Transformer-based encoder for improving perception in visual prosthesis

Julia Tomas-Barba Alejandro Perez-Yus Jesus Bermudez-Cameo
Jose J. Guerrero
Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain
{j.tomas,alperez,bermudez,josechu.guerrero}@unizar.es

Abstract

Visual prostheses have the potential to restore sight to the visually impaired, but current devices often deliver suboptimal visual experiences due to physical limitations and spatial and temporal distortions. To address these challenges, recent approaches have employed deep learning algorithms and computational models to optimize stimulation strategies. Additionally, some studies integrate patient-specific information to personalize the visual experience. Our work introduces a novel neural network architecture that incorporates a vision transformer to analyze both visual input and patient-specific parameters, aiming to reduce these distortions. Results demonstrate that Vision Transformers enhance patient perception on the MNIST dataset compared to previous methods and effectively handle more complex images from the CIFAR-10 dataset. These findings suggest that our approach could advance visual prosthetic technology, providing more effective and personalized solutions for visual restoration.

***Keywords*—** Computer Vision, Simulated Prosthetic Vision, Visually impaired assistance

1. Introduction

Visual impairment affects millions of individuals worldwide, significantly impacting their quality of life and ability to perform everyday tasks. Among the various causes of visual impairment, degenerative retinal diseases, such as retinitis pigmentosa and age-related macular degeneration, are some of the most prevalent [18]. These conditions gradually destroy the photoreceptor cells in the retina, leading to progressive vision loss and, in severe cases, complete blindness. In response to this growing health challenge, retinal prostheses have emerged as a promising solution to restore vision for individuals with such diseases. These advanced devices aim to replace the function of lost photoreceptors by delivering targeted electrical stimulation to the remaining retinal cells through electrodes implanted in the retina. This stimulation generates neural signals that the brain interprets as visual information, known as phosphenes.

Despite their potential, retinal prostheses face several challenges that must be addressed to improve their effectiveness and usability. First, technical limitations, such as restricted field of view (FOV) and low resolution complicate the visual experience [14]. To mitigate these issues, computer vision algorithms are developed to analyze visual information and highlight relevant information for the user, tailored to specific tasks, such as face recognition, object detection, or navigation [12, 17]. Secondly, users in clinical trials have described phosphenes as elongated shapes [2] with delayed onset and offset [6], which hinder environmental recognition, especially in dynamic scenes where quick and accurate perception is crucial. Additionally, all these effects vary across patients, highlighting the need for personalized approaches in designing retinal prostheses to optimize visual perception [8, 13]. In this context, two solutions have been proposed. On the one hand, some research focuses on electric pulse parameters [11] and precision in cell stimulation [10]. On the other hand, the combination of realistic models with deep learning algorithms is being used to mitigate both spatial and temporal distortions by optimizing stimulation strategies [3, 15, 19, 20].

Our current and ongoing work aligns with the latter approach, addressing variations across users. The goal is to predict the individual electrode stimulation signal in order to produce a visual perception that resembles the target image. Therefore, the deep network model is an autoencoder, whose encoder part predicts the input electrode stimuli from a given target image (*i.e*. the frequency, amplitude, and pulse duration of each electrode signal), while the decoder transforms such stimuli to a *percept*: a visual representation of what a real patient would perceive, which the network outputs as an image.

As previous studies show [1, 13] the elongations and deformations of the phosphenes depends not only on the stimulation signal, but also on some case-specific parameters related to the implant (*e.g*. location in the retina, size, electrodes positions) and the patient (*e.g*. body response to electrode stimulation, nerve fiber bundles). In our approach we use the parametrization and decoder network proposal from [7], which uses a physically-validated model with 13 parameters. Here, we keep the decoder fixed following the model and assume those parameters are an additional input to the network, although they could be estimated effectively with Bayesian approaches [5, 7].

Previous work has utilized fully-connected networks (FCNs) to predict stimuli from target images and parameter vectors [7]. However, FCNs are limited by their inefficiency with image data, as flattening images results in the loss of spatial relationships

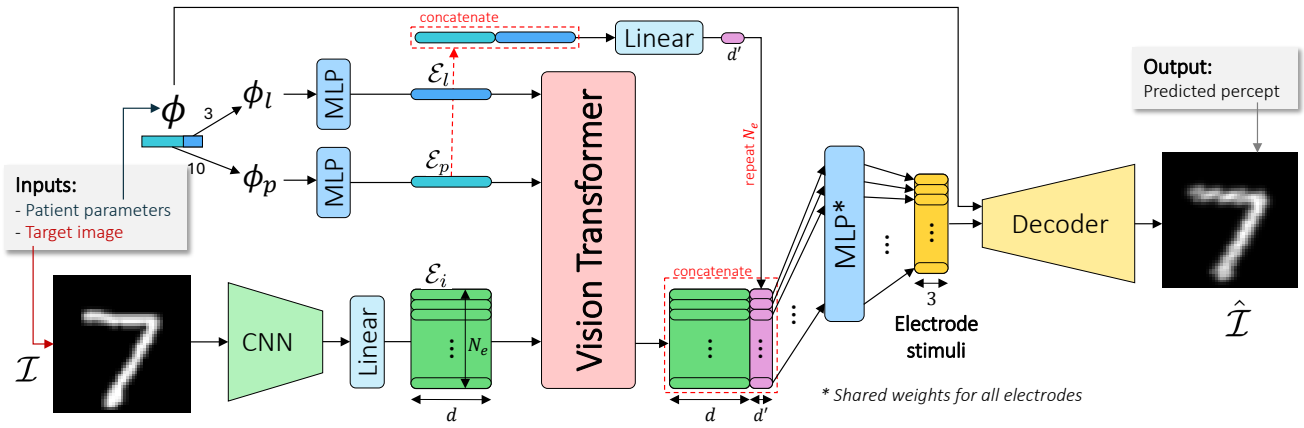

Figure 1. Schematic of the autoencoder architecture.

within the image. In addition, they can be computationally expensive and challenging to scale with high-dimensional, multi-modal data (*e.g.* images and parameter vectors). To address these limitations, we propose a transformer-based architecture [4, 16] as an encoder model for estimating electrode stimulation signals in prosthetic vision, which better preserves spatial relationships and manages complex multi-modal data. We introduce a new architecture tailored for visual prosthesis where each phosphene and parameter vector is represented by unique tokens within the transformer model, enabling it to dynamically adjust focus and improve learning from diverse data inputs. This helps the transformer to learn how the patient parameters interact with the phosphenes, and how the phosphenes interact with each other. Additionally, our method is well-positioned for future expansion to video data, since transformers excel at capturing long-range dependencies in sequential data, which will help to address temporal aspects of prosthetic vision. In this extended abstract, we outline our approach and present preliminary results demonstrating its effectiveness across different datasets.

2. Methods

In this section we are going to describe our proposed deep neural autoencoder. It consists of two main parts: the *encoder*, that produces the parameters of stimulation of each electrode from a given target image and the patient parameters, and the *decoder*, which takes the predicted stimuli and the patient parameters and produce a visual representation that resembles what a real patient would perceive. As stated above, the decoder is based on the analytical forward model from [7], and thus no parameters are trained or optimized. Therefore, our main contribution lies on the encoder part of the network. Inspired by [7], since their proposed forward model is neither linear nor invertible, our approach consists in training a neural network model to optimize the inverse model of the decoder. In Fig. 1 there is a diagram of our model to follow this section.

Our encoder takes as input a target image $\mathcal{I}$, of size $w_i \times h_i$ (note that we use grayscale input images since the produced percept is colorless), and the patient parameters vector ϕ, of 13 elements, and introduces them in our transformer-based encoder. To combine both modalities (image and vector), it is necessary to turn $\mathcal{I}$ and ϕ into token embeddings of hidden dimension d.

In our proposed design, the number of image tokens from $\mathcal{I}$ coincides with the number of electrodes in the prosthesis. Let us assume the size of the electrode grid is $w_e \times h_e$. Then, the number of image tokens will be $w_e \times h_e = N_e$. Our same model would work if we use another visual prosthesis: just the number of tokens in the input sequence will change, making our approach straightforwardly adaptable to different configurations. To obtain the token embeddings, we pass $\mathcal{I}$ through a trainable CNN with convolutional and pooling layers so that it is downsampled to $w_e \times h_e$. Then, the feature vector goes through a linear layer to obtain the input image embeddings tensor, $\mathcal{E}_i$, of shape $N_e \times d$.

The patient vector ϕ is converted to additional token embeddings that will be introduced to the transformers as class tokens. The number of parameter token embeddings will be two: one for the three parameters related to the location of the implant in the retina (ϕ_l), and other for the rest of patient-specific parameters (ϕ_p). This allows the model to pay special attention to the location of the phosphenes in the image and disambiguate this with parameters more related to the response to stimulation and collateral distortion effects. Thus, we split the parameter vector ϕ in ϕ_l and ϕ_p and pass both of them through several fully connected layers of higher dimensions to finally output two vectors of dimension d: $\mathcal{E}_l$ and $\mathcal{E}_p$ respectively.

Our transformer block is a Vision Transformer (ViT) [4], whose inputs are $\mathcal{E}_l$, $\mathcal{E}_p$ and $\mathcal{E}_i$. We take each of the output embeddings corresponding to the image tokens (and therefore, phosphenes) and pass them through a multi-layer perceptron (MLP) to finally obtain a 3-dimensional vector for each phosphene that corresponds to the pulse duration, frequency, and amplitude of the stimulus, respectively. Those stimuli are passed through the decoder, along with ϕ, producing the output percept $\hat{\mathcal{I}}$.

To train this model, we propose using a reconstruction loss defined as follows:

$$\mathcal{L}_{\text{reconstruction}} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\mathcal{I}}_i^* - \mathcal{I}_i^* \right)^2 \tag{1}$$

where $\mathcal{I}^*$ is the target image normalized by its average value,

$\hat{\mathcal{I}}^*$ is the predicted image normalized by its average value, and N is the number of pixels in the images. This loss function employs normalized images to emphasize the relative intensities between the target ($\mathcal{I}^*$) and predicted ($\hat{\mathcal{I}}^*$) images, rather than focusing solely on individual pixel values.

3. Results

In this section we show some experimental results in different datasets to show the effectiveness of our approach. For this study, replicating the experimental setup from [7], we conducted experiments using a visual prosthesis model with 225 phosphenes arranged in a 15×15 rectangular grid. Each phosphene had a radius of 75μm and was spaced 400μm apart. The model was evaluated using two datasets: MNIST, which served as a baseline for comparing our results with those in [7], and CIFAR-10 [9], a more complex and diverse dataset that better simulates the variety of visual experiences a prosthesis user might encounter in real-world scenarios.

In both datasets, input images are grayscaled and resized to 30×30, before being introduced to the CNN. The CNN has four convolutional layers followed by a max pooling layer, so that it outputs a feature vector of $15 \times 15 \times 64$. This vector is reshaped to 225×64 and passed through a linear layer to obtain the final desired shape of $\mathcal{E}_i$ ($225 \times d$). The parameter vectors ϕ_l and ϕ_p are processed through four fully connected layers, each with 1024 neurons, ultimately yielding d-dimensional embedding vectors $\mathcal{E}_l$ and $\mathcal{E}_p$.

Next, the reduced image is fed into a ViT, where the patient-specific parameter vectors are incorporated as class tokens. We experimentally choose a hidden dimension of $d = 128$. The output of the transformer retains a 128-dimensional hidden representation per phosphene. Subsequently, the patient parameter embeddings $\mathcal{E}_l$ and $\mathcal{E}_p$ are concatenated into a single 256-dimensional vector, which is then reduced to $d' = 16$ dimensions through a linear layer. The 16-dimensional vector is concatenated with each of the 225 (15×15) phosphene representations from the output of the ViT as shown in Fig. 1, effectively reintroducing patient-specific information at this stage. Finally, a multi-layer perceptron (MLP) is applied, reducing the $128 + 16$ channels to 3 output channels per phosphene, corresponding to the stimulation parameters.

For the MNIST dataset, we trained the encoder using the normalized L2 loss between the decoder output and the target image as the objective function. The encoder was trained for 20 epochs with a batch size of 16, using a cosine learning rate scheduler with initial learning rate of 1e-3. This training approach resulted in an L2 loss value of 0.032, representing an improvement over the previously reported perceptual loss of 0.05 in [7]. Additionally, we evaluated our trained encoder using a pretrained MNIST classifier with 98.65% accuracy. Our encoder achieved an accuracy of 97.45%, surpassing the previous model, which had an accuracy of 95.6%.

For the CIFAR-10 dataset, we employed the same architecture, but used different training parameters. The encoder was trained for 100 epochs with a batch size of 64 and a fixed learning rate of 2e-4. Although we do not have a direct comparison, the normalized L2 loss achieved was 0.033. In Fig. 2 there are some qualitative examples from both datasets.

4. Discussion

Our experiments highlight the significant potential of Vision Transformer (ViT) architectures in enhancing image processing for prosthetic vision. By effectively combining spatial information with patient-specific parameters, our approach has shown notable advancements in image reconstruction. Specifically, we achieved an L2 loss of 0.032 on the MNIST dataset, outperforming previous methods. Moreover, our model has demonstrated its capability to handle images from the CIFAR-10 dataset.

However, the model produced blurred reconstructions for complex images from the CIFAR-10 dataset. This suggests that future training should focus more on task-specific optimization to better align with user requirements. Additionally, incorporating image pre-segmentation techniques could potentially enhance the reconstruction fidelity of target images.

Future work will focus on addressing model limitations and implementing Bayesian optimization to refine hyperparameters and enhance calibration processes. Additionally, we will conduct experiments to evaluate the impact of these improvements on user experience.

Overall, this research advances the field of visual prosthetics by enhancing image processing techniques and incorporating patient-specific parameters. These contributions are expected to improve the practical applicability of visual prostheses, providing users with higher-quality visual information.

Acknowledgments

This work was supported by project PID2021-125209OB-I00 (MCIN/AEI/10.13039/501100011033 and FEDER/UE).

References

[1] David Avraham and Yitzhak Yitzhaky. Simulating the perceptual effects of electrode–retina distance in prosthetic vision. *Journal of Neural Engineering*, 19(3):035001, 2022. 1

[2] 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):9199, 2019. 1

[3] Jaap de Ruyter van Steveninck, Umut Güçlü, Richard Wezel, and Marcel van Gerven. End-to-end optimization of prosthetic vision. *Journal of Vision*, 22:20, 2022. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[5] Tristan Fauvel and Matthew Chalk. Human-in-the-loop optimization of visual prosthetic stimulation. *Journal of Neural Engineering*, 19(3):036038, 2022. 1

[6] Angélica Pérez Fornos, Jörg Sommerhalder, Lyndon da Cruz, Jose Alain Sahel, Saddek Mohand-Said, Farhad

Figure 2. **Qualitative results**: The first two rows present some results from the MNIST dataset, with target images and reconstructions, respectively. Similarly, the last two rows present results from the CIFAR-10 dataset, target images and reconstructions, respectively.

Hafezi, and Marco Pelizzone. Temporal properties of visual perception on electrical stimulation of the retina. *Investigative ophthalmology & visual science*, 53(6):2720–2731, 2012. 1

[7] Jacob Granley, Tristan Fauvel, Matthew Chalk, and Michael Beyeler. Human-in-the-loop optimization for deep stimulus encoding in visual prostheses. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 2, 3

[8] Kathleen E Kish, Alex Yuan, and James D Weiland. Patient-specific computational models of retinal prostheses. *Scientific Reports*, 13(1):22271, 2023. 1

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[10] Sasidhar S Madugula, Alex R Gogliettino, Moosa Zaidi, Gorish Aggarwal, Alexandra Kling, Nishal P Shah, Jeff B Brown, Ramandeep Vilkhu, Madeline R Hays, Huy Nguyen, et al. Focal electrical stimulation of human retinal ganglion cells for vision restoration. *Journal of neural engineering*, 19(6):066040, 2022. 1

[11] Javad Paknahad, Kyle Loizos, Mark Humayun, and Gianluca Lazzi. Targeted stimulation of retinal ganglion cells in epiretinal prostheses: A multiscale computational study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2548–2556, 2020. 1

[12] Alejandro Perez-Yus, Maria Santos-Villafranca, Julia Tomas-Barba, Jesus Bermudez-Cameo, Lorenzo Montano-Olivan, Gonzalo Lopez-Nicolas, and Jose J Guerrero. Raspv: A robotics framework for augmented simulated prosthetic vision. *IEEE Access*, 12:15251–15267, 2024. 1

[13] Galen Pogoncheff, Zuying Hu, Ariel Rokem, and Michael Beyeler. Explainable machine learning predictions of perceptual sensitivity for retinal prostheses. *Journal of Neural Engineering*, 21(2):026009, 2024. 1

[14] Shinyong Shim, Kyungsik Eom, Joonsoo Jeong, and Sung June Kim. Retinal prosthetic approaches to enhance visual perception for blind patients. *Micromachines*, 11(5): 535, 2020. 1

[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, 2024. 1

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[17] Jing Wang, Haiyi Zhu, Jianyun Liu, Heng Li, Yanling Han, Ruyan Zhou, and Yun Zhang. The application of computer vision to visual prosthesis. *Artificial Organs*, 45(10):1141–1154, 2021. 1

[18] WHO. Blindness and vision impairment. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment, 2023. Accessed on 20.08.2024. 1

[19] Yuli Wu, Ivan Karetic, Johannes Stegmaier, Peter Walter, and Dorit Merhof. A deep learning-based in silico framework for optimization on retinal prosthetic stimulation, 2023. 1

[20] Yuli Wu, Julian Wittmann, Peter Walter, and Johannes Stegmaier. Optimizing retinal prosthetic stimuli with conditional invertible neural networks. *CoRR*, abs/2403.04884, 2024. 1