SCAPE: Shift-variant Cortical-implant Adaptive Phosphene Encoding

Michelle Appel
Donders Institute for Brain, Cognition and Behaviour,
Radboud University, Nijmegen, The Netherlands
Email: michelle.appel@ru.nl

***Abstract*—Visual cortical implants aim to restore sight by electrically stimulating neurons through an array of electrodes. Existing encoding methods apply uniform or global image filters without accounting for the uneven spatial sampling imposed by the implant layout. We introduce SCAPE (Shift-variant Cortical-implant Adaptive Phosphene Encoding), a principled framework that adapts image processing to local electrode density. First, electrode coordinates are projected into visual field space and used to compute a continuous sampling density map via analytic magnification models or kernel density estimation. Next, Nyquist principles convert this density into a spatial scale map that specifies the highest resolvable frequency at each location. Finally, shift variant filtering applies a spatial kernel at each pixel whose width matches the local resolution limit. In an efficient example we implement a difference of Gaussians with separable convolutions, though SCAPE supports any kernel family. Integrated with a reconstruction decoder, SCAPE preserves structural detail and improves reconstruction accuracy across diverse natural scenes and implant configurations. By always presenting the appropriate amount of detail, SCAPE is compatible with any cortical implant scheme and paves the way for future behavioral and clinical studies.**

I. Introduction

Visual cortical prostheses offer a promising path to restore vision in individuals with severe visual impairment by directly stimulating populations of neurons in visual cortex. These devices, such as the Utah array and Neuralink's cortical implants, aim to bypass damaged retinal pathways and directly encode visual information into the brain. By stimulating neurons in a spatially organized manner, these implants can evoke perceptual phosphenes that correspond to visual stimuli.

However, a key challenge in designing effective cortical prostheses is the limited number of electrodes available for stimulation, which constrains the spatial resolution of the encoded visual information. This limitation necessitates careful consideration of how visual stimuli are processed and represented before being delivered to the implant. Current approaches to visual encoding for cortical implants often rely on uniform spatial filtering techniques, such as Sobel or Canny edge filtering, to reduce the dimensionality of visual input. While these methods can help manage the high spatial resolution of natural images, these conventional image preprocessing pipelines apply uniform filtering or feature extraction across the entire field of view, neglecting implant specific sampling density. This mismatch can produce oversmoothing in regions where electrodes are dense, degrading available detail, or unnecessary clutter in regions where electrodes are sparse, overwhelming limited channel capacity.

In this work we introduce SCAPE (Shift variant Cortical prosthesis Adaptive Phosphene Encoding), an adaptive encoding framework that tailors spatial filtering to the local resolvability of each implant configuration. SCAPE first estimates the sampling density from electrode or phosphene positions using analytic magnification models or kernel density estimation. It then maps density to a local spatial scale via Nyquist principles and applies shift variant filtering, implemented here as a difference of Gaussians, to the input image before phosphene rendering.

The main contributions of this paper are:

- A principled method for local sampling density estimation and shift variant spatial filtering tailored to cortical implant layouts.
- Comprehensive evaluation of SCAPE in simulation across multiple electrode configurations, including high density, Utah array, Neuralink shank, and receptive field based schemes.
- Benchmarking performance with low level fidelity metrics (SSIM, SR-SIM, VSI, FSIM, MDSI, Content, PIEAPP), representational similarity analysis, and reconstruction accuracy of an Attention UNet decoder.

II. Related Work

Efficiently encoding complex scenes for visual prostheses has progressed from simple contrast and edge detection to adaptive heuristics and end-to-end learned encoders. However, existing methods process the visual field uniformly and do not account for patient-specific electrode layouts or cortical magnification. This section reviews the evolution of image processing techniques for visual prostheses, highlighting the limitations of current approaches and the need for adaptive methods like SCAPE.

A. Image Processing for Visual Prostheses

Effective image processing is critical for maximizing the limited information conveyed by cortical visual prostheses. Traditional and recent methods have focused on extracting salient features or learning stimulus encoders, but none explicitly account for the patient's electrode layout or variations in cortical magnification. Ignoring these factors can lead to oversmoothing where electrodes are dense and visual clutter

where electrodes are sparse. SCAPE addresses this gap by adapting filtering to local sampling density.

1) Early Heuristic Methods: Under severe electrode count and bandwidth constraints, early approaches applied contrast and brightness enhancement, grayscale histogram equalization, and gradient-based edge detection operators such as Sobel and Canny to extract salient contours in low-resolution phosphene maps [8], [18]. These methods often binarized or morphologically filtered the output to suppress noise and highlight obstacles or object boundaries. However, they treated the visual field uniformly and did not consider implant-specific sampling schemes or the cortical magnification factor.

2) More Recent Learned Encodings: More recent strategies frame stimulus encoding as an optimization or learning problem. Relic et al. trained a convolutional neural network encoder in an end-to-end fashion with a differentiable phosphene simulator to predict electrode activation patterns for intelligible phosphenes on MNIST images [13]. Granley et al. developed hybrid neural autoencoders that invert neuroscientific forward models to generate patient-specific stimulation strategies, demonstrating improved fidelity over conventional encodings [4]. De Ruyter van Steveninck . proposed an end-to-end learned encoder that jointly optimizes spatial filtering and electrode selection, but still does not explicitly modulate processing based on local sampling density or cortical magnification [1].

B. Simulated Prosthetic Vision Pipelines

Simulated prosthetic vision pipelines convert visual inputs into phosphene maps by modeling the response to electrode stimulation. Early implementations often used static Gaussian phosphenes without explicit retinotopic mapping or cortical magnification, due to limited computational frameworks at the time. Immersive VR-SPV systems enabled user studies in virtual environments but relied on non-differentiable heuristics and uniform phosphene shapes, which constrained encoder development [6]. More recently, van der Grinten et al. released a fully differentiable PyTorch simulator that incorporates retinotopic projection, cortical magnification, current spread, temporal dynamics, and support for arbitrary electrode layouts, facilitating real-time rendering and gradient-based optimization of encoding algorithms [15].

C. Adaptive and Spatially-Varying Encoding Approaches

Several recent methods adapt encoding globally or per patient but still treat spatial filtering uniformly. Granley et al. introduced a human-in-the-loop Bayesian optimization framework that personalizes deep encoder parameters to individual patients, yielding improved perceived quality with minimal feedback [3]. Hybrid neural autoencoders have been used to invert biophysical forward models in an end-to-end fashion and generate patient-specific stimulation patterns [1], [4]. Despite these advances, no existing approach explicitly modulates the spatial scale of processing according to local electrode density or Nyquist limits. SCAPE fills this gap by continuously adapting its filtering kernel to the sampling density of each implant.

III. Methods

A. Phosphene Simulation Framework

To assess SCAPE under conditions that approximate clinical prosthetic vision we employ the simulator of van der Grinten et al. [15]. This pipeline models key aspects of cortical stimulation, including the retinotopic projection of electrodes into visual-field coordinates and the rendering of each electrode's percept as a Gaussian phosphene. The resulting phosphene image captures the spatial layout of perceptual activations and serves as the input for SCAPE's adaptive encoding stages.

1) Electrode Placement: We begin with E electrodes implanted in cortical tissue, each with a two-dimensional coordinate on the cortical surface:

$$\{(x_i, y_i)\}_{i=1}^{E}.$$

These positions are obtained from clinical implant schematics or patient-specific models.

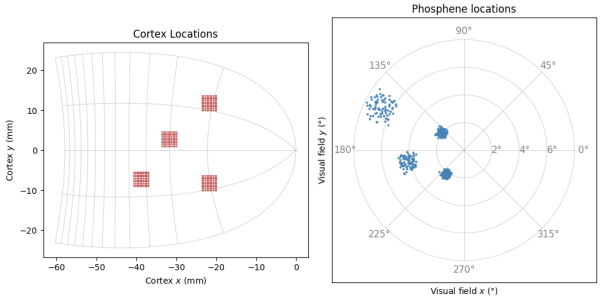

Fig. 1. Cortical electrode locations of 4 Utah Arrays (left) plotted in cortical x, y coordinates with overlaid retinotopic grid, and corresponding phosphene centers (right) in visual-field polar coordinates.

2) Retinotopic Projection and Phosphene Centers: Each cortical electrode (x_i, y_i) is mapped into visual-field coordinates $(\mu_{x,i}, \mu_{y,i})$ using the inverse of the Wedge–Dipole transform introduced by Polimeni *et al.* [11]. This model captures cortical magnification, in which the foveal region is allocated a larger cortical representation than the periphery. In complex notation the forward mapping from visual-field polar coordinates (r, θ) to cortical coordinate w is

$$w = k\big[\ln\big(r\,e^{i\alpha\theta} + a\big) - \ln\big(r\,e^{i\alpha\theta} + b\big)\big],$$

where r is eccentricity in degrees, θ is polar angle, k scales degrees to cortical millimeters, and a, b, α are fitted parameters. Analytically inverting this relation yields

$$r\,e^{i\alpha\theta} = \frac{b\,e^{\phi} - a}{1 - e^{\phi}}, \quad \phi = \frac{w}{k},$$

from which

$$r = \left|\tfrac{b\,e^{\phi}-a}{1-e^{\phi}}\right|, \quad \theta = \frac{1}{\alpha}\arg\!\left(\tfrac{b\,e^{\phi}-a}{1-e^{\phi}}\right).$$

Applying this inverse transform to each (x_i, y_i) and adding optional Gaussian angular noise and dropout produces $N \le E$ phosphene centers

$$\{(\mu_{x,i}, \mu_{y,i})\}_{i=1}^{N},$$

expressed in degrees of visual angle. These centers form the basis for SCAPE's density estimation and adaptive filtering

[15]. An example implant scheme is shown in Figure 1, where the left panel plots the cortical electrode locations of 4 Utah Arrays in x, y coordinates and the right panel shows the corresponding visual-field polar coordinates of the phosphenes.

3) Gaussian Blob Rendering: Empirical reports indicate that electrically evoked phosphenes are most often perceived as localized flashes of light with an approximately circular appearance [15]. Although some studies describe elongated or irregular forms, for simplicity we model each phosphene as an isotropic Gaussian. Note that SCAPE's core computations (density estimation and adaptive filtering) rely only on phosphene centers; the Gaussian shape is used downstream for visualization, evaluation, and amplitude normalization.

Formally, each phosphene center $(\mu_{x,i}, \mu_{y,i})$ in visual-field coordinates generates a two-dimensional Gaussian blob

$$G_i(x, y) = \exp\Big(-\frac{(x - \mu_{x,i})^2 + (y - \mu_{y,i})^2}{2\,\sigma^2}\Big),$$

where (x, y) are degrees of visual angle and σ is the nominal phosphene radius. The simulator rasterizes these Gaussians onto a regular image grid by converting $(\mu_{x,i}, \mu_{y,i})$ into pixel positions according to the chosen field-of-view and resolution, evaluating G_i at every grid point, and summing across all N phosphenes:

$$I_{\text{raw}}(x, y) = \sum_{i=1}^{N} G_i(x, y).$$

This raw phosphene map is then used for visualization, metric evaluation, and amplitude equalization but does not influence SCAPE's density or filter-scale computations. An example of this rendering is shown in Figure 2, where the left panel plots the centers of the phosphenes in visual-field coordinates and the right panel shows the corresponding Gaussian blobs.

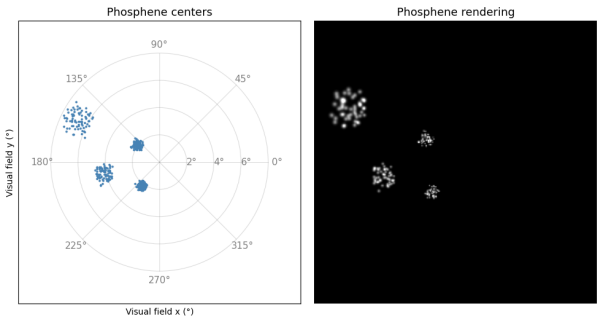

Fig. 2. Simulation of phosphene responses for a 4-Utah-array cortical implant. **Left:** Phosphene center locations in visual-field coordinates, shown on a polar grid spanning ±8° of eccentricity. **Right:** Corresponding Gaussian-blob rendering of phosphenes for a nominal stimulus amplitude (80 µA). This rendering illustrates how spatially discrete electrode activations translate into blurred perceptual spots that vary with cortical sampling density.

4) Activation-to-Electrode Mapping: To evoke a specific percept with a cortical implant, we ultimately need to decide which electrodes to turn on and how much current to send through each. A convenient intermediate representation is a continuous *activation map*

$$A(x, y) \in [0, 1]$$

defined over the visual field. Intuitively, brighter regions of this map correspond to stronger intended stimulation, and darker regions to weaker or no stimulation.

Given the N electrode centers $\{(\mu_i^x, \mu_i^y)\}_{i=1}^N$, we sample this map to produce a raw activation vector $\mathbf{a} = (a_1, \ldots, a_N)$. Two common sampling strategies are:

- **Point sampling:** $a_i = A(\mu_i^x, \mu_i^y)$.
- **Local pooling:** $a_i = \max_{(x,y)\in R_i} A(x, y)$, where R_i is a small neighborhood around (μ_i^x, μ_i^y) (e.g. a disk whose radius reflects local magnification).

In our simulator implementation, we build a direct correspondence between each electrode and pixel(s) of the activation map by precomputing, for every electrode i, a "distance map" on the visual-field grid. The simulator then uses these distance maps to (a) find the nearest pixel index for point sampling, or (b) define a binary mask for region pooling, entirely in pixel-space.

In a real implant system, the same correspondence can instead be obtained analytically: one applies the inverse retinotopic mapping (e.g. the wedge-dipole transform) to compute exactly which image coordinates fall under each electrode's receptive field, without requiring any precomputed pixel lookup.

Because $A(x, y)$ was normalized into $[0, 1]$, we then apply a global stimulus scale S (in Amperes) to obtain the actual per–electrode currents

$$I_i = S\, a_i.$$

Choosing S appropriately ensures that we stay within safe charge-delivery limits while preserving the relative pattern of activation. This current vector $\{I_i\}$ is then handed to the phosphene simulator, which applies temporal filtering, thresholding, and Gaussian-blob rendering to generate the final perceptual image.

In the end, our goal is to design or optimize the activation map $A(x, y)$ (and thus the stimulus vector $\{I_i\}$) so that the resulting phosphenes are as clear and interpretable as possible.

5) Amplitude Equalization: Phosphene brightness does not scale linearly with electrode current. Local electrode density and current spread interact with multiple nonlinear stages, such as overlapping Gaussian-blob summation, saturation in sigmoid transforms, thresholding and temporal filtering, which cause some regions of the percept to appear disproportionately bright or dim. Because these nonlinear effects accumulate, there is no practical closed-form inverse that maps a target brightness profile back to electrode currents.

To achieve a uniform perceptual dynamic range, we apply a simple gain-learning step after simulation. Each phosphene i has an initial amplitude $A_i = 1$. We first drive all electrodes at the same nominal current S, render the raw percept

$$I_{\text{raw}}(x, y) = \sum_{i=1}^{N} S\, G_i(x, y),$$

and measure each blob's peak intensity,

$$m_i = \max_{x,y} G_i(x, y).$$

We then optimize the gains $\{A_i\}$ by minimizing

$$\mathcal{L}(A) = \frac{1}{N}\sum_{i=1}^{N}\big(A_i\, m_i - m^*\big)^2,$$

updating each A_i by gradient descent with learning rate η and clamping to $[A_{\min}, A_{\max}]$. After convergence the normalized percept

$$I_{\text{norm}}(x,y) = \sum_{i=1}^{N} A_i\, S\, G_i(x,y)$$

has phosphenes with comparable peak brightness, improving visual consistency for evaluation and downstream decoding.

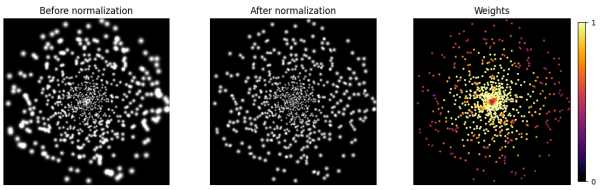

Fig. 3. Dynamic amplitude normalization example. *Left:* raw percept obtained by driving every electrode at the same current S. *Center:* normalized percept after learning per-electrode gains A_i. *Right:* learned gains $A_i \in [A_{\min}, A_{\max}]$. Normalization yields a more even brightness profile and reduces total current draw, conserving implant power and minimizing excessive charge delivery.

This calibration step yields a more uniform perceptual phosphene map, improving clarity and interpretability of encoded images. Compressing the amplitude range also lowers the total current required, which reduces neural fatigue and extends implant battery life. Although amplitude equalization is not required by SCAPE's core processing, we apply it here to standardize perceptual output for evaluation and comparison.

B. SCAPE Adaptive Encoding

The fundamental challenge in cortical prosthetic vision is that electrode arrays sample the visual field nonuniformly. Regions with dense electrode coverage can resolve fine detail while sparse regions cannot. SCAPE addresses this by adapting image filtering to the local sampling density. First, we estimate a continuous density map from the simulator-derived phosphene centers. Next, we convert density into a spatial scale map via sampling-theorem principles. Finally, we apply a shift-variant filter whose parameters vary continuously across the image. This adaptive encoding preserves maximal detail where it is supported and reduces visual clutter where it is not.

Figure 4 illustrates the full SCAPE pipeline at a glance: from electrode layout to phosphene centers (Panel 1), to density and σ-mapping (Panels 2–3), to the final shift-variant filter output (Panel 4).

1) Density Estimation: To guide adaptive filtering, SCAPE first computes a smooth sampling density $d(x,y)$ over the visual field. This density reflects how densely the implant probes each region and sets the local spatial resolution limit. Two practical estimators are used.

a) Analytic Cortical Magnification: In an idealized implant that uniformly samples cortex around the fovea, one can predict density purely from known retinotopy. Writing eccentricity $r = \sqrt{x^2+y^2}$ in degrees, the dipole-model magnification

$$M(r) = \frac{k}{2\pi}\Big(\frac{1}{r+a} - \frac{1}{r+b}\Big)$$

(with parameters k, a, b fit to human data [15]) gives a nominal density

$$d_{\text{analytic}}(x,y) = \frac{M(r)}{r}.$$

We then scale d_{analytic} so that its integral equals the total number of phosphenes N:

$$\iint d_{\text{analytic}}(x,y)\,\mathrm{d}x\,\mathrm{d}y = N.$$

b) Adaptive Kernel Density Estimation: For real implants with nonuniform phosphene layouts, we derive density directly from the simulator-produced phosphene centers $\{(\mu_{x,i}, \mu_{y,i})\}$. We place a two-dimensional Gaussian kernel at each center, choosing its bandwidth h_i based on local spacing. Specifically, let $d_{(i,k)}$ be the distance from point i to its kth nearest neighbor; then

$$h_i = \alpha\, d_{(i,k)},$$

where α (typically 1.0) controls smoothing. The density is

$$d_{\text{KDE}}(x,y) = \sum_{i=1}^{N} \frac{1}{2\pi\, h_i^2}\exp\Big(-\frac{(x-\mu_{x,i})^2 + (y-\mu_{y,i})^2}{2\, h_i^2}\Big).$$

Finally we normalize so that

$$\iint d_{\text{KDE}}(x,y)\,\mathrm{d}x\,\mathrm{d}y = N.$$

This adaptive KDE provides a flexible density estimate that captures local variations in electrode coverage.

2) σ-Mapping via Nyquist Principles: Once a smooth density map $d(x,y)$ is available, SCAPE computes a spatial-scale map $\sigma(x,y)$ that indicates the precise local sampling limit. By the Nyquist sampling theorem [9], the highest spatial frequency resolvable at (x,y) is

$$f_{\text{Nyq}}(x,y) = \tfrac{1}{2}\sqrt{\frac{d(x,y)}{\pi}},$$

since a local density of d phosphenes per unit area implies an average spacing of $\Delta \approx \sqrt{1/d}$. To match this limit exactly, we set

$$\sigma(x,y) = \frac{\kappa}{f_{\text{Nyq}}(x,y)} = 2\,\kappa\sqrt{\frac{\pi}{d(x,y)}},$$

where $\kappa > 0$ (typically 1) adjusts the transition sharpness. The resulting σ-map directly specifies the standard deviation of the local processing kernel at each location, thereby guiding where and at what scale spatial frequencies should be preserved or suppressed.

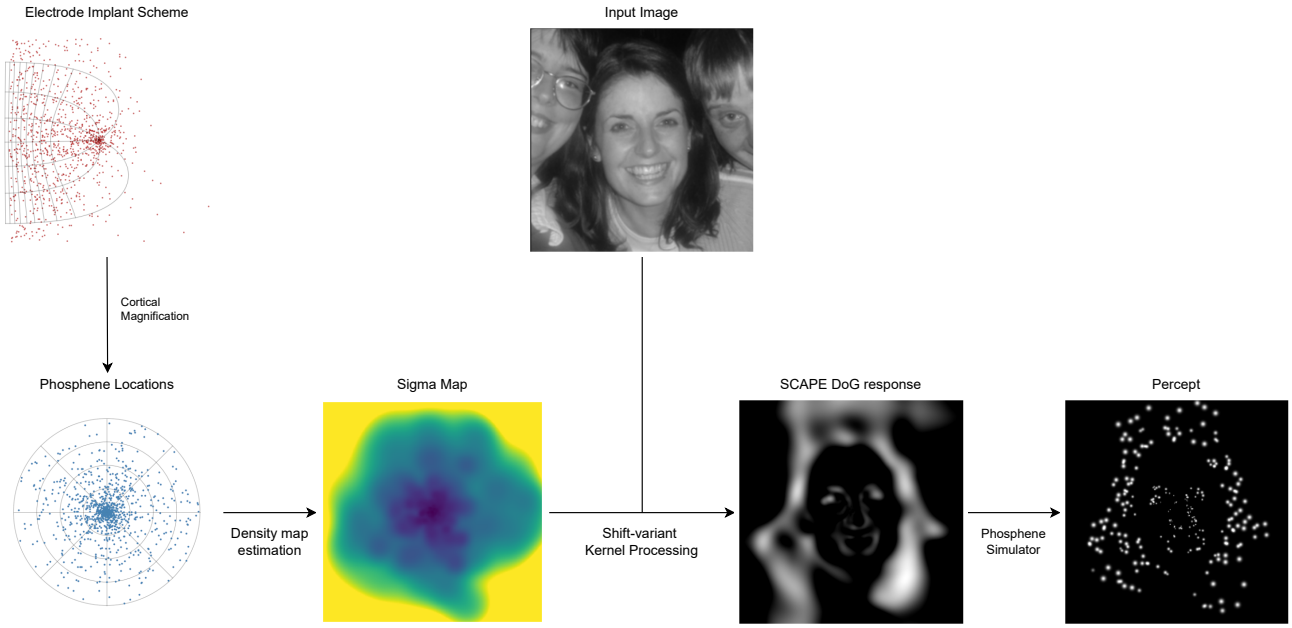

Fig. 4. Overview of the SCAPE pipeline. Starting from a cortical implant layout (top left) we obtain phosphene centers via the simulator, estimate a local density map, convert it into a spatial scale (σ) map, and then apply shift-variant filtering to produce an activation map for phosphene rendering.

3) Shift-variant Filtering: Having obtained a continuous filter-scale map $\sigma(x, y)$, SCAPE applies spatially adaptive filtering to the input image. At each location (x, y), a local kernel $K(\cdot; \sigma(x, y))$ is convolved with the image, yielding an output

$$I_{\text{filt}}(x, y) = \iint K\big(u, v; \sigma(x, y)\big)\, I_{\text{in}}(x - u, y - v)\, \mathrm{d}u\, \mathrm{d}v.$$

By tying the kernel's parameters to the local sampling density, shift-variant filtering preserves fine detail where electrodes are dense and reduces clutter where they are sparse. In the following sections we describe a concrete implementation using a difference of Gaussians and discuss how other kernel families can be incorporated within the same framework.

a) Difference-of-Gaussians Example: As a concrete illustration of SCAPE's adaptive filtering, we approximate the Laplacian-of-Gaussian (LoG) operator, widely used to model center–surround receptive fields in early vision [19], with a pair of Gaussian kernels. At each location (x, y), the LoG of the input image I would be

$$\mathrm{LoG}_\sigma(x, y) = \nabla^2 \big[G_\sigma * I\big](x, y),$$

where G_σ is a Gaussian of standard deviation $\sigma(x, y)$ and $*$ denotes convolution. Computing a full LoG at every pixel with its own σ has complexity $\mathcal{O}(n^2)$ per pixel, where n is the kernel size, making it impractical for real-time use. Instead, we approximate it by a Difference-of-Gaussians (DoG):

$$\mathrm{DoG}_\sigma(x, y) = \big[G_{\sigma_1} - G_{\sigma_2}\big] * I\,(x, y), \quad \sigma_2 = \lambda\, \sigma_1,$$

with $\lambda > 1$ (typically $\lambda = 1.6$) chosen so that $\mathrm{DoG} \approx \mathrm{LoG}$.

To implement this shift-variant DoG efficiently, we factor each Gaussian G_σ into separable one-dimensional kernels. This reduces the complexity to $\mathcal{O}(n)$ per pixel and allows us to modulate kernel width dynamically according to the local $\sigma(x, y)$ map. In practice this involves two passes of row- and column-wise convolutions with varying standard deviations, yielding real-time performance even on mobile hardware. This separable DoG serves as a simple yet powerful example of SCAPE's core idea: by varying filter scale across the image, we capture fine edges where phosphene density is high and suppress noise where it is low.

b) Extending to Other Kernels: While the DoG serves as a straightforward example, SCAPE's shift-variant framework accommodates any spatial filter family. For instance, one can replace the Gaussian kernels with orientation-tuned Gabor filters to emphasize contours aligned with cortical receptive-field preferences. More generally, the kernel $K(\cdot; \sigma(x, y))$ may be parameterized by a small set of basis functions (wavelets, steerable filters) whose shape adapts with σ. In future work one can even learn these kernels end-to-end: by embedding SCAPE into a differentiable reconstruction pipeline, the per-location filter weights can be optimized jointly with a decoder network. This flexibility allows SCAPE to capture both classical center–surround processing and more complex feature tuning while retaining its core principle of local, density-driven adaptation.

C. Reconstruction Decoder Integration

Human perceptual evaluation of phosphene encodings requires extensive behavioral studies and cannot be conducted at scale. Phosphene images are also fundamentally different from natural scenes, being sparse assemblies of localized blobs rather than continuous luminance patterns. To approximate how much visual information survives encoding, we train a

convolutional decoder to reconstruct the original scene from the phosphene map.

This decoder plays a role loosely analogous to downstream visual cortex: it must infer edges, textures, and object layouts from an abstract representation. Unlike the brain, however, the decoder can adjust all its parameters through gradient-based learning and is not constrained by biological priors. Despite these fundamental differences, a consistent improvement in reconstruction accuracy suggests that the encoding has preserved more of the scene's essential structure.

Our protocol is based on the end-to-end autoencoder framework of de Ruyter van Steveninck et al. [1], but here we fix the encoder to SCAPE and train only the decoder. By comparing reconstruction error and perceptual feature losses under identical training conditions, we derive a quantitative measure of how readily SCAPE's output can be interpreted, guiding future behavioral and clinical validation.

1) Attention-UNet Architecture: The reconstruction decoder employs an Attention-UNet, an extension of the original U-Net [14] with integrated attention and channel-recalibration mechanisms. Key components include:

- **Squeeze-and-Excitation (SE) blocks** to adaptively weight feature channels [5].
- **Dilated residual blocks** in the bottleneck for multi-scale context aggregation [20].
- **Spatial attention gates** on skip connections to emphasize salient regions [10].

Down-sampling is achieved via max-pooling and up-sampling via transposed convolutions. A final 1×1 convolution with sigmoid activation produces a reconstructed grayscale image. This configuration balances context integration and selective feature focus, making it effective for recovering scenes from sparse phosphene representations.

D. Evaluation Metrics

Comparing sparse phosphene encodings directly to natural images is inherently challenging, as conventional pixelwise measures assume dense, continuous luminance patterns. Nevertheless, low-level fidelity metrics provide useful insight into structural preservation when applied carefully to upsampled phosphene maps. To obtain a more complete assessment, we combine these fundamental measures with representational similarity analysis, which evaluates the geometry of stimulus relationships, and decoder-based reconstruction performance, which tests how accessible the encoded information is for downstream interpretation. Together, these complementary approaches offer a multifaceted evaluation of SCAPE's ability to convey visual content through the phosphene bottleneck.

1) Low-Level Fidelity Metrics: To quantify how faithfully SCAPE's phosphene output I_p preserves the original scene I_o, we apply full-reference image quality metrics that each compute a local similarity or error map and then pool these values into a single score. In the following paragraphs we describe how SSIM, VSI, MDSI and PIEAPP are formulated and computed for the pair (I_p, I_o).

a) Structural Similarity Index (SSIM): The Structural Similarity Index [17] evaluates local agreement in luminance, contrast, and structure. At each pixel (x, y) define

$$S_{\mathrm{SSIM}}(x,y) = \frac{2\,\mu_p(x,y)\,\mu_o(x,y) + C_1}{\mu_p(x,y)^2 + \mu_o(x,y)^2 + C_1} \times \frac{2\,\sigma_{po}(x,y) + C_2}{\sigma_p(x,y)^2 + \sigma_o(x,y)^2 + C_2},$$

where all statistics are computed over a small window around (x, y). The overall SSIM is

$$\mathrm{SSIM}(I_p, I_o) = \frac{1}{M\,N} \sum_{x=1}^{M} \sum_{y=1}^{N} S_{\mathrm{SSIM}}(x,y).$$

b) Spectral-Residual Saliency Index (SR-SIM): SR-SIM evaluates image fidelity by combining spectral-residual visual saliency with gradient-based structure similarity [21]. Given a reference I_o and a phosphene image I_p, let

$$\mathrm{SVRS}_o(x,y),\ \mathrm{SVRS}_p(x,y)$$

be their saliency maps computed via the spectral residual of the log-amplitude spectrum, and let

$$\mathrm{GM}_o(x,y),\ \mathrm{GM}_p(x,y)$$

be their gradient magnitude maps (Scharr-filtered). We form two local similarity measures

$$S_{\mathrm{S}}(x,y) = \frac{2\,\mathrm{SVRS}_o(x,y)\,\mathrm{SVRS}_p(x,y) + C_1}{\mathrm{SVRS}_o(x,y)^2 + \mathrm{SVRS}_p(x,y)^2 + C_1},$$

$$S_{\mathrm{G}}(x,y) = \frac{2\,\mathrm{GM}_o(x,y)\,\mathrm{GM}_p(x,y) + C_2}{\mathrm{GM}_o(x,y)^2 + \mathrm{GM}_p(x,y)^2 + C_2},$$

with constants C_1, C_2 for numerical stability. Defining $\mathrm{SVRS}_m(x,y) = \max\{\mathrm{SVRS}_o, \mathrm{SVRS}_p\}$, the fused local score is

$$S_L(x,y) = \big[S_{\mathrm{S}}(x,y)\big]\,\big[S_{\mathrm{G}}(x,y)\big]^{\alpha}\,\mathrm{SVRS}_m(x,y),$$

where α (typically 0.5) balances structure sensitivity. The global SR-SIM index is the normalized sum

$$\text{SR-SIM}(I_p, I_o) = \frac{\sum_{x,y} S_L(x,y)}{\sum_{x,y} \mathrm{SVRS}_m(x,y)}.$$

This metric captures both saliency-weighted structural agreement and overall contrast fidelity in a computationally efficient manner.

c) Visual Saliency Induced Index (VSI): VSI [22] augments structural similarity with a model of visual attention. First, saliency maps $v_o(x,y)$ and $v_p(x,y)$ are computed for the reference I_o and phosphene image I_p using the SDSP spectral-residual approach, which combines spectral residuals, spatial distribution and color cues to highlight regions likely to attract gaze. These maps are normalized to the range $[0, 1]$.

Next, three local similarity maps are formed:

$$s_{\mathrm{vs}}(x,y) = \mathrm{sim}\big(v_o(x,y),\, v_p(x,y)\big),$$
$$s_g(x,y) = \mathrm{sim}\big(\nabla I_o(x,y),\, \nabla I_p(x,y)\big),$$
$$s_c(x,y) = \mathrm{sim}\big(C_o(x,y),\, C_p(x,y)\big),$$

where $\mathrm{sim}(a,b) = \frac{2ab+C}{a^2+b^2+C}$ is the standard similarity kernel, ∇ denotes the Scharr-filtered gradient magnitude in luminance, and C_o, C_p are the two chromatic channels in a perceptual color space. These maps are fused into a single local score

$$S(x,y) = s_{\mathrm{vs}}(x,y)\,\big[s_g(x,y)\big]^{\alpha}\,\big[s_c(x,y)\big]^{\beta},$$

with empirically chosen exponents α and β. Finally, a per-pixel weight

$$W(x,y) \;=\; \max\big(v_o(x,y),\, v_p(x,y)\big)$$

emphasizes salient regions. The global VSI is the saliency-weighted average:

$$\mathrm{VSI}(I_p, I_o) = \frac{\sum_{x,y} W(x,y)\, S(x,y)}{\sum_{x,y} W(x,y)}.$$

d) Feature Similarity Index (FSIM): The Feature Similarity Index (FSIM) evaluates image fidelity by comparing low-level features known to align with human perception: phase congruency (PC) and gradient magnitude (GM) [23]. Given a reference image I_o and a phosphene rendering I_p, we first extract their PC maps $\mathrm{PC}_o(x,y)$, $\mathrm{PC}_p(x,y)$ and GM maps $\mathrm{GM}_o(x,y)$, $\mathrm{GM}_p(x,y)$. At each pixel (x,y) we compute two similarity measures:

$$S_{\mathrm{PC}}(x,y) = \frac{2\,\mathrm{PC}_o(x,y)\,\mathrm{PC}_p(x,y) + T_1}{\mathrm{PC}_o(x,y)^2 + \mathrm{PC}_p(x,y)^2 + T_1},$$

$$S_{\mathrm{G}}(x,y) = \frac{2\,\mathrm{GM}_o(x,y)\,\mathrm{GM}_p(x,y) + T_2}{\mathrm{GM}_o(x,y)^2 + \mathrm{GM}_p(x,y)^2 + T_2},$$

where T_1, T_2 stabilize the ratios. We then form a fused local score

$$S_L(x,y) = \big[S_{\mathrm{PC}}(x,y)\big]^{\alpha}\,\big[S_{\mathrm{G}}(x,y)\big]^{\beta},$$

with $\alpha = \beta = 1$. Finally, FSIM weights each location by the maximal phase congruency $\mathrm{PC}_m(x,y) = \max\{\mathrm{PC}_o(x,y), \mathrm{PC}_p(x,y)\}$, yielding

$$\mathrm{FSIM}(I_p, I_o) = \frac{\sum_{x,y} \mathrm{PC}_m(x,y)\, S_L(x,y)}{\sum_{x,y} \mathrm{PC}_m(x,y)}.$$

This formulation emphasizes perceptually salient features while penalizing structural deviations between I_p and I_o.

e) Mean Deviation Similarity Index (MDSI): MDSI [25] assesses image fidelity by combining gradient and chromaticity similarities with deviation-based pooling, making it robust to localized distortions. Given a reference I_o and a phosphene image I_p, let ∇I_o and ∇I_p be their gradient magnitude maps (computed via Prewitt filters) and let C_o, M_o and C_p, M_p be the two opponent color channels in a perceptual color space. Define at each pixel (x,y):

$$\mathrm{GS}(x,y) = \frac{2\,\nabla I_o(x,y)\,\nabla I_p(x,y) + c_1}{\nabla I_o(x,y)^2 + \nabla I_p(x,y)^2 + c_1},$$

$$\mathrm{CS}(x,y) = \frac{2\big(C_o(x,y)\,C_p(x,y) + M_o(x,y)\,M_p(x,y)\big) + c_3}{C_o(x,y)^2 + M_o(x,y)^2 + C_p(x,y)^2 + M_p(x,y)^2 + c_3}.$$

These maps are fused into a combined similarity

$$S(x,y) = \alpha\,\mathrm{GS}(x,y) + (1-\alpha)\,\mathrm{CS}(x,y),$$

with weighting α. To emphasize pixels where similarity deviates most from its mean, we compute

$$s_i = S(x_i,y_i)^{1/4}, \quad \bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i,$$

and pool by the Minkowski deviation

$$\mathrm{MDSI}(I_p, I_o) = \left[\frac{1}{N}\sum_{i=1}^{N}\big|\, s_i - \bar{s}\big|\right]^{1/4}.$$

Deviations from the mean highlight regions where the phosphene encoding departs most from the reference structure, yielding a sensitive measure of localized fidelity loss.

f) Content Loss: To capture perceptually meaningful differences beyond pixel-wise errors, we measure the dissimilarity between deep feature representations of the phosphene image I_p and the original scene I_o extracted from a pretrained convolutional network. Denote by $\phi_\ell(I)$ the activation tensor at layer ℓ of a VGG-type model. The content loss is then

$$\mathcal{L}_{\mathrm{content}}(I_p, I_o) = \big\|\phi_\ell(I_p) \;-\; \phi_\ell(I_o)\big\|_2^2.$$

By choosing mid- to high-level layers, this loss emphasizes preservation of object structure and semantic content rather than low-level textures [2], [24]. In our experiments we normalize inputs using ImageNet statistics and extract features from VGG19, averaging across spatial dimensions to yield a single content-distance score per image pair. Lower content loss indicates that SCAPE's encoding retains more of the high-level scene layout that a human observer would deem important.

g) Perceptual Image-Error Assessment through Pairwise Preference (PIEAPP): PIEAPP [12] is a learned full-reference metric designed to predict perceptual error by modeling human preferences between image pairs. It operates by decomposing the input images into overlapping patches and computing learned feature representations. Given a reference image I_o and a phosphene image I_p, each patch k yields a feature embedding difference Δf_k. A two-stage regression network maps these differences to an estimated perceptual error $d_k \in \mathbb{R}$ and an associated confidence weight $w_k > 0$. Formally, the model computes

$$\Delta f_k = \varphi\big(P_k(I_o)\big) \;-\; \varphi\big(P_k(I_p)\big),$$

where $P_k(\cdot)$ extracts the k-th patch and φ is the learned feature encoder. The per-patch perceptual error is then

$$d_k = g\big(\Delta f_k\big),$$

where g is a regression module that predicts the perceived dissimilarity. Weights are obtained as

$$w_k = h\big(\Delta f_k\big),$$

with h a second regression branch modeling reliability. The final PIEAPP score aggregates over all K patches:

$$\mathrm{PIEAPP}(I_p, I_o) = \frac{\sum_{k=1}^{K} w_k\, d_k}{\sum_{k=1}^{K} w_k}.$$

This approach has been shown to correlate strongly with human judgments of perceptual error in natural images. In our

experiments, we use PIEAPP as an approximate measure of perceptual fidelity by treating I_o as the original input image and I_p as the phosphene representation. Although PIEAPP was not trained on sparse phosphene patterns, lower scores still suggest that the encoding preserves more information relevant to human perception. For full details, see [12] and the reference implementation.

2) Representational Similarity Analysis: Representational Similarity Analysis (RSA) is a widely used framework in neuroscience for comparing the structure of internal representations across different systems [7]. Instead of requiring spatial alignment or direct correspondence between individual activation patterns, RSA abstracts each representation into a pairwise dissimilarity matrix. This approach enables comparisons across measurement modalities with very different spatial resolution and scales, such as fMRI voxel patterns, single-unit electrophysiology, and artificial neural networks.

In classic applications, each stimulus is represented as a vector r_i of activations in a brain region (for example, responses in V1). The representational dissimilarity matrix (RDM) is then defined by

$$D_{ij} = \delta\big(r_i, r_j\big),$$

where δ is a dissimilarity metric such as correlation distance

$$\delta_{\mathrm{corr}}(r_i, r_j) = \frac{1 - \mathrm{corr}(r_i, r_j)}{2}.$$

This RDM characterizes how different the representations of all pairs of stimuli are relative to each other.

Conceptually, RSA offers what Kriegeskorte et al. termed a "second-order" representational comparison: while first-order comparisons measure similarity between raw activations, second-order comparisons examine whether the geometric relationships between stimuli are preserved. In neuroscience, such second-order comparisons have been used to show that different cortical regions (e.g., V1 vs. higher visual areas) encode distinct relational structures that align with perceptual or semantic similarity [16].

This principle motivates our use of RSA for phosphene encoding. Phosphenes are, in a sense, direct artificial activations of V1. Although our encoding is constrained by electrode sampling, it still forms a structured response space, conceptually akin to biological V1 patterns. By building RDMs of phosphene renderings, we can ask whether the relational geometry of the stimulus space is retained. If it is, then different stimuli remain discriminable in the same way, even if their absolute fidelity is degraded.

Practically, for a set of N images, we flatten each phosphene rendering into a vector

$$r_i = \mathrm{vec}\big(I_p(i)\big),$$

and compute the correlation distance matrix

$$D_p(i, j) = \frac{1 - \mathrm{corr}\big(r_i, r_j\big)}{2}.$$

Similarly, we compute an RDM for the reference images I_o. To quantify second-order similarity, we correlate the upper triangles of these matrices:

$$\rho = \mathrm{corr}_{\mathrm{Spearman}}\Big(\mathrm{vec}\big(D_p^{\mathrm{upper}}\big),\ \mathrm{vec}\big(D_o^{\mathrm{upper}}\big)\Big).$$

High ρ indicates that the pairwise relationships between stimuli are well preserved by the encoding, even if pixelwise error is high. Low ρ suggests that the encoding distorts the relational structure, potentially impairing discriminability.

Although we compute dissimilarities here in pixel space, RSA can be applied in any feature space. For example, one could use embeddings from a pretrained neural network (e.g., VGG activations), perceptual models, or even behavioral similarity judgments. This flexibility allows RSA to bridge low-level and high-level aspects of representation in a unified framework.

In summary, RSA enables us to quantify whether phosphene encodings retain the relative distinctions between stimuli: a property that is often more relevant for perception and recognition than absolute pixel fidelity alone.

3) Reconstruction Performance: As a final evaluation axis, we assess how effectively a decoder network can reconstruct the original input image from the phosphene representation. Unlike the low-level fidelity metrics applied directly to the phosphene maps, this approach yields reconstructed natural images, making it more appropriate to apply conventional perceptual and pixel-based quality measures.

Specifically, for each phosphene-encoded input, the trained decoder produces a reconstructed image $\hat{I}_o$. We then compare $\hat{I}_o$ to the original reference image I_o using a suite of complementary metrics. These include pixelwise measures such as mean squared error (MSE) and SSIM, as well as perceptual similarity metrics such as LPIPS, VGG-based perceptual distance, DISTS, and PIEAPP. For completeness, we also report advanced full-reference metrics including FSIM, MDSI, VSI, and SRSIM. Together, these scores provide a multifaceted view of how much visual detail the encoding preserves in a form usable for end-to-end reconstruction.

Because the reconstructions are natural images, these metrics are applied without further adaptation. Higher perceptual similarity and lower pixelwise error indicate that the encoding retains more information useful for reconstructing recognizable content.

IV. Experiments

The experiments are designed to evaluate how well SCAPE adapts to different visual scenes, implant layouts, and spatial sampling densities compared to non-adaptive baselines. We assess performance across multiple datasets with diverse image statistics and test a range of implant schemes that vary in electrode count and distribution. All methods are processed through the same prosthetic vision simulation framework to ensure a fair comparison. Performance is quantified using complementary approaches: low-level fidelity metrics applied to phosphene renderings, representational similarity analysis to assess preservation of stimulus relationships, and reconstruction-based evaluation to gauge how well a learned decoder can recover the original scene from each encoding. This multi-faceted evaluation allows us to capture both the structural accuracy of the encoded images and their potential usability for downstream perception.

A. Datasets

We evaluate SCAPE using three publicly available datasets chosen to cover a wide range of visual statistics and content domains:

- **LaPa**: human face images, representing structured, high-contrast features and smooth shading.
- **MS COCO**: diverse natural scenes and objects, providing varied spatial frequencies and complex layouts.
- **SUN**: indoor and outdoor scene photographs, including both cluttered and open environments.

All images are first converted to grayscale to reflect the fact that brightness is the dominant cue available in prosthetic vision, where phosphenes primarily convey intensity rather than color. In early visual cortex, luminance signals are more strongly represented than chromatic signals, making brightness the most relevant dimension for encoding (citation to be added). Preliminary experiments with color confirmed that intensity alone preserved the majority of task-relevant structure.

Grayscale conversion is performed using **Rec. 709 perceptual luminance** (luma) weighting, which aligns with human brightness perception:

$$Y = 0.2126\,R + 0.7152\,G + 0.0722\,B, \tag{1}$$

where R, G, and B are normalized to $[0, 1]$. After conversion, perceptual normalization is applied to ensure a consistent dynamic range across images before encoding.

B. Implant Schemes

We evaluate SCAPE across five implant layouts that span large differences in sampling density, spatial extent, and geometric regularity. This set stresses density adaptation both in the foveal region and in the periphery, and it includes layouts used in recent prosthetic vision studies.

a) Uniform 1024: This layout provides a dense, near-uniform sampling of the visual field and serves as a standard reference in the literature on differentiable prosthetic vision. It allows direct comparison to prior work that uses a similar order of magnitude in channel count [1]. It also exposes whether SCAPE preserves fine structure when the density is sufficient across most of the field of view.

b) Four Utah arrays: This layout models a realistic multi-array configuration with gaps between arrays and moderate channel count. It introduces strong spatial inhomogeneity due to array borders and inter-array spacing, which is a common feature of practical cortical implants. It therefore tests whether SCAPE can suppress clutter in sparse regions while preserving detail within array footprints.

c) One Utah array: This layout matches a single-array setup and concentrates sampling near fixation with a very small field of view. It reflects ongoing experimental configurations used in current research programs at the Institute of Bioengineering of the Miguel Hernández University. It tests SCAPE under severe channel limits and minimal coverage, which is useful for understanding performance in constrained early clinical or preclinical settings.

d) Neuralink shank: This layout represents a high-site-count configuration that covers a wide field of view with many closely spaced sites. It stresses computational efficiency and scale selection, since local density varies with eccentricity and along the shank geometry. It helps determine whether SCAPE scales to thousands of sites while maintaining efficiency and stability.

e) Utah RFs: This layout is derived from receptive-field considerations anchored to a Utah-like geometry. It provides intermediate coverage between a single array and a four-array montage, and it yields a more graded eccentricity profile than a strict grid. It serves as a control that links analytic receptive-field placement to a hardware-inspired layout, which is useful for validating the density-to-scale mapping without strong grid artifacts.

Together these schemes span two orders of magnitude in channel count, from 94 to 4224. They also span field of view from a fraction of a degree to more than twelve degrees of eccentricity. This diversity is important because SCAPE is designed to set filter scale from local sampling density. The chosen set probes that mechanism in both uniform and highly inhomogeneous layouts, which is essential for a fair test of adaptive encoding.

C. Baselines

We compare SCAPE to simple and widely used encoders that do not adapt their spatial scale to local sampling density. Each baseline produces an activation map that is passed through the same simulator and the same amplitude normalization as SCAPE to ensure a fair comparison.

a) Perceptual luminance: A direct perceptual luminance image is used as a minimalist baseline. Images are converted to grayscale with Rec. 709 weights and then forwarded without additional filtering. This tests how much structure survives without any feature selection.

b) Canny edge detection: Canny is a standard choice in prosthetic vision pipelines because it preserves object boundaries under severe spatial resolution limits and often improves recognition in simulated conditions. We include a Canny baseline with standard smoothing and hysteresis thresholding, applied uniformly across the field of view. This represents a strong nonadaptive feature extractor centred on contours.

c) Random control: As a structure free control we sample a random activation pattern that matches the total activation energy of the method under test. This helps verify that improvements are not due to trivial differences in overall current or luminance.

All baselines use identical preprocessing, identical simulator settings, and identical amplitude normalization. This keeps the comparison focused on the effect of adaptive scale selection rather than on downstream rendering differences.

D. SCAPE Configuration

SCAPE builds a local density map from phosphene coordinates using adaptive kernel density estimation and converts that density into a spatial scale map $\sigma(x, y)$. The scale map

Scheme	View angle (deg)	Electrodes (N)	Eccentricity range (deg)
1 Utah array	0.4	94	0.009 to 0.203
Utah RFs	6.0	256	0.000 to 2.120
4 Utah arrays	16.0	320	1.632 to 7.931
Uniform 1024	16.0	1024	0.001 to 7.984
Neuralink	25.0	4224	0.000 to 12.095

TABLE I
IMPLANT LAYOUTS AND BASIC PROPERTIES. COUNTS REFLECT THE EFFECTIVE NUMBER OF ELECTRODES INSIDE THE SIMULATOR FIELD OF VIEW. ECCENTRICITY IS REPORTED IN DEGREES OF VISUAL ANGLE.

controls a shift variant Difference of Gaussians that runs with separable one dimensional passes.

a) Density estimation: All experiments use adaptive KDE over phosphene centers in visual field coordinates. Bandwidth h_i for point i equals α times the distance to its k-th nearest neighbor. We set $k = 16$ and $\alpha = 1.0$. The resulting density map is normalized so that its surface integral equals the total number of phosphenes of the scheme under test.

b) Mapping density to scale: We convert density $d(x, y)$ into a local scale through a Nyquist motivated rule. We use

$$\sigma_{\text{fov}}(x, y) \;=\; \frac{2}{\pi\sqrt{2}\,\beta}\,\frac{1}{\sqrt{d(x, y)}} \quad \text{with} \quad \beta = 0.55,$$

which yields a scale that decreases in regions with higher sampling density and increases in sparse regions. For filtering on an image grid we convert σ_{fov} from degrees to pixels using the simulator field of view and resolution.

c) Filter family and execution: We use a Difference of Gaussians with ratio $\lambda = 1.6$ to approximate a Laplacian of Gaussian while remaining efficient. The filter runs as two separable Gaussian passes per branch, first horizontal then vertical, at $\sigma_1(x, y)$ and $\sigma_2(x, y) = \lambda\,\sigma_1(x, y)$. Gaussian weights are truncated where their value falls below a small threshold and each one dimensional kernel is normalized to sum to one. Padding uses reflection at image borders. The half kernel radius is capped at a fixed maximum to bound runtime and memory.

d) Retinotopic model: Retinotopic projection follows the dipole form of the Polimeni wedge–dipole family with standard parameters. This model sets the visual field coordinate system used by the KDE and by the degree to pixel conversion.

e) Sampling to electrodes: The shift variant DoG produces an activation map on the simulator grid. Electrode currents are obtained by sampling the activation map at phosphene centers. The same sampling rule is used for every implant scheme.

f) Scope of configuration: Unless noted otherwise we do not apply extra smoothing or quantization to the σ map. All encoders, including baselines, share the same preprocessing, simulator settings, and the same amplitude normalization described in the next subsection.

E. Amplitude Normalization

Explain how amplitude equalization is applied and why.

F. Evaluation Protocol

Detail the metrics and evaluation procedures used, including:

- Low-level fidelity metrics
- Representational similarity analysis
- Reconstruction performance

V. RESULTS

A. Overview

Briefly summarize the key experimental findings and highlight the main trends across datasets and implant schemes.

B. Low-Level Fidelity Metrics

Present quantitative results for SCAPE and all baselines.

- Tables or bar plots comparing methods for each dataset.
- Highlight where SCAPE achieves the largest gains.
- Note any cases where baselines perform similarly or better.

C. Representational Similarity Analysis

Report RSA correlations between encoded and original image spaces.

- Show RSA scatter plots or correlation heatmaps.
- Compare correlations across implant schemes and datasets.
- Discuss preservation of relational structure.

D. Reconstruction Performance

Present decoder-based reconstruction results.

- Quantitative metrics for each dataset and encoding method.
- Example reconstructions for qualitative comparison.
- Highlight differences in how well structural detail is recovered.

E. Ablation and Sensitivity Analyses (Optional)

If performed, include:

- Effect of removing SCAPE components (e.g., density adaptation, amplitude normalization).
- Sensitivity to key parameters (e.g., λ in DoG).

F. Summary of Results

Provide a concise bullet-point recap of the main performance patterns and insights revealed by the experiments.

REFERENCES

[1] Jaap de Ruyter van Steveninck, Umut Güçlü, Richard van Wezel, and Marcel van Gerven. End-to-end optimization of prosthetic vision. *bioRxiv*, 2020.

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015.

[3] Jacob Granley, Tristan Fauvel, Matthew Chalk, and Michael Beyeler. Human-in-the-loop optimization for deep stimulus encoding in visual prostheses, 2023.

[4] Jacob Granley, Lucas Relic, and Michael Beyeler. Hybrid neural autoencoders for stimulus encoding in visual and other sensory neuroprostheses, 2022.

[5] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

[6] Justin Kasowski and Michael Beyeler. Immersive virtual reality simulations of bionic vision. In *Augmented Humans 2022*, AHs 2022, page 82–93. ACM, March 2022.

[7] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 02 2008.

[8] Wentai Liu, W. Fink, M. Tarbell, and M. Sivaprakasam. Image processing and interface for retinal visual prostheses. In *2005 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2927–2930 Vol. 3, 2005.

[9] H. Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928.

[10] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[11] J.R. Polimeni, M. Balasubramanian, and E.L. Schwartz. Multi-area visuotopic map complexes in macaque striate and extra-striate cortex. *Vision Research*, 46(20):3336–3359, 2006.

[12] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference, 2018.

[13] Lucas Relic, Bowen Zhang, Yi-Lin Tuan, and Michael Beyeler. Deep learning-based perceptual stimulus encoder for bionic vision, 2022.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[15] Maureen van der Grinten, Jaap de Ruyter van Steveninck, Antonio Lozano, Laura Pijnacker, Bodo Rueckauer, Pieter Roelfsema, Marcel van Gerven, Richard van Wezel, Umut Güçlü, and Yağmur Güçlütürk. Towards biologically plausible phosphene simulation for the differentiable optimization of visual cortical prostheses. *eLife*, 13:e85812, feb 2024.

[16] Xiaosha Wang, Yangwen Xu, Yuwei Wang, Yi Zeng, Jiacai Zhang, Zhen-Hua Ling, and Yanchao Bi. Representational similarity analysis reveals task-dependent semantic influence of the visual word form area. *Scientific Reports*, 8, 02 2018.

[17] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[18] Jungsuk Kim Yeonji Oh, Jonggi Hong. Retinal prosthesis edge detection (rped) algorithm: Low-power and improved visual acuity strategy for artificial retinal implants. *PLoS ONE*, 19(6):e0305132, 2024.

[19] Richard A. Young. The gaussian derivative model for spatial vision: I. retinal mechanisms. *Spatial Vision*, 2(4):273 – 293, 1987.

[20] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions, 2016.

[21] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE International Conference on Image Processing*, pages 1473–1476, 2012.

[22] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014.

[23] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

[24] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[25] Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016.