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 Corticali-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. 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. [9]. 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.

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.* [6]. 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

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

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 [9].

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 [9]. 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:

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

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.

4) Amplitude Equalization: Phosphene brightness varies with local electrode density and current spread, causing some regions to appear disproportionately bright or dim. To counteract this and present 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 render the raw phosphene map $I_{\text{raw}}(x, y)$ and measure the peak intensity of each blob,

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

We then adjust amplitudes by minimizing the squared error to a target intensity m^*:

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

Starting from $A_i = 1$, we update each gain by gradient descent with learning rate η and clamp A_i to the interval $[A_{\min}, A_{\max}]$. After a fixed number of iterations, the normalized map

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

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

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 1 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

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

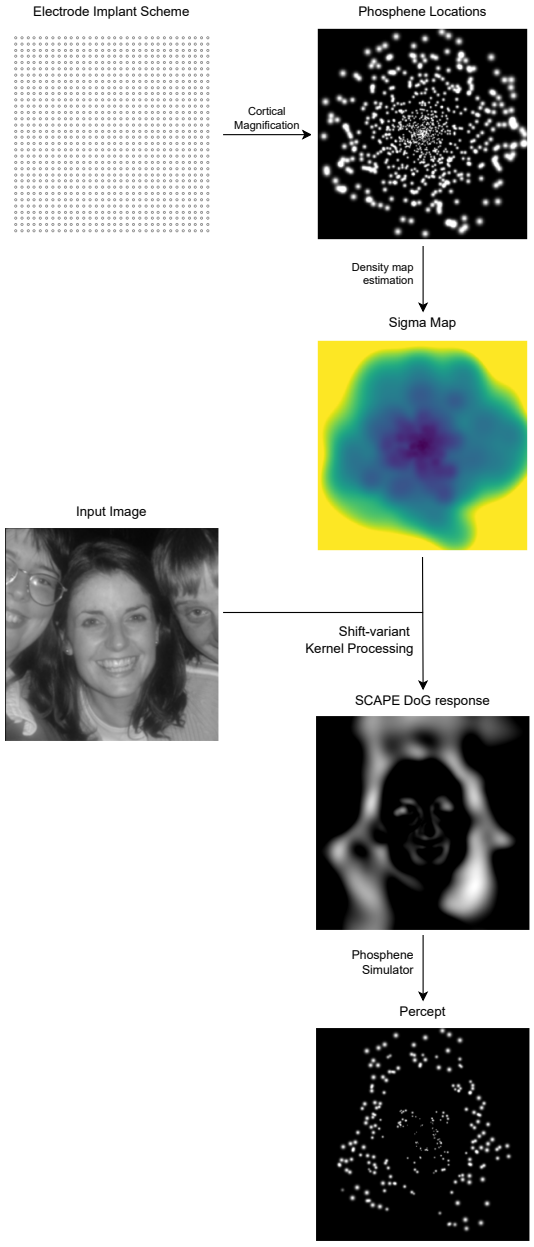

Fig. 1. Overview of the SCAPE pipeline. Starting from a cortical implant layout (top left) we obtain phosphene centers via the simulator, estimate a local density map, convert it into a spatial scale (σ) map, and then apply shift-variant filtering to produce an activation map for phosphene rendering.

(with parameters k, a, b fit to human data [9]) 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

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

Finally we normalize so that

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

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 corresponding spatial scale map $\sigma(x,y)$ that specifies the width of the local filter. By the Nyquist sampling theorem [4], the highest spatial frequency resolvable at (x,y) is

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

since a local density of d phosphenes per unit area implies an average spacing of $\sqrt{1/d}$. To suppress frequencies above f_{Nyq}, we choose a Gaussian filter whose standard deviation satisfies

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

where κ is a tuning constant (typically $\kappa = 1$) that controls the cutoff sharpness. This mapping ensures that filter width increases in regions of low density, reducing visual clutter where detail cannot be resolved, and decreases in regions of high density, preserving fine structure.

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

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

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 [12], 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 [8] with integrated attention and channel-recalibration mechanisms. Key components include:

- **Squeeze-and-Excitation (SE) blocks** to adaptively weight feature channels [2].
- **Dilated residual blocks** in the bottleneck for multi-scale context aggregation [13].
- **Spatial attention gates** on skip connections to emphasize salient regions [5].

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 [11] 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) Visual Saliency Induced Index (VSI): VSI [14] 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:

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

c) Mean Deviation Similarity Index (MDSI): MDSI [15] 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.

d) Perceptual Image-Error Assessment through Pairwise Preference (PIEAPP): PIEAPP [7] 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 [7] 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 [3]. 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 [10].

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.

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

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

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

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.