Hybrid Neural Autoencoder for Sensory Neuroprostheses and Its Applications in Bionic Vision

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sensory neuroprostheses are emerging as a promising technology to restore lost sensory function or augment human capacities. However, sensations elicited by current devices often appear artificial and distorted. Although current models can often predict the neural or perceptual response to an electrical stimulus, an optimal stimulation strategy solves the inverse problem: what is the required stimulus to produce a desired response? Here we frame this as an end-to-end optimization problem, where a deep neural network encoder is trained to invert a known, fixed forward model that approximates the underlying biological system. As a proof of concept, we demonstrate the effectiveness of our hybrid neural autoencoder (HNA) on the use case of visual neuroprostheses. We found that HNA is able to produce high-fidelity stimuli from the MNIST and COCO datasets that outperform conventional encoding strategies and surrogate techniques across all tested conditions. Overall this is an important step towards the long-standing challenge of restoring high-quality vision to people living with incurable blindness and may prove a promising solution for a variety of neuroprosthetic technologies.

1 Introduction

Sensory neuroprostheses are emerging as a promising technology to restore lost sensory function or augment human capacities [9, 16]. In such devices, diminished sensory modalities (e.g., hearing [45], vision [17, 26], cutaneous touch [41]) are re-enacted through streams of artificial input to the nervous system. For example, visual neuroprostheses electrically stimulate neurons in the early visual system to elicit neuronal responses that the brain interprets as visual percepts. Such devices have the potential to restore a rudimentary form of vision to millions of people living with incurable blindness.

However, all of these technologies face the challenge of interfacing with a highly nonlinear system of biological neurons whose role in perception is not fully understood. Due to the limited resolution of electrical stimulation, prostheses often create neural response patterns foreign to the brain. Consequently, sensations elicited by current sensory neuroprostheses often appear artificial and distorted [14, 29]. A major outstanding challenge is thus to identify a stimulus encoding that leads to perceptually intelligible sensations. Often there exists a numeric or symbolic forward model, f (Fig. 1A), constrained by empirical data, that can predict a neuronal or (ideally) perceptual response to the applied stimulus (see [5] for a recent review). To find the stimulus that will elicit a desired response, one essentially needs to find the inverse of the forward model, f^{-1}. However, realistic forward models are rarely analytically invertible, making this a challenging open problem for neuroprostheses.

Here we propose to approach this as an end-to-end optimization problem, where a deep neural network (*encoder*) is trained to invert a known, fixed forward model (*decoder*, Fig. 1B). The encoder is trained to predict the patterns of electrical stimulation patterns that elicit perception (*e.g.,* vision, audition) or neural responses (*e.g.,* firing rates) closest to the target. This hybrid neural autoencoder

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

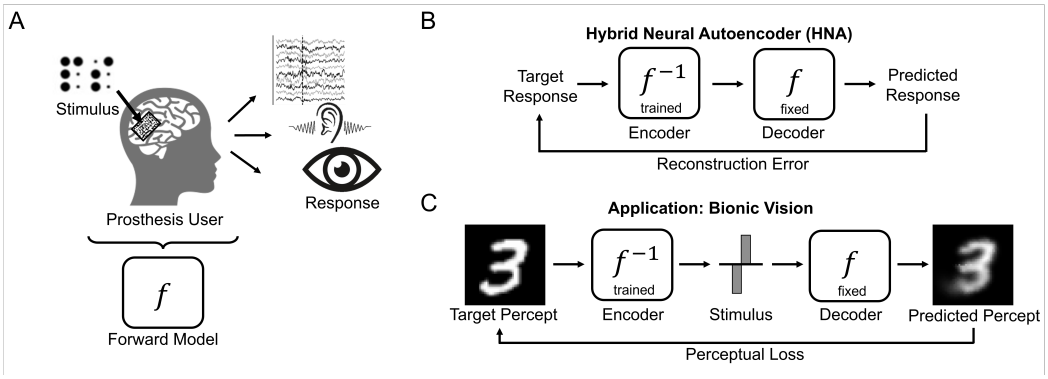

Figure 1: *A)* Sensory neuroprosthesis. A numeric or symbolic forward model (f) is used to approximate the neuronal or, ideally, perceptual response to electrical stimuli. *B)* Hybrid neural autoencoder (HNA). A deep neural encoder (f^{-1}) is trained to predict the patterns of electrical stimulation that elicit responses closest to the target. *C)* Bionic vision. Visual neuroprostheses are one prominent application of HNA, where electrical stimulation is used to elicit a desired visual percept.

(HNA) could in theory be used to optimize stimuli for any open-loop neuroprosthesis with a known forward model that approximates the underlying biological system.

In order to optimize end-to-end, the forward model must be differentiable and computationally efficient. When this is not the case, an alternative approach is to train a surrogate neural network, $\hat{f}$, to approximate the forward model [34, 28, 30, 31]. However, even well-trained surrogates may result in errors when included in our end-to-end framework, due to the encoders' ability to learn to exploit the surrogate model. We therefore also evaluate whether a surrogate approach to HNA is suitable.

To this end, we make the following contributions:

- We propose a hybrid neural autoencoder (HNA) consisting of a deep neural encoder trained to invert a fixed, numerical or symbolic forward model (decoder) to optimize stimulus selection. Our framework is general and addresses an important challenge with sensory neuroprostheses.
- As a proof of concept, we demonstrate the utility of HNA for visual neuroprostheses, where we predict electrode activation patterns required to produce a desired visual percept. We show that the HNA is able to produce high-fidelity, patient-specific stimuli representing handwritten digits and segmented images of everyday objects, drastically outperforming conventional approaches.
- We evaluate replacing a computationally expensive or nondifferentiable forward model with a surrogate, highlighting benefits and potential dangers of popular surrogate techniques.

2 Background

Sensory Neuroprostheses Recent advances in neural understanding, wearable electronics, and biocompatible materials have accelerated the development of sensory neuroprostheses to restore perceptual function to people with impaired sensation. Use of neuroprostheses typically requires invasive implants that elicit neural responses via electrical, magnetic, or optogenetic stimulation. Two of the most promising applications include cochlear implants, which stimulate the auditory nerve to elicit sounds [45], and visual implants (see next subsection) to restore vision to the blind. However, a variety of other devices are in development; for instance, to restore touch [41, 40] or motor function [6]. The latter differ from other sensory neuroprostheses in that they generate stimuli using motor feedback (*closed-loop*) [7, 43]. In the absence of feedback (*open-loop*), a proper stimulus encoding is paramount to the success of these devices.

Restoring Vision to the Blind For millions of people who are living with incurable blindness, a visual prostheses (*bionic eye*, Fig. 2, *left*) may be the only treatment option [17]. Analogous to cochlear implants, these devices electrically stimulate surviving cells in the visual pathway to evoke visual percepts (*phosphenes*), which can support simple behavioral tasks [26, 39, 23].

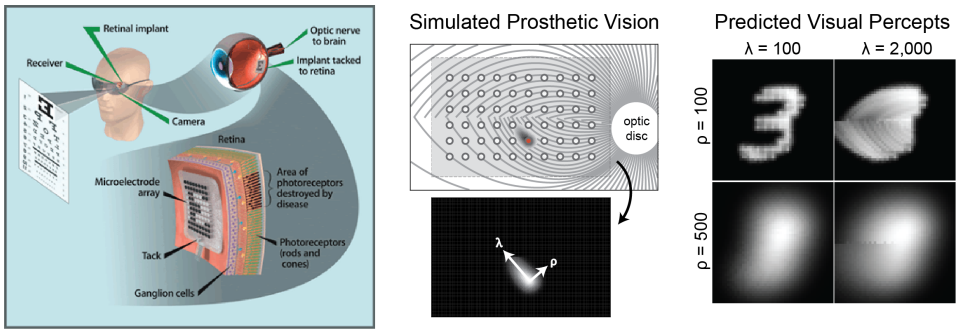

Figure 2: *Left*: Visual prosthesis. Incoming target images are transmitted from a camera to an implant in the retina, which encodes the image as an electrical stimulus pattern. *Center*: Electrical stimulation (red disc) of a nerve fiber bundle (NFB) (gray lines) leads to elongated tissue activation (gray shaded region). The resulting phosphene appears elongated as well; its shape can be described by two parameters, λ (axonal spread) and ρ (radial spread). *Right*: Predicted visual percepts for a particular MNIST digit for varying ρ and λ values.

A common misconception is that each electrode in the microelectrode array can be thought of as a pixel in an image; to generate a complex visual experience, one then simply needs to turn on the right combination of pixels [13]. However, recent evidence suggests that phosphenes often appear distorted (*e.g.*, as simple geometric shapes such as lines, wedges, and blobs) and vary drastically across subjects and electrodes [14, 17].

Phosphene appearance has been best characterized in epiretinal implants, where inadvertent activation of nerve fiber bundles (NFBs) in the optic fiber layer of the retina leads to elongated phosphenes [3, 35] (Fig. 2, *center*). To this end, Granley *et. al* [19] developed a forward model to predict phosphene shape as a function of both neuroanatomical parameters (*i.e.*, retinal location of the stimulating electrode) and stimulus parameters (*i.e.*, pulse frequency, amplitude, and duration). With this model, phosphenes are primarily characterized by two main parameters, ρ and λ, which dictate the size and elongation of the elicited phosphene, respectively (Fig. 2, *right*). Stimulation from multiple electrodes is nonlinearly integrated into a combined perception, and if two electrodes happen to activate the same NFB, they might not generate two distinct phosphenes. In addition, ρ and λ vary drastically across patients [3], making stimulus encoding a particularly challenging problem.

3 Related Work

The conventional 'naive' encoding strategy simply sets the amplitude of each electrode to the brightness of the corresponding pixel in the target image [8, 26, 38], essentially making the stimulus a downsampled version of the target. Although simple, this encoding strategy only works with near-linear forward models and cannot account for phosphene distortions encountered by real prosthesis patients [3, 14]. When used with a more sophisticated forward model, this strategy often produces unrecognizable percepts (Fig. 2, *right*).

To provide an alternative, Shah *et al.* [36] used a greedy approach to iteratively select the stimuli that best recreate a desired neural activity pattern over a given temporal window, assuming that the brain would integrate them into a coherent visual percept. Ghaffari *et al.* [18] used a neural network surrogate model combined with an interior point algorithm to optimize for localized, circular neural activation patterns for individual electrodes. Fauvel *et al.* [15] used human in-the-loop Bayesian optimization to achieve encodings perceptually favored by the subject. Spencer *et al.* [38] proposed framing the problem of stimulus encoding as inversion of a forward model of neural activation patterns. To approximate the inverse, their approach either requires simplification or is NP-hard [38]. Furthermore, it cannot predict the perceptual consequences of the predicted neural activity.

van Steveninck *et al.* [10] proposed an end-to-end optimization strategy with a fixed phosphene model, similar to HNA. However, their approach crucially differs from ours in that they included a secondary deep neural network to post-process the predicted phosphenes. This is a critical limitation, because a low reconstruction loss does not reveal whether a high-fidelity encoder was learned or whether the

secondary decoder network simply learned to correct for the encoder's mistakes. In addition, they used an unrealistic phosphene model that simply upscales and smooths a binary stimulus pattern. It is therefore not clear whether their results would generalize to real visual prosthesis patients.

Relic *et al.* [34] also utilized the end-to-end approach, but without the secondary decoder network used in [10]. They used a more realistic phosphene model, which accounts for some spatial distortions (*e.g.*, axonal streaks), but not the effects of stimulus parameters. Since including a nontrivial phosphene model in the loop is not straightforward, they instead trained a surrogate neural network to approximate the forward model. We re-implemented Relic's 'surrogate' approach in this paper as a baseline method to compare against, as described in Section 4.

Taken together, we identified three main limitations of previous work that this study aims to address:

1) **Unrealistic forward models.** Most previous approaches (*e.g.*, [10, 36, 38]) use an overly simplified forward model that cannot account for empirical data [3, 14]. We overcome this limitation by developing (and inverting) a differentiable version of a neurophysiologically informed and psychophysically validated phosphene model [19] that can account for the effects of stimulus amplitude, frequency, and pulse duration on phosphene appearance.
2) **Optimization of neural responses.** Most previous approaches (*e.g.*, [36, 38]) focus on optimizing neural activation patterns in the retina in response to electrical stimulation ("bottom-up"). However, the visual system undergoes extensive remodeling during blinding diseases such as retinitis pigmentosa [27]. Thus the link between neural activity and visual perception is unclear. We overcome this limitation by inverting a phenomenological ("top-down") model constrained by behavioral data that predicts visual perception directly from electrical stimuli [3, 19].
3) **Objective function** Most previous approaches rely on minimizing mean squared error (MSE) between the target and reconstructed image. Although simple and efficient, MSE is known to be a poor measure of perceptual dissimilarity for images [44] and does not align well with human assessments of image quality [46]. We overcome this limitation by proposing a joint perceptual metric that combines mean absolute error (MAE), VGG, and Laplacian smoothing losses.

4 Methods

Problem Formulation We consider a system where there is some known forward process f mapping stimuli to responses $f : \mathcal{S} \mapsto \mathcal{R}, f(\mathcal{S}) \subset \mathcal{R}$. In the case of visual prostheses, f may map stimuli to visual percepts. In spinal cord rehabilitation devices, f might map electrical stimuli to neural activation patterns. f may be parameterized by patient-specific parameters ϕ.

Finding the best stimulus for an arbitrary target response $\mathbf{t} \in \mathcal{R}$ is therefore equivalent to finding the inverse of f. However, since not every response can be perfectly reproduced by a stimulus, the true inverse of f is not well defined. We therefore seek the pseudoinverse (still denoted as f^{-1} for simplicity) instead, which outputs the stimuli which produces the closest response to $\mathbf{t}$ under some distance metric d:

$$f^{-1}(\mathbf{t}, \phi) \coloneqq \underset{\mathbf{s} \in \mathcal{S}}{\arg\min}\, d(f(\mathbf{s}; \phi), \mathbf{t}). \tag{1}$$

This problem is straightforward to solve using an autoencoder approach, where a learned encoder f^{-1} is trained to invert the fixed decoder f (i.e., forward model).

Encoder We approximate the pseudoinverse f^{-1} with a deep neural network encoder $\hat{f}^{-1}(\mathbf{t}, \phi; \theta)$ with weights θ, trained to minimize the distance d between the target image $\mathbf{t}$ and predicted image $\hat{\mathbf{t}}$:

$$\min_{\theta,\, \phi \sim p(\phi)} d(\mathbf{t}, \hat{\mathbf{t}}) \tag{2}$$

$$\hat{\mathbf{t}} = f(\, \hat{f}^{-1}(\mathbf{t}, \phi;\, \theta);\ \phi), \tag{3}$$

where ϕ is sampled from a uniform random distribution spanning the empirically observed range of patient-specific parameters [3, 19].

We use a simple architecture consisting solely of fully connected (FC) and batch normalization (BN) [21] layers. First, the target $\mathbf{t}$ is flattened and input to one FC layer. In parallel, the patient parameters ϕ are input to a BN layer before two hidden FC layers. Then, the outputs of these

two paths are concatenated, and the combined vector input through two FC layers, producing an intermediate representation $\mathbf{x}$. Amplitudes are predicted from $\mathbf{x}$ with a FC layer. The amplitudes are then concatenated with $\mathbf{x}$, put through a BN layer, and used to predict frequency and pulse duration, each with a FC layer. The encoder outputs are merged into one stimulus matrix $\hat{\mathbf{s}}$. All layers uses a leaky ReLU activation functions, except for stimulus outputs, which use ReLu activation, enforcing nonnegativity. In total there were 1.4M trainable parameters. We also tried using convolutional, transformers, and other attention layers, but the fully connected architecture performed best.

Decoder For our decoder f, we use a reformulated but equivalent version of the model described in [19]. We use a square 15×15 array of 150μm electrodes, spaced 400μm apart and centered on the fovea. The size and scale of this device was motivated by similar designs in real epiretinal implants.

f takes as input a stimulus matrix $\mathbf{s} \in \mathbb{R}_{\geq 0}^{n_e \text{x} 3}$, where the stimulus on each electrode ($\mathbf{s_e}$) is a biphasic pulse train described by its frequency, amplitude, and pulse duration. A simulated map of retinal NFBs is used to predict phosphene shape. Following [3], each retinal ganglion cells' activation is assumed to be the maximum stimulation intensity along its axon. Each axon's sensitivity to stimulation is assumed to decay exponentially with i) Euclidean distance to the stimulating electrode (radial decay rate, ρ) and distance to the soma along the curved axon (axonal decay rate, λ). To account for stimulus properties [19], ρ, λ, and the per-electrode brightness are scaled by three simple equations $F_{\text{size}}(\mathbf{s_e}, \phi)$, $F_{\text{streak}}(\mathbf{s_e}, \phi)$, and $F_{\text{bright}}(\mathbf{s_e}, \phi)$, respectively.

The brightness of a pixel located at the point $\mathbf{x} \in \mathbb{R}^2$ in the output image is given by

$$f(\mathbf{s}; \phi) = \max_{\mathbf{a} \in A} \sum_{e \in E} F_{\text{bright}}(\mathbf{s_e}, \phi) \exp\left(\frac{-||\mathbf{x} - \mathbf{e}||_2^2}{2\rho^2 F_{\text{size}}(\mathbf{s_e}, \phi)} + \frac{-d_s(\mathbf{x}, \mathbf{a})^2}{2\lambda^2 F_{\text{streak}}(\mathbf{s_e}, \phi)} \right) \tag{4}$$

where A is the cells' axon trajectory, E is the set of electrodes, $\phi = \{\rho, \lambda, ...\}$ is a set of 12 patient-specific parameters, and d_s is the path length along the axon trajectory [22]from $\mathbf{a}$ to $\mathbf{x}$:

$$d_s(\mathbf{x}, \mathbf{a}) = \int_{\mathbf{a}}^{\mathbf{x}} \sqrt{A(\theta)^2 + \left(\frac{dA(\theta)}{d\theta} \right)^2} \, d\theta. \tag{5}$$

This model (f) can be fit to individual patients; however, it is computationally slow and not differentiable. For more details on these equations, see [19]. We therefore considered two approaches.

- **Differentiable Model:** We reformulated equations 4 and 5 into an equivalent set of parallelized matrix operations, implemented in Tensorflow [1]. Significant efforts were put towards developing a model optimized for XLA engines on GPU, resulting in speedups of up to 5000x compared to the model as presented in [19], enabling large-scale gradient descent. To enforce differentiability, we numerically approximated d_s using $|A| = 500$ axon segments per axon.
- **Surrogate Model:** We also implemented the surrogate approach from [34] as a baseline method, where f is approximated with another deep neural network $\hat{f}_\phi(\mathbf{s}; \theta_f)$ with weights θ_f. To achieve this we generated 50,000 percepts using randomly selected stimuli passed through f and fit a deep neural network to produce identical images. As f is highly dependent on patient-specific parameters ϕ, we generated new data and fit a separate surrogate model for each ϕ in our experimental set. Specific implementation details of the surrogate are presented in Appendix A. Our implementation improves upon [34] by using the more advanced phosphene model described above, which accounts for effects of stimulus properties and allows for optimization of stimulus pulse duration and frequency, in addition to amplitude.

Metrics To measure perceptual similarity, we use a joint perceptual objective consisting of a VGG [37] perceptual similarity term, a mean absolute error term, and a smoothness regularization term. The mean absolute error term is given by

$$L_{MAE} = ||\mathbf{t} - \hat{\mathbf{t}}||_1. \tag{6}$$

The VGG term aims to capture higher-level, abstract differences between images [10, 24]. Both the target image and the reconstructed phosphene are input to a VGG-19 network pretrained for classification on ImageNet [11], and the MSE between the activations on a downstream convolutional

layer is computed. Let V_l be a function that extracts the activations of the l-th convolutional layer for a given image. The VGG loss is then defined as:

$$L_{VGG} = ||V_l(\mathbf{t}) - V_l(\hat{\mathbf{t}})||_2^2. \tag{7}$$

We also include a smoothing regularization term. This term imposes a loss on the second spatial derivative of the predicted image. A low second derivative implies that where the target image does change, it changes slowly. We found this encouraged smoother, more connected phosphenes. To approximate the second derivative, we convolve the image with a Laplacian filter [32] of size k, denoted by $Lap(\cdot, k)$, and compute the mean. The smoothness loss is given by:

$$L_{Smooth} = \frac{1}{|\hat{\mathbf{t}}|} \sum_i (\frac{\partial^2}{dx^2} \hat{\mathbf{t}})_i = \frac{1}{|\hat{\mathbf{t}}|} \sum_i Lap(\hat{\mathbf{t}}, k)_i. \tag{8}$$

Our final objective is the weighted sum of the three individual losses, given by Eq. 9, where α and β are hyperparameters controlling the relative weighting of the three terms:

$$d = L_{MAE} + \alpha L_{Smooth} + \beta L_{VGG}. \tag{9}$$

We also implement a secondary metric to quantify phosphene recognizability, applicable only for the MNIST reconstruction task. We first pre-train a classifier network on the MNIST targets until it reaches 99% test accuracy, and then fix the weights. The relative accuracy (RA) is then defined as the ratio of the classifiers accuracy on the reconstructed images to its accuracy on the targets $RA = ACC/ACC(\mathbf{t})$. A perfect encoder would result in $RA = 100\%$. A similar process was not possible for the COCO task, due to the possibility of having multiple objects in each target image.

Training/Optimization We trained using Tensorflow 2.7 [1] on a single NVIDIA RTX 3090 with 24GB memory. Stochastic gradient descent with Nesterov momentum was used to minimize the joint perceptual loss. We used a batch size of 16 due to memory limitations imposed by f. The amplitude, frequency predictions are scaled by 2, 20 respectively, while the pulse duration predictions were shifted by 1e-3 prior to being fed through the decoder. This encourages the initial predictions of the network to be in a reasonable range. The Laplacian filter size k is set to 5. We choose l to be first convolutional layer in the last block using cross validation (see Appendix D). Similarly, we perform cross validation to find the best values for α and β. Instead of using one value, we found that incrementally increasing the weighting of the VGG loss (β) from 0 while simultaneously decreasing the initially high weight on the smoothing constraint (α) was crucial for performance, especially when the range of allowed ϕ values was large (see Appendix D).

Datasets We first evaluated on handwritten digits from MNIST [12], enabling comparison to previous works [34]. Images preprocessing consisted of resizing the target images to the same shape as the output of f (49x49). We also evaluate on more realistic images of common objects from the MS-COCO [25] dataset. We selected a subset of 25 of the MS-COCO object categories deemed more likely to be encountered by blind individuals (*e.g.* people, household objects), and use only images that contain at least one instance of these objects. We further filter out images by various other criteria, such as being too cluttered or too dim. This process results in a total of approximately 47K training images and 12K test images. See Appendix C for a full description of the selection process.

Natural images often contain too much detail to be encoded with prosthetic vision. While scene simplification strategies exist [20], we focus on the encoding algorithm, so we simply used the ground-truth segmentation masks to segment out the objects of interest. The images were then converted to grayscale, and resized to (49, 49).

5 Results

5.1 MNIST

The phosphenes produced from the HNA, surrogate, and naive encoders on the MNIST test set are shown in Fig. 3 and performance is summarized in Table 1. For each MNIST sample, the target image is input to the encoder, which predicts a stimulus. The stimulus is fed through the true forward model f, and the predicted phosphene is shown. Since the surrogate method must be retrained for each ϕ, results are only shown for 4 simulated patients. Our proposed approach outperformed the baselines across all metrics (see Appendix B for a comparison of stimuli).

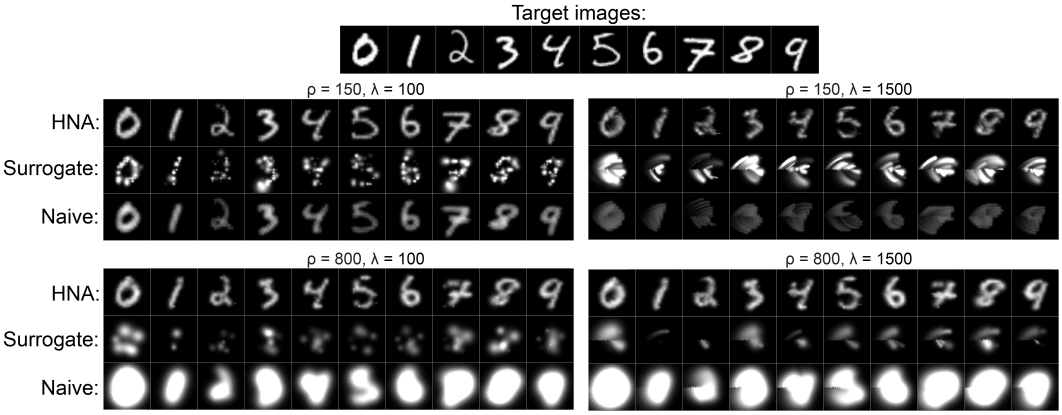

Figure 3: Reconstructed MNIST targets for HNA, surrogate, and naive encoders across 4 specific simulated patients. Note that the brightness of the naive encoder is clipped for display

Table 1: MNIST performance

Encoding	ρ=150 λ=100			ρ=150 λ=1500			ρ=800 λ=100			ρ=800 λ=1500		
	Joint Loss	MAE	RA	Joint Loss	MAE	RA	Joint Loss	MAE	RA	Joint Loss	MAE	RA
Naive	1.161	0.1855	90.3	1.442	0.214	78.1	8.152	1.500	34.8	8.780	1.726	28.8
Surrogate	2.509	0.1351	53.8	3.118	0.2431	30.7	1.692	0.2135	19.9	1.694	0.2237	18.1
HNA	**0.559**	**0.064**	**98.1**	**1.029**	**0.1412**	**89.3**	**0.913**	**0.113**	**95.9**	**0.957**	**0.126**	**94.8**

5.2 COCO

The phosphenes produced by HNA and the naive encoder for the segmented COCO dataset are shown in Fig. 4. We omit the surrogate results due to its poor perceptual performance on MNIST. Averaged across all ϕ, HNA had a joint loss of 0.713 on the test set and MAE of 0.1408, while the naive encoder had a joint loss of 1.873 and MAE of 0.2830.

5.3 Modeling Patient-to-Patient Variations

MNIST encoder performance across simulated patients (ϕ) is shown in Fig. 5. Since the surrogate encoder has to be retrained for each patient, comparison is infeasible. To visualize the effects of changing ρ and λ on the produced phosphenes, Fig. 5A shows the result of encoding two example MNIST digits, both using the naive method and our encoder. As λ increases, the naive phosphenes appear increasingly elongated, and as ρ increases, the phosphenes become increasingly large and blurry. The phosphenes from HNA are slightly too dim and disconnected at low ρ, but are relatively stable across other values of ρ and λ.

To compare performance across the entire dataset, we computed the average test set loss across the same range of ρ and λ (Fig. 5B). The encoder performs well across a wide range of simulated patients, with larger loss only at low ρ. The naive method performs well only on a limited set of ϕ, with small λ and $\rho \approx 200$. The naive loss was higher than the learned encoder at every simulated point. Random sampling of ρ and λ for each image results in a joint loss of 0.921, MAE of 0.120, and RA of 94.0% for HNA, while the naive encoder results in a joint loss of 3.17, MAE of 0.596, and RA of 63.6%. The same analysis yielded similar results on COCO (Appendix F). An analysis across other parameters is presented in Appendix E.

In order for prosthetic vision to be useful, different instances of the same objects would ideally produce similar phosphenes, allowing for consistent perception. To evaluate whether our model achieves this, we cluster the target images and resulting phosphenes using T-SNE [42] shown in Fig. 5C. The ground truth images form clusters corresponding to the digits 0-9. The phosphenes from our encoder roughly form similar, slightly less separated groupings, whereas the naive phosphenes do not. To ensure that this was not the result of bad T-SNE hyperparameters, we repeated the clustering across different perplexities and learning rates, obtaining similar or worse results.

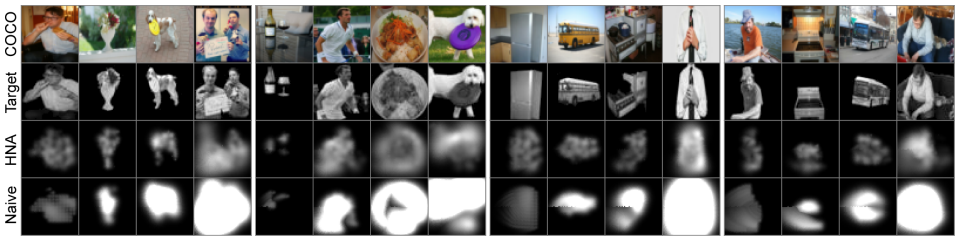

Figure 4: Original (*top row*), segmented (*second row*), and reconstructed targets for the COCO dataset, for both HNA (*third row*) and naive encoders (*bottom row*). Left to right within each block of 4 images, ρ takes values of 200, 400, 600, 800. Left to right across blocks, λ takes values of 250, 750, 1250, 2000. Note that the brightness of the naive method is clipped for display.

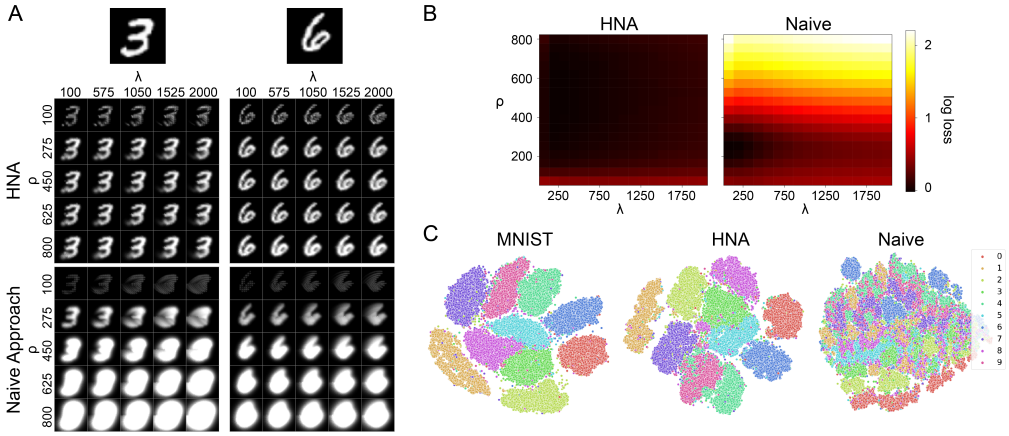

Figure 5: Encoder performance across simulated patients (varying ρ and λ) on the MNIST dataset. ***A***: Target, HNA encoder, and naive encoder phosphenes for two example digits. ***B***: Heatmaps showing the log joint loss across ρ and λ for HNA and naive encoders. ***C***: T-SNE clusterings on original MNIST targets, HNA reconstructed phosphenes, and naive reconstructed phosphenes.

5.4 Joint Perceptual Error Ablation Study

To show that the joint perceptual metric performs better than any of its individual components, we train models using just the VGG loss (Eq. 7), and just MAE (Eq. 6). Shown are values for ρ=150 and λ=600. As mentioned previously, encoders trained using just VGG loss fail to converge, thus we pretrain the VGG encoder using MAE and smoothing loss, then transition to using only VGG. We do not consider ablating the smoothing term (Eq. 8) because it is simply a regularization term. Fig. 6 shows the phosphenes produced by HNA trained on the joint, VGG-only, and MAE-only loss.

The VGG encoder had a test VGG loss of 4% lower than the joint model, and the phosphenes it produces have the correct high level perceptual details, but are oversmoothed and blurry. The MAE encoder had a final test MAE of 9% lower than the joint model. Despite a lower MAE, the reconstructed images appear to be lower quality. The joint model had a RA of 99.0% The VGG encoder had a RA of 95.9%, while the joint model had a RA of 77.6%

6 Discussion

Visual Prostheses We found that HNA is able to produce high-fidelity stimuli from the MNIST and COCO datasets that outperform conventional encoding strategies across all tested conditions. Importantly, HNA produces phosphenes that are consistent across representations of the same object (Fig. 5C), which is critical to allowing prosthesis users to learn to associate certain visual patterns with specific objects. On the MNIST task, HNA produced high quality reconstructions, nearly matching the targets (Figure 3. On the harder COCO task, HNA significantly outperformed the naive encoder,

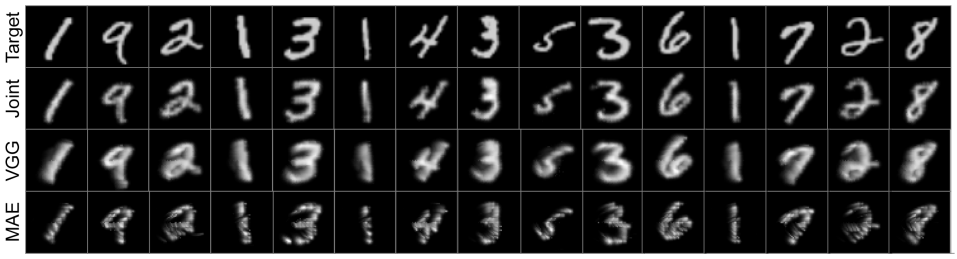

Figure 6: MNIST images for HNA encoders trained using the joint, VGG-only, and MAE-only loss.

but was still unable to capture all of the detail in the images. In Appendix G, we demonstrate that this is mainly due to the limited spatial resolution and not a fundamental limitation of HNA.

Another advantage of the HNA is that it can be trained to predict stimuli across a wide range of patient-specific parameter values ϕ, whereas the conventional naive encoder works well only for small values of ρ and λ. This may be one reason why the naive encoding strategy has been shown to lead to substantial individual differences in visual outcomes [39, 33]. Our results suggest that stimuli produced with HNA may be able to reduce at least some amount of this patient-to-patient variability.

Furthermore, HNA also proved superior to a surrogate forward model. The latter offer an alternative when the forward model is computationally expensive or not differentiable. Understandably, any inaccuracies in the surrogate model will propagate to the learned encoder during training. However, we observed this phenomenon even when the surrogate performed well individually (see Appendix A), suggesting that even for well trained surrogates, the encoder may still learn to exploit the inexact surrogate instead of learning to invert the true model. We suspect that the surrogate method's inferior performance here compared to [34] can be explained by our larger stimulus search space. Thus, we cannot currently suggest HNA for forward models where a surrogate model must be used, unless the forward model is sufficiently simple or has a small stimulus space.

Broader Impacts While our work is presented in the context of visual prostheses, the HNA framework may apply to any sensory neuroprosthesis where stimulus selection can be informed by a numeric or symbolic forward model. For example, cochlear implants [45] could use HNA to choose stimuli that result in a desired sound, and spinal cord implants [6] could use HNA to find the best way to relay neural signals through a damaged section of the spinal cord. However, HNA might not apply to all neural interfaces, such as systems without a clear neural or perceptual target (*e.g.*, deep brain stimulation for the treatment of Parkinson's [2]) or closed-loop systems [4, 7].

Limitations Our end-to-end architecture allows for exact optimization of the forward model, but it also imposes limitations on the stability and speed of training. For example, we observed that convergence of the encoder required customized learning rate scheduling, careful tuning of metric hyperparameters, and modification of the encoder output layer initialization. Additionally, including the forward model in-the-loop required large amounts of memory and compute time. Notably, these limitations are only applicable during training. Although this may be application-specific, we expect any complex forward model to encounter similar issues, hindering ease of implementation.

Safety Concerns The use of a deep neural network for stimulus encoding does raise safety concerns. Since we cannot examine the process by which stimuli are chosen, it is possible that an HNA encoder might produce potentially harmful stimuli that could lead to serious adverse events (*e.g.*, seizures). However, this concern is mitigated by the fact that most current neuroprostheses are equipped with firmware responsible for verifying that all stimuli stay within FDA-approved safety limits.

7 Conclusion

In summary, this paper proposes a hybrid autoencoder structure as a general framework for stimulus optimization in sensory neuroprostheses and, as a proof of concept, demonstrates its utility on the prominent example of visual neuroprostheses, drastically outperforming conventional encoding strategies. This may prove a promising solution for a variety of neuroprosthetic technologies.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] Alim Louis Benabid. Deep brain stimulation for Parkinson's disease. *Current Opinion in Neurobiology*, 13(6):696–706, December 2003. ISSN 0959-4388. doi: 10.1016/j.conb.2003.11.001. URL `https://www.sciencedirect.com/science/article/pii/S0959438803001739`.

[3] Michael Beyeler, Devyani Nanduri, James D. Weiland, Ariel Rokem, Geoffrey M. Boynton, and Ione Fine. A model of ganglion axon pathways accounts for percepts elicited by retinal implants. *Scientific Reports*, 9(1):1–16, June 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-45416-4. URL `https://www.nature.com/articles/s41598-019-45416-4`.

[4] Bardia Bozorgzadeh, Douglas R. Schuweiler, Martin J. Bobak, Paul A. Garris, and Pedram Mohseni. Neurochemostat: A Neural Interface SoC With Integrated Chemometrics for Closed-Loop Regulation of Brain Dopamine. *IEEE Transactions on Biomedical Circuits and Systems*, 10(3):654–667, June 2016. ISSN 1940-9990. doi: 10.1109/TBCAS.2015.2453791. Conference Name: IEEE Transactions on Biomedical Circuits and Systems.

[5] Bingni W. Brunton and Michael Beyeler. Data-driven models in human neuroscience and neuroengineering. *Current Opinion in Neurobiology*, 58:21–29, October 2019. ISSN 0959-4388. doi: 10.1016/j.conb.2019.06.008. URL `http://www.sciencedirect.com/science/article/pii/S0959438818302502`.

[6] Marco Capogrosso, Tomislav Milekovic, David Borton, Fabien Wagner, Eduardo Martin Moraud, Jean-Baptiste Mignardot, Nicolas Buse, Jerome Gandar, Quentin Barraud, David Xing, Elodie Rey, Simone Duis, Yang Jianzhong, Wai Kin D. Ko, Qin Li, Peter Detemple, Tim Denison, Silvestro Micera, Erwan Bezard, Jocelyne Bloch, and Grégoire Courtine. A brain–spine interface alleviating gait deficits after spinal cord injury in primates. *Nature*, 539(7628):284–288, November 2016. ISSN 1476-4687. doi: 10.1038/nature20118. URL `https://www.nature.com/articles/nature20118`.

[7] Christopher A. R. Chapman, Noah Goshi, and Erkin Seker. Multifunctional Neural Interfaces for Closed-Loop Control of Neural Activity. *Advanced Functional Materials*, 28(12):1703523, 2018. ISSN 1616-3028. doi: 10.1002/adfm.201703523. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201703523`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adfm.201703523.

[8] S. C. Chen, G. J. Suaning, J. W. Morley, and N. H. Lovell. Simulating prosthetic vision: I. Visual models of phosphenes. *Vision Research*, 49(12):1493–506, June 2009. ISSN 1878-5646 (Electronic) 0042-6989 (Linking).

[9] Caterina Cinel, Davide Valeriani, and Riccardo Poli. Neurotechnologies for Human Cognitive Augmentation: Current State of the Art and Future Prospects. *Frontiers in Human Neuroscience*, 13:13, January 2019. ISSN 1662-5161. doi: 10.3389/fnhum.2019.00013. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6365771/`.

[10] Jaap de Ruyter van Steveninck, Umut Güçlü, Richard van Wezel, and Marcel van Gerven. End-to-end optimization of prosthetic vision. *Journal of Vision*, 22(2):20, February 2022. ISSN 1534-7362. doi: 10.1167/jov.22.2.20. URL `https://doi.org/10.1167/jov.22.2.20`.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

[12] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[13] Wm H. Dobelle. Artificial Vision for the Blind by Connecting a Television Camera to the Visual Cortex. *ASAIO Journal*, 46(1):3–9, February 2000. ISSN 1058-2916. URL `https://journals.lww.com/asaiojournal/Fulltext/2000/01000/Artificial_Vision_for_the_Blind_by_Connecting_a.2.aspx`.

[14] Cordelia Erickson-Davis and Helma Korzybska. What do blind people "see" with retinal prostheses? Observations and qualitative reports of epiretinal implant users. *PLOS ONE*, 16(2):e0229189, February 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0229189. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229189. Publisher: Public Library of Science.

[15] Tristan Fauvel and Matthew Chalk. Human-in-the-loop optimization of visual prosthetic stimulation. preprint, Neuroscience, November 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.11.24.469867.

[16] Eduardo Fernandez. Development of visual Neuroprostheses: trends and challenges. *Bioelectronic Medicine*, 4(1):12, August 2018. ISSN 2332-8886. doi: 10.1186/s42234-018-0013-8. URL https://doi.org/10.1186/s42234-018-0013-8.

[17] Eduardo Fernández, Arantxa Alfaro, Cristina Soto-Sánchez, Pablo Gonzalez-Lopez, Antonio M. Lozano, Sebastian Peña, Maria Dolores Grima, Alfonso Rodil, Bernardeta Gómez, Xing Chen, Pieter R. Roelfsema, John D. Rolston, Tyler S. Davis, and Richard A. Normann. Visual percepts evoked with an intracortical 96-channel microelectrode array inserted in human occipital cortex. *Journal of Clinical Investigation*, 131 (23):e151331, December 2021. ISSN 1558-8238. doi: 10.1172/JCI151331. URL https://www.jci.org/articles/view/151331.

[18] Dorsa Haji Ghaffari, Yao-Chuan Chang, Ehsan Mirzakhalili, and James D. Weiland. Closed-loop Optimization of Retinal Ganglion Cell Responses to Epiretinal Stimulation: A Computational Study. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 597–600, May 2021. doi: 10.1109/NER49283.2021.9441437. ISSN: 1948-3554.

[19] Jacob Granley and Michael Beyeler. A Computational Model of Phosphene Appearance for Epiretinal Prostheses. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 4477–4481, November 2021. doi: 10.1109/EMBC46164.2021.9629663. ISSN: 2694-0604.

[20] Nicole Han, Sudhanshu Srivastava, Aiwen Xu, Devi Klein, and Michael Beyeler. Deep Learning–Based Scene Simplification for Bionic Vision. In *Augmented Humans Conference 2021*, AHs'21, pages 45–54, New York, NY, USA, February 2021. Association for Computing Machinery. ISBN 978-1-4503-8428-5. doi: 10.1145/3458709.3458982. URL https://doi.org/10.1145/3458709.3458982.

[21] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Technical Report arXiv:1502.03167, arXiv, March 2015. URL http://arxiv.org/abs/1502.03167. arXiv:1502.03167 [cs] type: article.

[22] N. M. Jansonius, J. Nevalainen, B. Selig, L. M. Zangwill, P. A. Sample, W. M. Budde, J. B. Jonas, W. A. Lagrèze, P. J. Airaksinen, R. Vonthein, L. A. Levin, J. Paetzold, and U. Schiefer. A mathematical description of nerve fiber bundle trajectories and their variability in the human retina. *Vision Research*, 49(17):2157–2163, August 2009. ISSN 0042-6989. doi: 10.1016/j.visres.2009.04.029. URL http://www.sciencedirect.com/science/article/pii/S0042698909002922.

[23] Lewis Karapanos, Carla J. Abbott, Lauren N. Ayton, Maria Kolic, Myra B. McGuinness, Elizabeth K. Baglin, Samuel A. Titchener, Jessica Kvansakul, Dean Johnson, William G. Kentler, Nick Barnes, David A. X. Nayagam, Penelope J. Allen, and Matthew A. Petoe. Functional Vision in the Real-World Environment With a Second-Generation (44-Channel) Suprachoroidal Retinal Prosthesis. *Translational Vision Science & Technology*, 10(10):7–7, August 2021. ISSN 2164-2591. doi: 10.1167/tvst.10.10.7. URL https://tvst.arvojournals.org/article.aspx?articleid=2776583. Publisher: The Association for Research in Vision and Ophthalmology.

[24] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal Style Transfer via Feature Transforms. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/49182f81e6a13cf5eaa496d51fea6406-Abstract.html.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. Technical Report arXiv:1405.0312, arXiv, February 2015. URL http://arxiv.org/abs/1405.0312. arXiv:1405.0312 [cs] type: article.

[26] Yvonne Hsu-Lin Luo and Lyndon da Cruz. The Argus® II Retinal Prosthesis System. *Progress in Retinal and Eye Research*, 50:89–107, January 2016. ISSN 1350-9462. doi: 10.1016/j.preteyeres.2015.09.003. URL https://www.sciencedirect.com/science/article/pii/S1350946215000701.

[27] Robert E Marc, Bryan W Jones, Carl B Watt, and Enrica Strettoi. Neural remodeling in retinal degeneration. *Progress in Retinal and Eye Research*, 22(5):607–655, 2003. ISSN 1350-9462. doi: https://doi.org/10.1016/S1350-9462(03)00039-9. URL https://www.sciencedirect.com/science/article/pii/S1350946203000399.

[28] David Montes de Oca Zapiain, James A. Stewart, and Rémi Dingreville. Accelerating phase-field-based microstructure evolution predictions via surrogate models trained by machine learning methods. *npj Computational Materials*, 7(1):1–11, January 2021. ISSN 2057-3960. doi: 10.1038/s41524-020-00471-8. URL https://www.nature.com/articles/s41524-020-00471-8. Number: 1 Publisher: Nature Publishing Group.

[29] Craig D. Murray. Embodiment and Prosthetics. In Pamela Gallagher, Deirdre Desmond, and Malcolm MacLachlan, editors, *Psychoprosthetics*, pages 119–129. Springer, London, 2008. ISBN 978-1-84628-980-4. doi: 10.1007/978-1-84628-980-4_9. URL https://doi.org/10.1007/978-1-84628-980-4_9.

[30] Mohammad Amin Nabian and Hadi Meidani. A Deep Neural Network Surrogate for High-Dimensional Random Partial Differential Equations. *Probabilistic Engineering Mechanics*, 57:14–25, July 2019. ISSN 02668920. doi: 10.1016/j.probengmech.2019.05.001. URL http://arxiv.org/abs/1806.02957. arXiv:1806.02957 [physics, stat].

[31] Stefanos Nikolopoulos, Ioannis Kalogeris, and Vissarion Papadopoulos. Non-intrusive surrogate modeling for parametrized time-dependent partial differential equations using convolutional autoencoders. *Engineering Applications of Artificial Intelligence*, 109:104652, March 2022. ISSN 0952-1976. doi: 10.1016/j.engappai.2021.104652. URL https://www.sciencedirect.com/science/article/pii/S0952197621004541.

[32] Sylvain Paris, Samuel W Hasinoff, and Jan Kautz. Local Laplacian Filters: Edge-aware Image Processing with a Laplacian Pyramid. page 11.

[33] Eli Peli. Testing Vision Is Not Testing For Vision. *Translational Vision Science & Technology*, 9(13):32–32, December 2020. ISSN 2164-2591. doi: 10.1167/tvst.9.13.32. URL https://tvst.arvojournals.org/article.aspx?articleid=2772087. Publisher: The Association for Research in Vision and Ophthalmology.

[34] Lucas Relic, Bowen Zhang, Yi-Lin Tuan, and Michael Beyeler. Deep Learning–Based Perceptual Stimulus Encoder for Bionic Vision. In *Augmented Humans 2022*, AHs 2022, pages 323–325, New York, NY, USA, March 2022. Association for Computing Machinery. ISBN 978-1-4503-9632-5. doi: 10.1145/3519391.3524034. URL https://doi.org/10.1145/3519391.3524034.

[35] J. F. Rizzo, J. Wyatt, J. Loewenstein, S. Kelly, and D. Shire. Perceptual efficacy of electrical stimulation of human retina with a microelectrode array during short-term surgical trials. *Invest Ophthalmol Vis Sci*, 44 (12):5362–9, December 2003. ISSN 0146-0404 (Print) 0146-0404 (Linking).

[36] Nishal P. Shah, Sasidhar Madugula, Lauren Grosberg, Gonzalo Mena, Pulkit Tandon, Pawel Hottowy, Alexander Sher, Alan Litke, Subhasish Mitra, and E.J. Chichilnisky. Optimization of Electrical Stimulation for a High-Fidelity Artificial Retina. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 714–718, March 2019. doi: 10.1109/NER.2019.8716987. ISSN: 1948-3554.

[37] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. Technical Report arXiv:1409.1556, arXiv, April 2015. URL http://arxiv.org/abs/1409.1556. arXiv:1409.1556 [cs] type: article.

[38] Martin J. Spencer, Tatiana Kameneva, David B. Grayden, Hamish Meffin, and Anthony N. Burkitt. Global activity shaping strategies for a retinal implant. *Journal of Neural Engineering*, 16(2):026008, January 2019. ISSN 1741-2552. doi: 10.1088/1741-2552/aaf071. URL https://doi.org/10.1088/1741-2552/aaf071. Publisher: IOP Publishing.

[39] Katarina Stingl, Ruth Schippert, Karl U. Bartz-Schmidt, Dorothea Besch, Charles L. Cottriall, Thomas L. Edwards, Florian Gekeler, Udo Greppmaier, Katja Kiel, Assen Koitschev, Laura Kühlewein, Robert E. MacLaren, James D. Ramsden, Johann Roider, Albrecht Rothermel, Helmut Sachs, Greta S. Schröder, Jan Tode, Nicole Troelenberg, and Eberhart Zrenner. Interim Results of a Multicenter Trial with the New Electronic Subretinal Implant Alpha AMS in 15 Patients Blind from Inherited Retinal Degenerations. *Frontiers in Neuroscience*, 11, 2017. ISSN 1662-453X. doi: 10.3389/fnins.2017.00445. URL https://www.frontiersin.org/articles/10.3389/fnins.2017.00445/full. Publisher: Frontiers.

[40] Gregg A. Tabot, John F. Dammann, Joshua A. Berg, Francesco V. Tenore, Jessica L. Boback, R. Jacob Vogelstein, and Sliman J. Bensmaia. Restoring the sense of touch with a prosthetic hand through a brain interface. *Proceedings of the National Academy of Sciences*, 110(45):18279–18284, November 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1221113110. URL https://pnas.org/doi/full/10.1073/pnas.1221113110.

[41] Daniel W. Tan, Matthew A. Schiefer, Michael W. Keith, James Robert Anderson, Joyce Tyler, and Dustin J. Tyler. A neural interface provides long-term stable natural touch perception. *Science Translational Medicine*, 6(257):257ra138–257ra138, October 2014. doi: 10.1126/scitranslmed.3008669. URL https://www.science.org/doi/10.1126/scitranslmed.3008669. Publisher: American Association for the Advancement of Science.

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] Fabien B. Wagner, Jean-Baptiste Mignardot, Camille G. Le Goff-Mignardot, Robin Demesmaeker, Salif Komi, Marco Capogrosso, Andreas Rowald, Ismael Seáñez, Miroslav Caban, Elvira Pirondini, Molywan Vat, Laura A. McCracken, Roman Heimgartner, Isabelle Fodor, Anne Watrin, Perrine Seguin, Edoardo Paoles, Katrien Van Den Keybus, Grégoire Eberle, Brigitte Schurch, Etienne Pralong, Fabio Becce, John Prior, Nicholas Buse, Rik Buschman, Esra Neufeld, Niels Kuster, Stefano Carda, Joachim von Zitzewitz, Vincent Delattre, Tim Denison, Hendrik Lambert, Karen Minassian, Jocelyne Bloch, and Grégoire Courtine. Targeted neurotechnology restores walking in humans with spinal cord injury. *Nature*, 563(7729):65–71, November 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0649-2. URL https://www.nature.com/articles/s41586-018-0649-2.

[44] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Processing Magazine*, 26(1):98–117, January 2009. ISSN 1558-0792. doi: 10.1109/MSP.2008.930649. Conference Name: IEEE Signal Processing Magazine.

[45] Blake S. Wilson, Charles C. Finley, Dewey T. Lawson, Robert D. Wolford, Donald K. Eddington, and William M. Rabinowitz. Better speech recognition with cochlear implants. *Nature*, 352(6332):236–238, July 1991. ISSN 1476-4687. doi: 10.1038/352236a0. URL https://www.nature.com/articles/352236a0. Number: 6332 Publisher: Nature Publishing Group.

[46] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63(11):211301, November 2020. ISSN 1674-733X, 1869-1919. doi: 10.1007/s11432-019-2757-1. URL https://link.springer.com/10.1007/s11432-019-2757-1.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] we highlight both the limitations of our general framework and our specific proof of concept, see Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6, last paragraph
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] We do not present any theoretical results
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include the trained models, and code to reproduce the main result figures in the supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] detailed instructions are given in 4, training paragraph on hyperparameter selection, learning rates, and many other training details. Data splits were predetermined by the datasets (MNIST and COCO already have splits), with some more details given in Appendix C
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We do not train our models multiple times. Once trained, the networks produce consistent results.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This is specified in Section 4, training paragraph
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We used models from [19, 3, 34], which are all cited whenever used in the paper.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] , we provide saved weights for our final trained models, along with code to display results shown on main results figures.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We do not use any new personal data

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data used does not contain any identifying information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

See supplementary materials for appendix