Deep learning approach for image classification-optimized retinal implant stimulation

Ismaïl Sahbane (326464), Oskar Boesch (341380), Mattia Moro (395819)
CS-503 Final Project Report

***Abstract*—While retinal implants can restore a certain level of grayscale vision in patients with degenerative retinal conditions, the resulting percepts are typically of very low resolution and often feature complex phosphene shapes. This project explores the optimization of stimulation patterns for PRIMA and Argus II retinal implants within an image classification pipeline, leveraging recent advances in percept simulation using the Pulse2Percept framework. The results demonstrate the potential of this method to bypass the need for human studies and rapidly evaluate large perceptual datasets. However, the approach also reveals limitations inherent to the classification task and susceptibility to adversarial attacks with encoders models not only reaching the classification baseline but surpassing it. Finally, this report highlights the need for more constrained encoders and robust, brain-inspired models to generate percepts that are more relevant to human vision.**

I. Introduction

Vision loss due to degenerative retinal conditions, such as retinitis pigmentosa (RP), profoundly affects millions worldwide, leading to significant impairments in daily life [1]. Visual prosthetics have made remarkable progress in restoring a form of vision in these patients. Among the most promising advancements is the PRIMA System, developed by Pixium Vision. This retinal implant restores a degree of vision by electrically stimulating the remaining retinal ganglion cells via an array of 378 electrodes which are activated based on images captured by camera glasses. Although promising, this technology, in addition to producing only grayscale percepts, faces a significant challenge: resolution [2]. Very little information is available on the exact video processing that occurs between the camera glasses and the electrode array in retinal prostheses, but it usually involves basic edge detection and contrast enhancement [3], resulting in visual percepts that often fail to convey meaningful or task-relevant information.

To better understand and simulate the percepts generated by these implants, Beyeler et al. developed Pulse2Percept, a Python-based framework that models the visual experience resulting from electrical stimulation of specific retinal implants [4]. This tool has been validated against patient-reported percepts, offering researchers a way to simulate and evaluate stimulation strategies without requiring in vivo testing [5].

Motivated by these challenges, this report investigates how optimizing electrode array stimulation can improve image representation for prosthetic vision. By leveraging recent advances in computer vision, we explore stimulation strategies that improve image classification performance of simulated percepts using benchmark datasets such as CIFAR-10 [6] and Imagenette [7].

II. Related Work

The enhancement of retinal prosthetic vision through deep learning has recently seen significant advances. A notable contribution is the work by Wu et al. [8], who introduced a fully neural framework comprising a CNN encoder, a perceptual simulation module approximating Pulse2Percept, and a VGG classifier. Their approach significantly outperformed bilinear downsampling on MNIST, illustrating the potential of end-to-end optimization in prosthetic vision systems. However, the use of simplistic data and limited perceptual evaluation restricted broader applicability.

Our project builds on this idea by adopting transformer and U-Net-based encoders, and by targeting more complex image distributions. Unlike Wu et al., our approach generates either spatially downscaled percepts or full stimulation vectors for devices like PRIMA. More generally, the need for realistic percept simulations has been underscored by van der Grinten et al. [9], who developed *DiffPercept*, a differentiable and biologically plausible simulator of cortical prosthetic vision. While our work is focused on retinal implants and utilizes static simulations from Pulse2Percept [4], their work highlights the importance of neurophysiological constraints and real-time simulation fidelity. Other deep learning approaches have aimed at improving retinal health assessments rather than perceptual simulation. For example, Christopher et al. [10] showed that structural metrics like GCIPL and RNFL thickness can be predicted from OCT using DL, reducing sample size requirements in glaucoma clinical trials. Complementary efforts in explainability, such as those by Shah et al. [11], leverage routine clinical measures to predict electrode deactivation in prosthetic users. This integration of explainable AI (XAI) enhances trust and reliability in device diagnostics. Finally, at the interface of material science and vision restoration, Rahmani and Eom [12] proposed the use of plasmonic silver nanoparticles in organic photovoltaic implants, emphasizing that optimizing perception requires innovations across both computation and hardware.

III. Method

Our approach revolves around what we call our Pipeline, which comprises several steps relating 3 models. Our end goal is to train an encoder that, given a high-resolution image, outputs a low-resolution stimuli such that the resulting percepts (passed through pulse2percept) are useful to human patients, enabling them to understand what's on the image. To build the pipeline allowing us to do that, we start from a (dataset, classification model) pair, where the latter can classify the former. We then insert between them our encoder, and the pulse2percept model, and freeze all of the pipeline except the encoder. We can then train the encoder on the classification loss. This project studies whether the percepts solving the classification task can also be useful to humans.

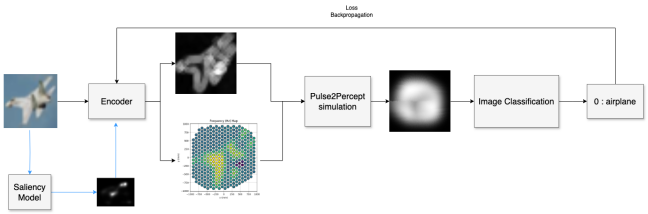


Figure 1: Graphical representation of the pipeline with the three core models : the Encoder, the Pulse2Percept Simulation and the Image Classification.

A. Datasets & Classification Models

We opted for image classification over object detection, as predicting bounding boxes over a small bottleneck percept as in PRIMA or ARGUS (both in cases of images or stimuli) is impractical, especially for small objects. Classification focuses optimization on preserving class-relevant features, indirectly encouraging spatial consistency.

To train and evaluate our pipelines, we used CIFAR-10 [6] and Imagenette [7], as they offer a balanced trade-off between complexity and computational feasibility. Larger datasets like complete ImageNet [13] were excluded due to resource constraints, while overly simplistic datasets like MNIST risked encouraging degenerate encoding strategies, as the model could solve classification without preserving perceptual details. CIFAR-10 has 50000 train images and 10000 test, each with size 32×32 in RGB format and belonging to one out of 10 total classes. Imagenette is a subset of ImageNet with 9469 training samples and 3925 test, each of size 128×128 in RGB format, with 10 possible classes also in this case. Both datasets include sufficient visual variability and resolution to make the down-scaling task meaningful, while still being accessible for training.

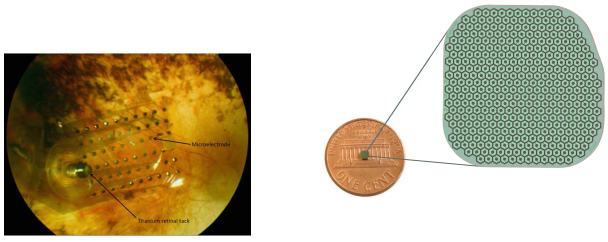
Each encoder received two inputs: (1) the RGB image, representing input from prosthetic camera glasses, and (2) a saliency map derived from OpenCV's spectral residual method. Saliency was used to guide the encoder toward perceptually important regions, inspired by findings demonstrating that user gaze and attention play a crucial role in prosthetic vision [14]. Though deep saliency models offer higher accuracy [15], we prioritized efficiency to maintain training speed and reproducibility.

B. Encoder

The first and main objective of this report is the transformation of an input image (as captured by the camera setup of a retinal implant) into specific stimulation patterns in the dimensions of retinal implants. Both Argus II (see Figure 2a) and PRIMA (see Figure 2b) implants will be studied. For this transformation, we propose two approaches. The first encodes an input image into a tensor matching the shape of the selected implant (e.g., 6×10 for Argus II) with a single channel representing the intensity of stimulation. The second approach, which we call parameterEncoder, applies the same transformation but takes into account three stimulation parameters for each electrode: frequency (Hz), amplitude (Th), and pulse duration (ms). This second approach results in three output channels.

Many encoders models could be used, ranging from CNN- to transformer-based architectures. We first designed and tested a custom UNet-based encoder [16] to preserve spatial locality while remaining lightweight and interpretable—critical for mapping stimuli to retinal implant layouts. Unlike standard U-Nets, our version omits the decoder and integrates downsampled copies of both the input image and its saliency map (inspired by DUCKNet [17]), encouraging detail preservation. The encoder includes three convolutional blocks with max pooling, progressively reducing resolution and increasing depth. After reaching a compact representation, a final CNN head reduces the feature depth to one channel and spatially compresses the output, which is then scaled to match the implant size or the stimulation parameters' dimensions. A schematic representation of this approach is shown in Supplementary Figure 6. This approach was only used for image encoding and not for parameter encoding. A second approach was used to maximize the flexibility of the encoder, employing a simple transformer encoder architecture (nn.TransformerEncoder) with both patch embedding and positional embedding. This approach was used for both image encoding and parameter encoding.

As a means of comparison, two baselines were established. The first, the DummyEncoder, simply takes an image and downscales it to match the implant size. The second, the DummyParameterEncoder, uses grid interpolation to predict the frequency parameter of a specific electrode from nearby pixels and fixes the pulse duration and amplitude values to 0.5.



(a) (b)

Figure 2: Comparison of Argus II and PRIMA retinal implants.Fundus photograph demonstrating an Argus II microelectrode array (6x10 electrodes, 3×3.1 mm) in situ (a) [18].PRIMA implant microelectrode array (378 electrodes, 2×2 mm) representation (b)[19].

C. Pulse2Percept Simulation

Pulse2Percept is a Python framework developed in 2017 by Beyeler et al. [20], which simulates the vision of implanted patients according to different models of retinal cell stimulation and perception. This report focuses on the axonal propagation model, which is, to date, the most realistic representation [5], where streaks of light following the axon structure are produced from single-electrode stimulation (see Supplementary Figure 7).

The AxonMapModel takes as input a tensor of shape matching the implant shape with a single channel representing intensities and produces percepts at a chosen field of view size and resolution. This model simulate the axon propagation seen in Figure 7. Pulse2Percept also provides a more complex version, the BiphasicAxonMapModel, based on the same theory, but takes as input, for each electrode, the three stimulation parameters discussed in the Encoder section. This latter model enables a more realistic and precise transformation and simulation of vision. Both models were used with a resolution of 0.1 degrees and a visual field ranging from -15 to 15 degrees for Argus II, and from -6 to 6 degrees for PRIMA, in both the x and y dimensions.

In order to train our encoder models, the pipeline must be fully differentiable and support end-to-end backpropagation. However, the Pulse2Percept models are not implemented in PyTorch, so we decided to train transformer models to approximate their percept predictions. The transformer models use the same architecture as the transformer encoder models described in the Encoder section, and the same input and output formats as the Pulse2Percept models.

In addition, the AxonMapModel introduces significant complexity due to the spread of activation along axonal pathways. To evaluate a simpler version of the problem, we also trained an additional encoder to match the predictions of the ScoreboardModel, a more optimistic model of percepts in retinal patients.

D. Human Study

To assess the quality of the various generated stimuli, 3 pipelines with significant results were selected. 10 corresponding percepts for each pipeline were shown to human subjects ($n = 2$) and they were asked to guess the class label. The final accuracy was measured for each pipeline.

IV. Experiments

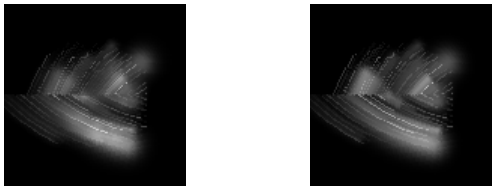
A. Pulse2Percept Transformer Training

The reliable generation of percepts and the generalization of these models are crucial for producing meaningful results in our pipeline. To this end, a first AxonMapTransformer was trained on two versions of CIFAR-10—one inverted and one non-inverted—both with increased contrast, as well as on a dataset of Perlin noise (see Supplementary Material). This resulted in 150,000 training samples transformed through the AxonMapModel. Two separate models were trained to match the percepts generated by the PRIMA and Argus II implants. To train BiphasicTransformers that match the predictions of the BiphasicAxonMapModel, we generated a dataset of 64,000 samples with random stimulation parameters. These parameters were sampled within common stimulation ranges (frequency: 0–30 Hz, amplitude: 0–1, duration: 0–1 ms). However, fully random sampling led to overstimulated percepts due to the high spread of large stimulation values. To mitigate this issue, parameters were instead sampled from a beta distribution with $\alpha = 1$ and $\beta = 5$, biasing the values toward lower intensities. To simulate the ScoreboardModel from Pulse2Percept, a simple Torch model was built, interpolating the grid of pixels to the electrode positions and producing a Gaussian distribution depending on the intensity of the input pixels. To compare the results obtained, a random baseline was used, generating random images with the same output shape. However, as seen with the AxonMapModel, the percepts are highly similar across different inputs. Therefore, two additional baselines were introduced: one that always predicts the percept resulting from complete and maximum stimulation, and another that predicts the percept resulting from no stimulation (a black image).

All pipelines using Pulse2Percept-like models trained on the AxonMapModel did not yield significant results, and their analysis will thus be excluded from this report. Importantly, evaluation of the Pulse2Percept-like models trained on the BiphasicAxonMapModel shows that both models trained to match the percepts of Argus II and PRIMA perform not only better than the random model, but also significantly better than both the complete stimulation and no stimulation baselines (see Table I). These results confirm the capacity of the trained torch models to produce coherent percepts, allowing them to be further used in the proposed pipelines. (See Figure 3)

Implant	Model	MSE	SSIM	L-SSIM
PRIMA	Random Model	4128.01	2.5e-03	-2.0e-03
PRIMA	Complete Stimuli Model	266.66	0.51	0.34
PRIMA	No Stimuli Model	4563.12	0.40	0.35
PRIMA	Transformer	6.28	0.93	0.86
ArgusII	Random Model	29847.66	1.9e-04	-1.42e-05
ArgusII	Complete Stimuli Model	41.31	0.88	0.79
ArgusII	No Stimuli Model	6636.25	0.62	0.59
ArgusII	Transformer (8,4,1)	13.88	0.94	0.89

Table I: Evaluation results of the pulse2percept-like models trained to match percepts generated by the BiphasicAxonMapModel for the PRIMA and ArgusII implants, as well as their respective baseline models. L-SSIM represents the lowest SSIM score obtained in the evalutation



(a) Model Output (b) Ground Truth

Figure 3: Comparison between model output and target image for the pulse2percept-like model trained to match the prediction of the BiphasicAxonMapModel and the PRIMA implant.

B. Pipeline Training

The results of various pipeline variants are shown in table II, in the order they are discussed.

1) CIFAR-10: As a proof of concept, two pipelines were trained on the classification task of CIFAR-10. The first pipeline used the PRIMA implant, the BiphasicAxonMap model from Pulse2Percept, and the parameter encoder (Pipeline 2). This method is the pipeline including percept simulation with the most degrees of freedom, transforming an input of size 32×32 to an output of size (378×3). To compare, a second pipeline (Pipeline 1) was trained on CIFAR-10 without percept simulation and with a smaller bottleneck of 16×16 (compared to the effective size of PRIMA, 22×19). The training results first demonstrate the complexity introduced by percept simulation, with a baseline accuracy of 0.142 (almost random) compared to 0.398 without simulation. Additionally, the final accuracy obtained by both models highlights the potential of the method: the accuracy of the first pipeline eventually matches that of the second pipeline with enough training. However, both accuracies remain quite low compared to the upper baseline for CIFAR-10 (0.846), reflecting the limited freedom in downscaling due to the small input image size of the dataset (32×32) and the U-Net architecture used.

2) Imagenette: To assess the importance of the input image size on the encoder's expressivity, two pipelines were trained with the same configuration but using Imagenette instead of CIFAR-10—respectively Pipeline 3 and Pipeline 4. The results show accuracies around 0.65 for both pipelines, nearly reaching the upper baseline of 0.75. This demonstrates the encoder's potential to learn relevant downscaling for the classification task when given a sufficiently large input size (128×128 for Imagenette). As expected, the greater downscaling amplitude leads to lower baselines compared to CIFAR-10 for both pipelines.

3) Implants: While the PRIMA implant theoretically enables greater resolution with an increased number of electrodes, its complex shape and pattern could introduce additional challenges. Furthermore, all electrodes are much smaller and positioned closer together, resulting in more spread of information. However, the comparison results with ArgusII (Pipeline 5 achieving 0.46 accuracy compared to Pipeline 3's 0.65) confirm that the final accuracy of our pipelines seems to be mainly correlated with the expressivity range of the encoder, and that the complex pattern of PRIMA does not significantly impact it.

4) Percept Simulation: To further investigate the importance of percept transformation in the final classification, four additional pipelines were trained with increasing clarity of percepts. The first (Pipeline 6) implements the ScoreBoard model, while the other three do not simulate percepts and simply classify the encoder output at sizes 10×10, 20×20, and 40×40, respectively. Interestingly, Pipeline 6 performs worse than its AxonMap equivalent (Pipeline 3), supporting the observation that the encoder's range of free-

N°	Implant	P2P Model	Encoder	Stimuli	Dataset	Accuracy (baseline)	Val CE error (baseline)
1	Prima	AxonMap	Transformer	Parameter	CIFAR10	0.41(0.142)	1.86(2.29)
2	16x16	-	UNet	Image	CIFAR10	0.436(0.398)	1.65(1.81)
3	Prima	AxonMap	Transformer	Parameter	imagenette	0.654(0.102)	1.18(7.32)
4	16x16	-	UNet	Image	imagenette	0.647(0.139)	1.06(4.20)
5	Argus	AxonMap	Transformer	Parameter	imagenette	0.460(0.104)	1.90(3.59)
6	Prima	Scoreboard	Transformer	Image	imagenette	0.512(0.093)	1.61(4.26)
7	10x10	-	Transformer	Image	imagenette	0.564(0.118)	1.38(4.77)
8	20x20	-	Transformer	Image	imagenette	0.893(0.162)	0.505(3.95)
9	40x40	-	Transformer	Image	imagenette	0.982(0.314)	0.172(2.89)

Table II: Results and parameters for all trained pipelines. The implant column dictates the bottle-neck size with PRIMA 22x19 or 378 electrodes and ArgusII 6x10 or 60 electrodes. When no P2P Model is specified the pipeline simply classify the output of the encoder without percept simulation.

dom is more relevant to pipeline performance than the actual clarity of the generated percepts. However, the results from pipelines without percept simulation show that introducing percepts still complicates the task. Notably, Pipelines 8 and 9 performed better than the upper baseline, achieving accuracies of 0.893 and 0.982, respectively, compared to the baseline of 0.75.

C. Human Study

Although many pipelines reached similar or even higher accuracies than the classification models themselves, this does not ensure the main aim of this report to produce human-relevant percepts. In fact, modern deep learning models are known to not always learn human-relevant features and this concern is confirmed by the results of our human study, shown in table III. High accuracy is obtained for the unaltered CIFAR10 and Imagenette datasets, while very low accuracy is obtained for pipelines 3 and 9, which are representative of our experiments (pipeline 9 had accuracy 0.98 with its respective model!). Finally, pipeline 2 yields acceptable accuracy, and is the only one which produced spatially coherent percepts without being adversarial, which we interpret as a consequence of its smaller encoder size and relatively larger bottleneck. Figure 4 shows percepts for the three evaluated pipelines, along with their dataset input and label.

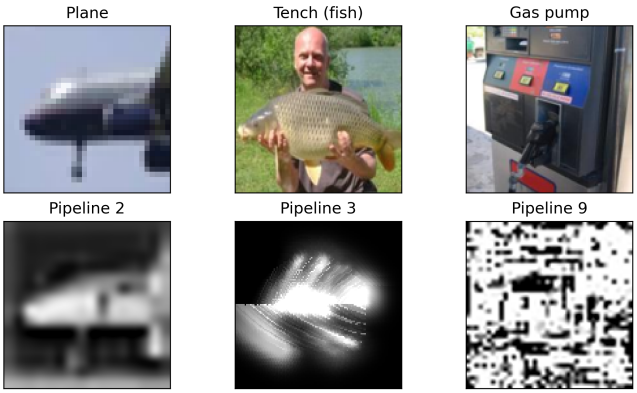


Figure 4: Dataset inputs and labels (top) and predicted percepts (bottom) for pipelines used in the human study.

N°	Human Accuracy
2	0.4
3	0.1
9	0.15

Table III: Accuracy results of the human study averaged over the two participants. 10 samples of each pipeline were classified by the participants. CIFAR-10 and Imagenette were additionally tested without any transformation as means of comparison obtaining respectively 0.75 and 0.9 accuracy.

V. Discussion

The human study result demonstrates that the percepts generated by most of our models are completely useless to potential patients, even though the corresponding accuracies are often very high. In the extreme case of pipeline 9, with a 40×40 bottleneck, the accuracy obtained is much higher than the accuracy of the classification model itself (0.982 vs 0.752). Our interpretation is that the encoder, instead of learning a down-sampling operation, learned to first solve the image classification problem latently, and then generate, depending on the label, adversarial stimuli that make the fixed classification model output the desired label. Its high accuracy suggests that the encoder solved the classification task significantly better than the classification model itself, which completely negates the usefulness of our pipeline. A reason enabling this outcome is that the encoder has a more powerful architecture than the image classifier, and that the task to solve (classification with 10 classes) is too easy, such that this adversarial behavior is the best strategy for our encoder.

Several weak-points in the pipeline may be exploited by the encoder. First is the capacity of the trained pulse2percept-like models to generalize to new form of inputs. In fact, in order to allow the proposed encoders to explore the full range of possibility, their stimuli need to produce coherent percepts. Although the trained pulse2percept-like models showed good results, they might generalize poorly to some form of new inputs which would make the results obtained completely incoherent and unrealistic, and those errors could be exploited by a powerful encoder. Secondly and crucially, the classification model we used is very simple, and is likely vulnerable to adversarial inputs [21]. In fact, it is only trying to partition image space into 10 regions, hence most images classified in a given class will not resemble the desired object to humans. Given this, the unusable percepts we obtain are expected.

VI. Conclusion and Future Work

To help with the adversarial problem presented above, a more difficult vision task could be used in the pipeline, such as classification with many more classes ($1000+$), object detection, segmentation, or even navigation. This would make the task too difficult for the transformer to solve it itself and force it to actually perform a smart downsampling operation, that would then allow the downstream model to solve it. Even then, it would not guarantee that low-resolution stimuli that are useful to the downstream model will be useful for patients, as most models are very different from the human visual system. Using models made to imitate the way the human brain works, such as models scoring high on the BrainScore [22] could be worth exploring.

REFERENCES

[1] N. Cross, C. van Steen, Y. Zegaoui, A. Satherley, and L. Angelillo, "Retinitis pigmentosa: Burden of disease and current unmet needs," *Clin. Ophthalmol.*, vol. 16, pp. 1993–2010, Jun. 2022.

[2] D. Palanker, Y. Le Mer, S. Mohand-Said, M. Muqit, and J. A. Sahel, "Photovoltaic restoration of central vision in atrophic age-related macular degeneration," *Ophthalmology*, vol. 127, no. 8, pp. 1097–1104, Aug. 2020.

[3] J. Park, A. K. Goldstein, Y. Zhuo, N. Jensen, and D. Palanker, "Simulation of prosthetic vision with prima system and enhancement of face representation," 2025. [Online]. Available: https://arxiv.org/abs/2503.11677

[4] M. Beyeler, G. M. Boynton, I. Fine, and A. Rokem, "pulse2percept: A python-based simulation framework for bionic vision," in *Proceedings of the 16th Python in Science Conference (SciPy)*, 2017, pp. 81–88.

[5] M. Beyeler, D. Nanduri, J. D. Weiland, A. Rokem, G. M. Boynton, and I. Fine, "A model of ganglion axon pathways accounts for percepts elicited by retinal implants," *Scientific reports*, vol. 9, no. 1, p. 9199, 2019.

[6] A. Krizhevsky, "Learning multiple layers of features from tiny images," https://www.cs.toronto.edu/~kriz/cifar.html, 2009, accessed: 2025-05-26.

[7] J. Howard, "Imagenette: A smaller subset of imagenet," https://github.com/fastai/imagenette, 2019, accessed: 2025-05-26.

[8] Y. Wu, F. Wang, U. P. Froriep, and M. Beyeler, "A deep learning-based in silico framework for optimization on retinal prosthetic stimulation," *arXiv preprint arXiv:2302.03570*, 2023.

[9] R. van der Grinten, A. Lahiri, and M. Beyeler, "Diffpercept: A differentiable phosphene simulator for visual neuroprostheses," *eLife*, vol. 12, p. RP85812, 2023.

[10] M. Christopher, C. Bowd, J. A. Proudfoot *et al.*, "A deep learning approach to improve retinal structural predictions and aid glaucoma neuroprotective clinical trial design," *Frontiers in Cellular Neuroscience*, vol. 17, p. 10507809, 2023.

[11] M. Shah, M. Beyeler *et al.*, "Explainable ai for retinal prostheses: Predicting electrode deactivation from routine clinical measures," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021.

[12] M. Rahmani and J.-H. Eom, "Plasmonic silver nanoparticle-enhanced organic photovoltaic retinal implants for improved visual prostheses," *Frontiers in Cellular Neuroscience*, vol. 18, p. 1385567, 2024.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[14] M. P. Barry and G. Dagnelie, "Use of fixation and scanning movements improves low-resolution vision simulated with a retinal prosthesis," *Journal of Vision*, vol. 16, no. 3, pp. 21–21, 2016.

[15] S. Wang, J. Fu, T. Mei, and J. Luo, "A survey on deep learning for named object saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3416–3445, 2021.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[17] R.-G. Dumitru, D. Peteleaza, and C. Craciun, "Using duck-net for polyp image segmentation," *Scientific Reports*, vol. 13, no. 1, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41598-023-36940-5

[18] Y. H.-L. Luo and L. Da Cruz, "The argus® ii retinal prosthesis system," *Progress in retinal and eye research*, vol. 50, pp. 89–107, 2016.

[19] Science.xyz, "Prima: Photovoltaic retinal implant for vision restoration," https://science.xyz/technologies/prima/, 2025, accessed: 2025-05-26.

[20] M. Beyeler, G. M. Boynton, I. Fine, and A. Rokem, "pulse2percept: A python-based simulation framework for bionic vision," Jun. 2017. [Online]. Available: http://dx.doi.org/10.1101/148015

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[22] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger *et al.*, "Brain-score: Which artificial neural network for object recognition is most brain-like?" *BioRxiv*, p. 407007, 2018.

VII. Individual Contributions

Oskar brought the initial idea, researched the field of neural implants and computational percept simulations, implemented the pulse2percept framework and built and trained various transformers to simulate it in different contexts (and generated the datasets to train them), and designed the human study. In general, he was responsible for the general direction of the project, bringing propositions on which experiments to conduct. Mattia designed the UNet encoder, implemented support for saliency maps, and developed the support for the CIFAR10 dataset. He also explored extensively and developed partly an experiment using navigation task and RL, which wasn't included in the final report, and added support for object detection and the COCO dataset, which we didn't use in the end. Ismaïl developed the full pipeline, developed the environment for training it, using config files, designed evaluation metrics and baselines, implemented the transformer encoder, specified and evaluated the different versions of the pipeline and baselines, and implemented the imagenette dataset and the classification models.

VIII. Appendix

A. Prima Complex Pattern

It is important to note that all percept-prediction models from Pulse2Percept already perform a form of grid interpolation for PRIMA, as the implant is not a perfect grid but a complex pattern of hexagonal shapes. This explains why the shape of PRIMA is considered to be 22×19—a square-like approximation of the implant that does not match the actual number of electrodes, which is 378.

B. RL for Navigation Task

For learning paradigms, Habitat Lab enables training the encoder in a more dynamic, task-driven setting. Our idea is to train a reinforcement learning agent for PointGoal navigation while equipping it with a virtual, learnable retinal implant. The agent, placed in a Gibson room, has to reach a random target using a custom "implant policy": RGB and depth inputs (mimicking external cameras) are processed by our encoder and Pulse2Percept simulation, then passed to Habitat's ResNet+RNN policy (Figure 6). The agent thus navigates as if perceiving through a retinal implant and is trained end-to-end, learning to produce percepts that enhance task performance. This setup allows it to evolve task-specific representations, potentially yielding more adaptive and meaningful visual transformations than those obtained through static, classification-based training alone.

IX. Supplementary Figures

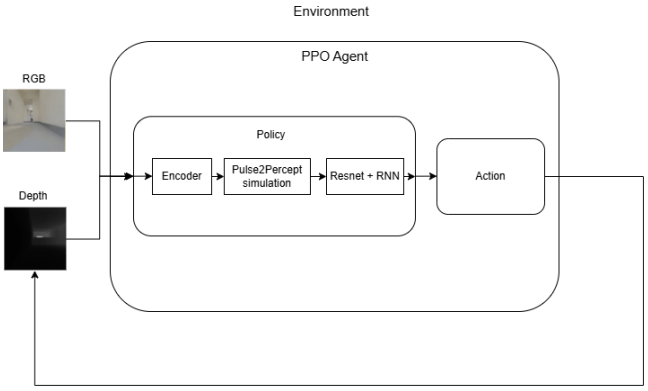


Figure 5: DRL hypothetical setting

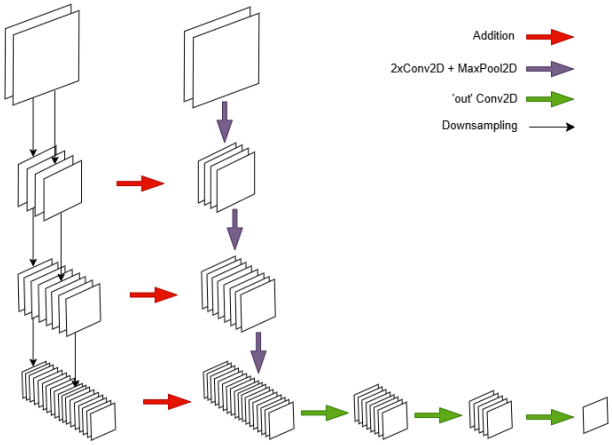


Figure 6: Unet Encoder Architecture

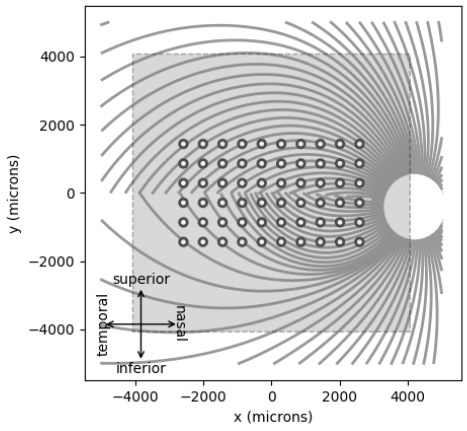


Figure 7: AxonMapModel Visualization using pulse2percept with the relative posiiton and size of ArgusII electrodes.