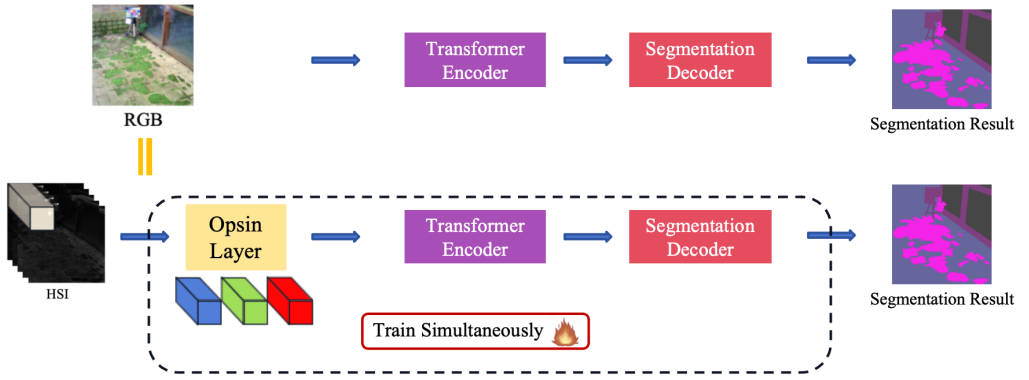


Figure 7: Training stage with Opsin Layer, Transformer Encoder and Segmentation Decoder.

A.1.2 Camouflage Score

Specifically, we utilize the reconstruction fidelity score proposed by Hala Lamdouar *et al.* [48] as the metric to evaluation the ability of the colour vision to recognize camouflage. For the LMS image F output from the opsin layer and segmentation mask m, we first get the foreground region and the

background region by erosion and dilation operations on the mask,

$$F_{fg}, F_{bg} = \text{erode}(m) \odot F, (1 - \text{dilate}(m)) \odot F \tag{5}$$

We calculate the reconstruction fidelity score, which is, briefly, the proportion of pixels in the foreground that can be successfully reconstructed from the background [48],

$$S_R = \frac{1}{N_{fg}} \sum_{(i,j) \in F_{fg}} R_f(i,j) \tag{6}$$

$$R_f(i,j) = \left[||F_{fg} - \Phi(F_{fg})||_2 < t||F_{fg}||_2 \right] \tag{7}$$

where N_{fg} is the number of pixels in F_{fg}, Φ is the reconstruction operation and t is a threshold parameter. A lower S_R value indicates a superior capability of the model to break camouflage, suggesting that the simulated animal vision is better adapted to survival pressures. In Sec.4.3, we adjust the value of t based on the lighting conditions, as $||F_{fg}||_2$ is small in dim light, making the reconstruction criterion for the foreground region more stringent. Specifically, we set $t = 0.2$ for bright conditions, $t = 1.2$ for dark (0.1), and $t = 1.6$ for darker conditions (0.05).

A.1.3 Data Preparation

A.1.3.1 Datasets Used in Experimental Platform for Evolution

- **LIB-HSI** [30] comprises hyperspectral images and corresponding RGB images of building façades, incorporating associated vegetation, bare soil, and ground surfaces. This composition provides critical spectral-spatial features that effectively simulate key environmental parameters required for our experiments.
- **MinneApple** [35] It comprises various apple tree species, closely matching the fruit-picking conditions in our experiment. We selected red fruits and extracted image regions containing only leaves and fruits for analysis.
- **Mars-Seg** [51] contains downsampled (560 × 500 px, bilinear interpolation) Mars surface images from NASA's Curiosity rover, representing the most comprehensive and authoritative collection of diverse Martian terrains for scientific research.

A.1.3.2 Datasets Used in Camera Design

- **Mars-Seg** See Sec. A.1.3.1.
- **MHSI Choledoch** [108] is a comprehensive collection of microscopy hyperspectral and colour images focused on cholangiocarcinoma, a type of bile duct cancer.
- **MODID** [13] is a in-vivo multispectral oral cancer dataset comprising 243 images with 16 spectral bands from 460–600 nm.

A.1.3.3 HSI Generation Due to the scarcity of high-quality annotated HSI datasets, we generate synthetic HSI data from RGB images using the state-of-the-art method described in [7]. This approach is applied to MinneApple and Mars-Seg datasets, enabling us to model photoreceptor spectral sensitivities across diverse real-world scenarios.

A.1.3.4 Noise of Eyes in Dim-Light Environment We simulate dim-light environments in Sec.4.1, Sec.4.3 and Sec.4.4 by adding a noise layer after opsin layer. We simplify the noise in eye as the Poisson fluctuations of dark light [24][47]. The variance of noise σ_N^2 is shown in Eq.8, where I is the intensity of dark light and τ is the noise factor. We referred the value of the signal-to-noise ratio (SNR) under dim light for single cone cell in human and rod cell in macaque monkeys measured by H.Richard Blackwell [4] and D.M.Schneeweis *et al.* [75] correspondingly, hence set the noise factor $\tau = 0.1$ for cone cells in Sec.4.1, Sec.4.3 and $\tau = 0.02$ for rod cells in Sec.4.4.

$$\sigma_N^2 = \tau \times I \tag{8}$$

A.1.4 Ablation Method Details

Let $\mathbf{X} \in \mathbb{R}^{B \times H \times W}$ denote a hyperspectral image with B spectral bands. The center wavelength of the b-th band is denoted as λ_b. All compared modules convert $\mathbf{X}$ into a three-channel representation $\mathbf{Y} \in \mathbb{R}^{3 \times H \times W}$ before it is fed into the downstream segmentation network (MiT Encoder).

A.1.4.1 Fixed Opsin Layer. The proposed opsin layer represents each photoreceptor-like channel by a Gaussian spectral response centered at a biologically motivated peak wavelength $\lambda_c^{\max}$. For the three-kernel opsin layer, the three peak wavelengths are fixed. The response of the c-th channel to the b-th band is

$$g_{c,b} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\lambda_b - \lambda_c^{\max})^2}{2\sigma^2}\right), \tag{9}$$

where σ nm is fixed. The three-channel output is obtained by a spectral 1×1 convolution:

$$Y_{c,h,w} = \sum_{b=1}^{B} g_{c,b} X_{b,h,w}. \tag{10}$$

This design imposes an explicit smoothness and unimodality prior on the spectral responses, mimicking opsin-like sensitivity curves.

A.1.4.2 Free Spectral Convolution. We further consider a fully data-driven spectral mixing baseline implemented as an unconstrained 1×1 convolution from the B bands to three channels:

$$Y_{c,h,w} = \sum_{b=1}^{B} W_{c,b} X_{b,h,w}, \tag{11}$$

where $W \in \mathbb{R}^{3\times B}$ is learned directly from the segmentation loss. This baseline contains only $3 * B$ trainable spectral parameters and does not assume Gaussianity, smoothness, or biological interpretability. Although a free convolution does not explicitly define $\lambda_{\max}$, we extract an equivalent wavelength from the learned spectral response. Since raw convolution weights may be negative, we first map them to non-negative response values:

$$r_{c,b} = \mathrm{softplus}(W_{c,b}) = \log(1 + \exp(W_{c,b})). \tag{12}$$

The hard peak wavelength is then defined as

$$\lambda_c^{\mathrm{peak}} = \lambda_{\arg\max_b r_{c,b}}. \tag{13}$$

A.1.4.3 PCA Projection. We include PCA as a classical non-learnable band-compression baseline. PCA is fitted only on the training split to avoid information leakage. Specifically, spectral pixels are sampled from the training hyperspectral cubes and reshaped into a matrix

$$\mathbf{S} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N\times B}, \tag{14}$$

where each $\mathbf{x}_n \in \mathbb{R}^B$ is a spectral vector at one spatial location. PCA estimates the leading eigenvectors of the empirical covariance matrix

$$\mathbf{C} = \frac{1}{N-1}(\mathbf{S} - \bar{\mathbf{S}})^\top(\mathbf{S} - \bar{\mathbf{S}}), \tag{15}$$

and keeps the first three principal components:

$$\mathbf{P} = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]^\top \in \mathbb{R}^{3\times B}. \tag{16}$$

Each hyperspectral pixel is then projected to a three-dimensional representation:

$$\mathbf{y}_{h,w} = \mathbf{P}\mathbf{x}_{h,w}. \tag{17}$$

In our implementation, the PCA matrix is estimated from the training set and reused unchanged for the validation/test set, ensuring that all splits share the same projection space. The projected channels are min–max normalized independently. This baseline has no trainable parameters in the segmentation model and evaluates whether a variance-preserving linear projection is sufficient for downstream semantic segmentation. For interpretability, the spectral characteristics of the i-th principal component can be summarized by the magnitude of its loading vector. Its dominant wavelength is

$$\lambda_i^{\mathrm{peak}} = \lambda_{\arg\max_b |P_{i,b}|}, \tag{18}$$

A.1.4.4 Gumbel-Softmax Differentiable Band Selection. Hard band selection is a natural baseline for hyperspectral compression because it preserves real sensor bands rather than synthesizing channels by dense spectral mixing. However, directly selecting bands with an $\arg\max$ operator is non-differentiable and cannot be optimized end-to-end using gradient descent. We therefore use a Gumbel-Softmax relaxation, which provides a differentiable approximation to categorical band selection during training while converging toward discrete one-hot choices as the temperature decreases. For each output channel c, we maintain learnable logits $\boldsymbol{\alpha}_c \in \mathbb{R}^B$. During training, a relaxed band-selection vector is sampled as

$$q_{c,b} = \frac{\exp((\alpha_{c,b} + \gamma_{c,b})/\tau)}{\sum_{b'=1}^{B} \exp((\alpha_{c,b'} + \gamma_{c,b'})/\tau)}, \tag{19}$$

where τ is the temperature and $\gamma_{c,b}$ is Gumbel noise sampled by

$$\gamma_{c,b} = -\log(-\log u_{c,b}), \qquad u_{c,b} \sim \mathcal{U}(0,1). \tag{20}$$

The selected three-channel representation is computed by

$$Y_{c,h,w} = \sum_{b=1}^{B} q_{c,b} X_{b,h,w}. \tag{21}$$

The temperature is annealed after each epoch:

$$\tau_t = \max(\tau_{\min}, \tau_0 \exp(-\eta t)), \tag{22}$$

where $\tau_0 = 1.0$, $\tau_{\min} = 0.1$, and $\eta = 0.003$ in our implementation. At inference time, the relaxed distribution is replaced by a hard one-hot selection:

$$b_c^\star = \arg\max_b \alpha_{c,b}, \qquad q_{c,b} = \mathbb{I}[b = b_c^\star]. \tag{23}$$

The corresponding selected wavelength is therefore

$$\lambda_c^{\max} = \lambda_{b_c^\star}. \tag{24}$$

This baseline is reasonable because it bridges classical discrete band selection and end-to-end neural optimization: during training it remains differentiable, while during inference it yields an interpretable subset of actual MODID wavelengths.

A.1.5 Training Details

We have eight RTX 3090 GPUs to run the experiments. We train the segmentation task on various datasets by minimizing the cross-entropy loss and using the Adam optimizer [42] to update the parameters of the opsin layer, encoder, and decoder.

In training to reconstruct the evolutionary transition to dichromatic vision in mammals (Sec.4.1), we use the Adam optimizer to train the opsin layer with a base learning rate of 2×10^{-2} and the encoder and decoder with a base learning rate of 5×10^{-4}. Additionally, a cosine learning schedule is applied to mitigate over-fitting.

In training to simulate the gene duplication process in primates leading to the acquisition of trichromatic vision (Sec.4.2), we use the Adam optimizer to train the opsin layer with a base learning rate of 5×10^{-2} and the encoder and decoder with a base learning rate of 5×10^{-4}. Additionally, a cosine learning schedule is applied to mitigate over-fitting.

In training to compare the performance of colour blindness and normal trichromatic vision (Sec.4.3), we use the Adam optimizer to train the opsin layer, the encoder and decoder all with a base learning rate of 5×10^{-4}. Additionally, a cosine learning schedule is applied to mitigate over-fitting.

In training to reconstruct the blue-shift phenomenon of rod opsins in squirrelfishes (Sec.4.4.1), we use the Adam optimizer to train the opsin layer with a base learning rate of 5×10^{-3} and the encoder and decoder with a base learning rate of 5×10^{-4}. Additionally, a cosine learning schedule is applied to mitigate over-fitting.

In training to quantitatively analyse the evolutionary development of multiple rod opsins in certain deep-sea species (Sec.4.4.2), we use the Adam optimizer to train the opsin layer with a base learning

rate of 1×10^{-2} and the encoder and decoder with a base learning rate of 5×10^{-4}. Additionally, a cosine learning schedule is applied to mitigate over-fitting.

To quantitatively analyse the potential colour vision of hypothetical creatures adapted to Martian environments (Sec.4.5), we use the Adam optimizer to train the opsin layer with a base learning rate of 5×10^{-2} and the encoder and decoder with a base learning rate of 5×10^{-4}. Additionally, a cosine learning schedule is applied to mitigate over-fitting.

For the camera design section (Sec.5), we use the Adam optimizer to train the opsin layer with a base learning rate of 3×10^{-2} and the encoder and decoder with a base learning rate of 6×10^{-5}. Furthermore, we incorporate an adaptive training strategy, inspired by [81], to dynamically optimize the wavelengths during model training. This approach enables the network to learn task-specific spectral sensitivities, enhancing feature discriminability for oral cancer detection.

The reported results represent mean performance over multiple independent random seeds. The default training duration is 500 epochs; the multi-rod trajectories below are reported through 200 epochs, and the hypothetical Martian-vision trajectories through 1000 epochs. These reported durations should not be interpreted as a common stopping rule, and mean scores alone do not quantify uncertainty.

A.2 Supplementary Results for Evolution

A.2.1 Multi Rod Opsins: Result of Bioluminescence

As shown in Figure 8, we simulate a deep-sea environment with bioluminescence by retaining the labelled region without underwater attenuation. At epoch 200, the five peak wavelengths span 0.33 nm without bioluminescence (Table 9) and 3.98 nm with bioluminescence (Table 10). The separation is qualitatively consistent with the multi-rod hypothesis [63], but its magnitude remains smaller than in the biological observations.

This provides quantitative evidence supporting the role of bioluminescence in the development of multiple rod opsins. However, the separation between rod opsins remains minimal compared to biological experimental data [63], aligning with hypotheses suggesting the involvement of more complex mechanisms in the evolution of multiple rod opsins [63].

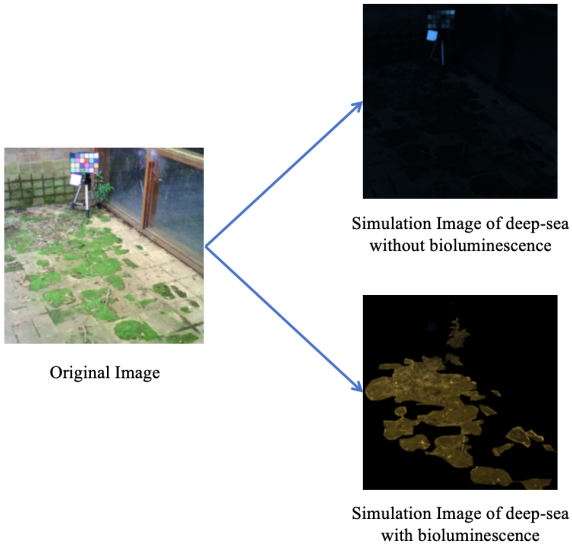


Figure 8: Environment simulation with and without bioluminescence.

A.2.2 Colour Vision for Hypothetical Martian Life

In this section, we present predictions of possible colour vision capabilities for hypothetical Martian life, including dichromatic, trichromatic, and tetrachromatic vision.

Table 7: 1-kernel opsin layer without bioluminescence.

Epoch	λ_{max}	mIoU %
0	481.00	-
50	471.86	6.69
100	466.09	7.88
150	461.62	10.17
200	457.23	11.14

Table 8: 1-kernel opsin layer with bioluminescence.

Epoch	λ_{max}	mIoU %
0	481.00	-
50	482.66	8.26
100	482.13	10.09
150	481.41	12.01
200	480.83	15.17

Table 9: 5-kernel opsin layer without bioluminescence.

Epoch	λ_{1max}	λ_{2max}	λ_{3max}	λ_{4max}	λ_{5max}	mIoU %
0	481.00	481.00	481.00	481.00	481.00	-
50	475.41	475.41	475.26	475.56	475.29	8.44
100	472.89	472.89	472.74	473.04	472.76	12.16
150	470.40	470.38	470.24	470.56	470.26	13.02
200	467.76	467.73	467.59	467.92	467.62	13.77

Firstly, we train a 2-kernel opsin layer to optimize segmentation performance on the Mars-Seg dataset, aiming to predict the optimal dichromatic vision. As shown in Tab.11, the initial wavelengths are set to values similar to human vision, 560 nm and 425 nm.

Secondly, we apply the gene duplication process to the previously trained 2-kernel opsin layer to analyse the evolutionary transition from dichromatic to trichromatic vision adapted to the Martian environment. Specifically, as shown in Tab.12, the initial wavelengths are set to 588.81, 588.81, and 429.37 nm, corresponding to the optimal parameters obtained from the trained 2-kernel layer. It is worth noting that if the low-wavelength-sensitive opsin is allowed to duplicate, no duplication occurs, or more specifically, the two wavelengths remain indistinguishable and do not separate.

Similarly, we allow gene duplication to simulate the hypothetical evolutionary transition from trichromatic to tetrachromatic vision in the Martian environment, as illustrated in Tab.13.

The quantitative results indicate that dichromatic vision achieves the best performance in the Martian environment. This suggests that hypothetical Martian creatures may evolve to have dichromatic vision if the environmental conditions remain consistent.

A.3 Supplementary Results for Camera Design

A.3.1 Camera for Martian Environment

In this section, we present the design of cameras with varying numbers of filters, optimized for operation in the Martian environment. As shown in Tab.14, R, G, and B represent the red, green, and blue colour filters, respectively. The initial parameters for the maximum sensitivity wavelengths of these filters are set to 600 nm, 540 nm, and 460 nm, respectively, aligning with the values commonly used in general cameras [58]. Therefore, we train various configurations of the opsin layer with learning rate 3×10^{-2} to obtain the colour response functions of filters optimized for the best segmentation performance. As shown in Tab.14, we propose specific camera filter designs. For a minimalist Martian camera, the two-filter design with peaks at 624.22 and 459.67 nm offers a

Table 10: 5-kernel opsin layer with bioluminescence.

Epoch	λ_{1max}	λ_{2max}	λ_{3max}	λ_{4max}	λ_{5max}	mIoU %
0	481.00	481.00	481.00	481.00	481.00	-
50	481.78	482.40	480.63	481.46	482.12	10.20
100	481.60	482.53	479.68	481.12	481.76	13.17
150	481.37	482.48	478.96	480.74	481.40	17.67
200	481.73	482.96	478.98	480.98	481.69	21.26

Table 11: Dichromatic Vision on Mars.

Epoch	λ_{1max}	λ_{2max}	mIoU
0	560.00	425.00	-
100	571.49	428.05	18.16
200	575.26	428.68	24.86
300	578.98	429.02	27.54
400	580.68	429.21	31.36
500	582.98	429.34	34.11
600	582.52	429.24	36.58
700	583.21	429.21	37.48
800	584.79	429.23	39.25
900	586.70	429.34	39.15
1000	588.81	429.37	41.53

Table 12: Trichromatic Vision on Mars.

Epoch	λ_{1max}	λ_{2max}	λ_{3max}	mIoU
0	588.81	588.81	429.37	-
100	596.48	584.76	431.02	16.44
200	603.89	580.89	432.35	23.79
300	609.40	576.38	433.12	26.22
400	615.76	573.13	434.03	29.77
500	621.05	573.67	434.30	32.71
600	625.98	574.24	434.57	35.30
700	630.70	575.09	434.83	36.38
800	637.33	577.62	435.22	37.77
900	642.24	579.63	435.55	39.14
1000	647.87	582.86	435.80	39.88

Table 13: Tetrachromatic Vision on Mars.

Epoch	λ_{1max}	λ_{2max}	λ_{3max}	λ_{4max}	mIoU
0	647.87	647.87	582.86	435.80	-
100	646.60	644.15	577.73	434.19	18.62
200	647.90	643.76	576.71	434.22	26.45
300	648.76	643.28	575.91	434.24	29.72
400	649.90	643.35	575.27	434.32	30.44
500	651.01	643.49	574.80	434.37	34.62
600	651.91	643.63	574.44	434.45	34.78
700	653.35	644.21	573.90	434.63	36.20
800	654.66	644.73	573.73	434.80	35.51
900	655.83	645.11	573.14	434.82	39.33
1000	657.08	645.72	572.63	434.95	41.07

channel-count/performance trade-off: 40.78% mIoU versus 41.60% for four filters, a difference of 0.82 percentage points.

Table 14: Camera Specialised for Martian Environment.

Filters	λ_{1max}	λ_{2max}	λ_{3max}	λ_{4max}	mIoU %
R	613.31	-	-	-	36.26
R/B	624.22	459.67	-	-	40.78
R/G	622.91	517.07	-	-	40.61
R/R/B	617.27	615.00	457.84	-	40.37
R/R/B/B	629.36	623.09	457.93	454.86	41.60

A.3.2 Camera for Cancer Detection

Similar to Sec.A.3.1, we train various configurations of the opsin layer to obtain the colour response functions of filters optimized for the best segmentation performance on MHSI dataset. As shown in Tab.15, we propose specific camera filter designs.

In addition, for the segmentation task, we replace the MiT encoder and segmentation decoder used previously with a U-Net model [72] featuring a ResNet-34 backbone [32] and a U-Net decoder, similar to the baseline model in [107]. The opsin layer is integrated into the U-Net architecture, and the opsin layer, backbone, and U-Net decoder are trained simultaneously. As illustrated in Tab.16, we propose a 3-filter camera design optimized using the U-Net architecture, demonstrating that our evolutionary conservation optimization framework is robust and effective across different model architectures, including both transformer-based models with MiT encoders and convolutional architectures like U-Net.

Due to the limited dataset size in our oral cancer detection task, we employ 5-fold cross-validation to robustly evaluate our method. To ensure a fair comparison between conventional RGB channels

Table 15: Camera Specialised for Cancer Detection.

Filters	λ_{1max}	λ_{2max}	λ_{3max}	mIoU %
R	612.95	-	-	54.22
R/B	620.43	477.91	-	56.97
R/G	623.44	538.26	-	56.64
R/R/B	621.11	619.56	472.21	54.53

Table 16: Train With U-Net Model.

Camera	λ_{1max}	λ_{2max}	λ_{3max}	Iou for cancer %
Standard	600.00	540.00	460.00	52.33
Design	610.39	555.09	456.02	54.97

and our proposed adaptive camera design, we maintain identical experimental parameters (e.g., model architecture, learning rates) across both configurations. Each fold is repeated three times with randomized initializations, and we report the mean performance metrics to mitigate variance and enhance statistical reliability.

A.4 Discussion, Limitation and Future Works

In this paper, we present a paleoinspired vision framework that models retinal visual transduction and simulates evolutionary transitions of visual opsins. Our approach enables the quantitative analysis of opsin evolution with high computational efficiency, addressing a core challenge in traditional evolutionary biology. While methods like forward genetics experiments often suffer from resource-intensive stochastic searches for difficult-to-quantify traits [31], our framework provides algorithmic tools to streamline biological research. By implementing processes like evolutionary pressure analysis, we can quantify the direction and magnitude of trait changes in successive generations and simulate the specific environmental conditions of extinct species within an optimization-based paradigm. Beyond theoretical insights, our framework establishes practical paradigms for high-performance camera design. The evolutionary conservatism embedded in our optimization supports manufacturing-friendly filter designs [88], while our minimalist approach, maximizing task-specific adaptability, aligns with emerging trends in computational imaging [44]. Furthermore, the integration of direct photoreceptor control (Oz principle) suggests potential for engineering spectral sensitivities that bypass biological constraints to detect "impossible" colors [27]. Such innovations could revolutionize surveillance, biomedical imaging, and the development of bio-compatible spectral encoding for brain-computer interfaces (BCIs), especially as the demand for vision-restoring neural chips continues to grow [15, 65].

Despite advancing evolutionary insights, this study faces several constraints. First, the scarcity of functional or behavioral data limits the experimental validation of competing hypotheses, while relying on segmentation as a fitness proxy may underspecify the multifaceted nature of evolutionary pressures. Second, our environmental simulations and phylogenetic sampling remain simplified. Furthermore, the reliance on RGB-to-HSI reconstruction (via MST++), though state-of-the-art, is inherently ill-posed; the resulting spectral artifacts and ambiguities could introduce spurious correlations that confound evolutionary interpretations.

Future work will focus on several directions to address these limitations. On the evolutionary modeling front, we aim to enhance the framework by incorporating more sophisticated environmental and anatomical variables, and by collecting biologically relevant hyperspectral datasets to enable more rigorous paleontological analyses. On the camera design front, we plan to collaborate with materials scientists and optical engineers to prototype and validate the proposed camera designs, with careful consideration of manufacturing feasibility constraints including spectral bandwidth tolerances, fabrication precision, and quantum efficiency trade-offs. These efforts will bridge the gap between computational simulation and practical implementation, ultimately advancing both evolutionary biology research and task-specific imaging system development.

A.5 Task-Based Evolutionary Modelling: Rationale and Evidence

Our framework is intended as a computational hypothesis-testing model of visual adaptation, rather than a literal replay of evolutionary history. The central question is whether a biologically constrained visual phenotype—here, the number and spectral tuning of opsins—moves toward independently observed biological configurations when it is optimized under an explicit ecological demand. This section explains the origin of that modelling paradigm, why spatial recognition tasks are informative for the cases studied here, and what conclusions the resulting agreement can and cannot support.

A.5.1 Origin in Computational Visual Ecology

Using task performance to operationalize selection pressure predates this work. Morgenstern *et al.* evolved biologically inspired networks against the statistics of natural spectra and obtained colour contrast and constancy without specifying them as separate objectives [62]. Matthews *et al.* combined measured fruit and leaf spectra with their spatial arrangement, then searched cone sensitivities that best preserved fruit–foliage patterns; the predicted M-cone optimum was close to the biological value [57]. These studies establish a computational visual-ecology paradigm in which an ecological function is translated into a measurable visual objective and the resulting phenotype is compared with independent biological evidence.

Our work also follows the experimental logic of paleoinspired and robotics-inspired biology: a controllable surrogate is used to isolate a trait and evaluate its performance landscape when the corresponding historical system cannot be directly manipulated [29, 38]. Our distinct contribution is to expose opsin number and peak sensitivity as the controlled phenotype, embed retinal spectral transduction in a trainable perception system, and test several documented transitions under matched environmental assumptions.

Subsequent independent work by Tiwary *et al.* provides further support for this direction. Their embodied agents co-evolve eye morphology, neural processing, and behaviour under navigation, food–poison discrimination, and tracking objectives, and show that different task pressures lead to different visual systems [90]. That study does not validate our individual reconstructions, but its results strengthen the broader premise that task-conditioned computational evolution can serve as a scientific sandbox for testing hypotheses about vision.

A.5.2 Why Use a Semantic Task?

We do not treat generic semantic segmentation accuracy as biological fitness. Instead, each task instantiates a specific survival-relevant visual demand. In MinneApple, apples are partially occluded by and spectrally confusable with foliage [35]. Recovering their category, location, boundary, and spatial extent therefore operationalizes fruit–foliage discrimination during foraging. Field observations provide an independent biological link: wild trichromatic capuchins consume conspicuously coloured fruit more efficiently than dichromats [59], while colour-vision phenotype also changes how capuchins combine visual and olfactory cues during food selection [60]. Matthews *et al.* further show that preserving the spatial pattern of fruit against leaves is informative for predicting cone tuning [57].

Segmentation is useful here because it jointly measures discrimination and figure–ground organization rather than only image-level recognition. Figure–ground separation is a fundamental operation in biological vision [96, 69]; in machine vision, promptable segmentation also transfers mask prediction across image distributions and downstream tasks [43]. These observations motivate segmentation as a spatially informative interface, but they do not make mIoU a direct measurement of reproductive success. In our experiments, mIoU is an operational score for a controlled ecological task, and the learned spectral response is interpreted only after comparison with external biological observations.

A.5.3 Distinguishing Adaptation from a Task Artifact

A task-optimized spectrum could reflect the selected dataset or architecture rather than a general adaptive pressure. We therefore require converging evidence. First, the reconstructed direction must agree with measurements that were not used as optimization targets, such as documented opsin loss or duplication, spectral divergence, and depth-dependent RH1 blue shift. Second, the direction should

persist when the downstream model changes. Third, related ecological objectives should produce comparable spectral solutions.

The experiments in Sec. A.6 directly evaluate the latter two requirements. Across six segmentation encoders, the duplicated long-wave channels separate in the same direction. Segmentation, detection, and counting also recover closely located spectral optima while improving their respective task metrics. This does not prove that segmentation performance equals ecological fitness, but it substantially reduces the alternative explanation that the reported spectral adaptation is an artifact of one loss function or one backbone. The quantitative squirrelfish comparison provides a complementary test against specimen-based measurements rather than another machine-vision metric.

A.5.4 Evidence Levels and Claim Boundary

The biological cases do not have equal evidential status. Mammalian opsin loss and primate LWS/MWS duplication are strongly supported by molecular and comparative studies [9, 6, 104]; the causal claim that frugivory was the primary pressure behind trichromacy is less certain [19]. Confidence is highest for the direction of squirrelfish RH1 blue shift because specimen measurements, phylogenetic reconstruction, and molecular analysis agree with the predicted trend [105]. The dim-light advantage of colour-vision deficiency remains uncertain because behavioural findings are limited, while multiple RH1 genes in deep-sea fish are well established but their specific causal relationship to bioluminescence is not directly proven [63]. The Martian-life experiment is explicitly a speculative extrapolation with no biological validation.

Accordingly, agreement between optimized and measured spectra supports the framework as a controlled method for testing whether a proposed pressure is sufficient to produce a known direction of adaptation. It does not establish that the pressure was historically necessary, uniquely causal, or equivalent to organism-level fitness. Direct behavioural, genetic, phylogenetic, and physiological evidence remains the standard for those stronger claims.

A.6 Additional Experimental Evidence

This section evaluates the robustness of task-driven spectral adaptation across recognition architectures and objectives, and quantifies agreement with measured squirrelfish spectral sensitivities. General denotes the fixed initial responses and Design denotes the optimized responses within each comparison. These experiments test the dependence of the findings on the downstream model and task; they do not identify a unique historical evolutionary trajectory.

A.6.1 Evolutionary Reconstruction Across Backbones

The MinneApple experiments start from peaks of $(425, 560, 560)$ nm and compare six segmentation encoders. Design improves mIoU and reduces the camouflage score S_R for every tested encoder (Table 17). The two longer-wavelength responses diverge in a similar direction, despite differences in the perception network.

The reported optimized short-wavelength peaks also differ from 425 nm. We therefore present the complete reported peak vectors without describing that channel as fixed in these additional experiments. The initialization is shared, but a fixed-channel constraint cannot be inferred from the table. The reported values are not accompanied by seed-level uncertainty estimates.

A.6.2 Reconstruction Across Recognition Tasks

With the original MiT backbone, the additional experiments compare segmentation, detection, and fruit counting on MinneApple (Table 18). Detection AP@50/AP@75 increases from 0.592/0.411 to 0.617/0.435, and counting MAE/RMSE decreases from 13.1/18.3 to 11.1/17.3. The learned longer-wavelength peaks remain similar across the three tasks. Thus, the observed direction of spectral adaptation is not confined to segmentation in these experiments. AP, F1, and counting errors retain their source units and are not converted into mIoU.

A.6.3 Camera Design Across Backbones

On MODID and MHSI, Design improves mIoU over General in all twelve dataset–encoder combinations (Table 19). Values in this table are fractions on $[0, 1]$; for example, MiT on MODID improves

Table 17: Dichromacy-to-trichromacy reconstruction across segmentation encoders on MinneApple. Lower S_R is better.

Encoder	Opsin	$\lambda_{3\,\max}$	$\lambda_{2\,\max}$ (nm)	$\lambda_{1\,\max}$	mIoU (%)	S_R
MiT	General	425.00	560.00	560.00	47.29	0.276
	Design	440.65	547.34	582.10	47.99	0.239
PVT-v2	General	425.00	560.00	560.00	49.21	0.265
	Design	430.33	542.15	578.22	50.67	0.222
ConvNeXt	General	425.00	560.00	560.00	46.89	0.270
	Design	434.42	545.93	584.41	47.52	0.239
SwinV2	General	425.00	560.00	560.00	48.66	0.255
	Design	425.56	544.73	584.01	50.05	0.225
InternImage	General	425.00	560.00	560.00	47.89	0.262
	Design	420.10	533.26	576.12	48.65	0.238
MambaVision	General	425.00	560.00	560.00	51.15	0.244
	Design	426.64	545.03	579.91	52.21	0.219

Table 18: Dichromacy-to-trichromacy reconstruction across downstream tasks on MinneApple.

Task	Opsin	$\lambda_{3\,\max}$	$\lambda_{2\,\max}$ (nm)	$\lambda_{1\,\max}$	Task-specific performance
Segmentation	General	425.00	560.00	560.00	mIoU $= 47.29\%$, $S_R = 0.276$
	Design	440.65	547.34	582.10	mIoU $= 47.99\%$, $S_R = 0.239$
Detection	General	425.00	560.00	560.00	AP@50/75 $= 0.592/0.411$, mAP $= 0.398$, F1 $= 0.6805$
	Design	439.22	545.24	581.45	AP@50/75 $= 0.617/0.435$, mAP $= 0.409$, F1 $= 0.7021$
Counting	General	425.00	560.00	560.00	MAE/RMSE $= 13.1/18.3$
	Design	435.65	546.16	580.83	MAE/RMSE $= 11.1/17.3$

from 0.5642 to 0.5867, equivalent to 2.25 percentage points. This supports complementarity between spectral optimization and backbone selection in the reported setting, without establishing significance in the absence of variances or confidence intervals.

Table 19: Camera-design robustness across segmentation encoders.

Encoder	Opsin	MODID mIoU	MHSI mIoU
MiT	General	0.5642	0.5435
	Design	0.5867	0.5636
PVT-v2	General	0.5924	0.5670
	Design	0.6057	0.5767
ConvNeXt	General	0.5350	0.5186
	Design	0.5403	0.5250
SwinV2	General	0.5141	0.5539
	Design	0.5426	0.5600
InternImage	General	0.5725	0.5672
	Design	0.5871	0.5812
MambaVision	General	0.6155	0.5849
	Design	0.6301	0.5946

A.6.4 Quantitative Agreement for Squirrelfish RH1

The learned RH1 peaks at depths of 0, 10, 50, and 70 m are 496.19, 488.73, 484.69, and 481.10 nm, respectively (Table 4). Their strictly decreasing rank order gives Spearman $\rho = -1$, and their endpoint difference is 15.09 nm. The predicted range falls within the reported 481–502 nm range across nine squirrelfish species; the 70 m prediction differs by 0.10 nm from the reported 481 nm value for the deep-dwelling *N. aurolineatus* [105]. This is an additional analysis of the existing depth experiment, not an independent training study or a species-matched estimate of historical causality.

A.6.5 Interpreting the Supplementary Comparisons

The full multi-rod trajectories in Tables 7–10 show a consistent interaction between channel count and simulated bioluminescence. At epoch 200, one/five rods achieve 11.14/13.77% mIoU without bioluminescence and 15.17/21.26% with it. The five-rod terminal peak spans are 0.33 and 3.98 nm, respectively. The larger separation supports a direction of adaptation within the model, while its small magnitude motivates richer ecological and photoreceptor models.

Filter-count experiments also expose task-dependent trade-offs. For MHSI, a single red filter yields 54.22% mIoU and an R/B pair yields 56.97%, an increase of 2.75 percentage points (Table 15). For Mars, two filters yield 40.78% and four yield 41.60% (Table 14). These camera-design comparisons use camera-specific initializations and should not be pooled with the hypothetical Martian evolutionary trajectories.

The U-Net result in Table 16 is labelled cancer-class IoU, not mIoU. Its Standard/Design scores are 52.33/54.97%. The source table does not identify a dataset and split unambiguously, so this result is retained as a separate architectural example rather than pooled with the MODID/MHSI backbone benchmark. Likewise, the stated five-fold, three-initialization protocol for oral cancer does not supply per-fold uncertainty and should not be interpreted as a significance test or automatically assigned to every additional experiment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see Abstract and Introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Sec. A.4

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The work includes proof of theory in Experiments in Sec.4, Sec.5 and Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We detail all the techniques for reproducing the results of our work in Appendix A.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code repository is `https://github.com/luyijie-research/PaleVision`. Access is currently restricted, and the repository does not yet provide all materials required to reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We detail all the Experimental setting/details for reproducing the results of our work in Appendix A.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: No such settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Appendix A.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work does not involve social impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not release data or models with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Appendix A.1

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our work does not introduce new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not study participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We use LLM only for writing and editing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.