UC Santa Barbara
UC Santa Barbara Previously Published Works

Title

Deep Learning--Based Scene Simplification for Bionic Vision

Permalink

https://escholarship.org/uc/item/8c12d0wc

Authors

Han, Nicole
Srivastava, Sudhanshu
Xu, Aiwen
et al.

Publication Date

2021-01-30

Peer reviewed

Deep Learning–Based Scene Simplification for Bionic Vision

Nicole Han
University of California
Santa Barbara, CA, USA
xhan01@ucsb.edu

Sudhanshu Srivastava*
University of California
Santa Barbara, CA, USA
sudhanshu@ucsb.edu

Aiwen Xu*
University of California
Santa Barbara, CA, USA
aiwenxu@ucsb.edu

Devi Klein
University of California
Santa Barbara, CA, USA
dklein@ucsb.edu

Michael Beyeler
University of California,
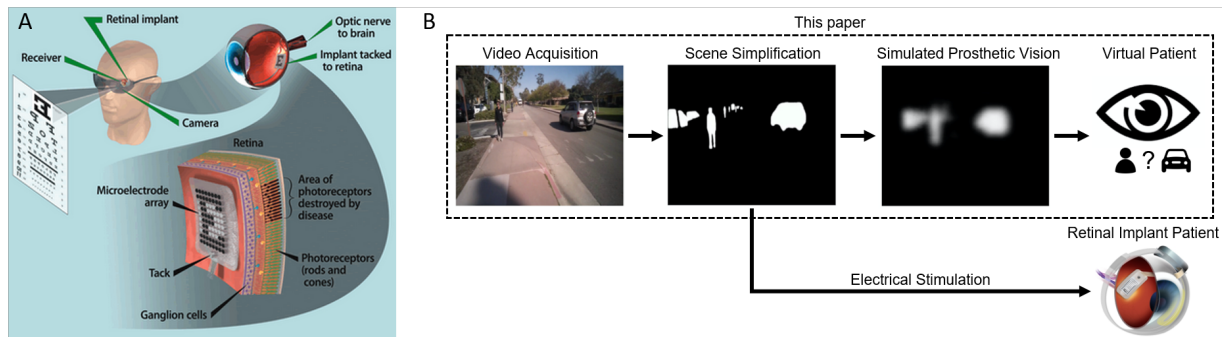Santa Barbara, CA, USA
mbeyeler@ucsb.edu

Figure 1: Retinal implant ('bionic eye') for restoring vision to people with visual impairment. *A)* **Light captured by a camera is transformed into electrical pulses delivered through a microelectrode array to stimulate the retina (adapted with permission from [39]).** *B)* **To create meaningful artificial vision, we explored deep learning–based scene simplification as a preprocessing strategy for retinal implants (reproduced from doi:10.6084/m9.figshare.13652927 under CC-BY 4.0). As a proof of concept, we used a neurobiologically inspired computational model to generate realistic predictions of simulated prosthetic vision (SPV), and asked sighted subjects (i.e., virtual patients) to identify people and cars in a novel SPV dataset of natural outdoor scenes. In the future, this setup may be used as input to a real retinal implant.**

ABSTRACT

Retinal degenerative diseases cause profound visual impairment in more than 10 million people worldwide, and retinal prostheses are being developed to restore vision to these individuals. Analogous to cochlear implants, these devices electrically stimulate surviving retinal cells to evoke visual percepts (phosphenes). However, the quality of current prosthetic vision is still rudimentary. Rather than aiming to restore "natural" vision, there is potential merit in borrowing state-of-the-art computer vision algorithms as image processing techniques to maximize the usefulness of prosthetic vision. Here we combine deep learning–based scene simplification strategies with a psychophysically validated computational model of the retina to generate realistic predictions of simulated prosthetic vision, and measure their ability to support scene understanding of sighted subjects (virtual patients) in a variety of outdoor scenarios. We show that object segmentation may better support scene understanding than models based on visual saliency and monocular depth estimation. In addition, we highlight the importance of basing theoretical predictions on biologically realistic models of phosphene shape. Overall, this work has the potential to drastically improve the utility of prosthetic vision for people blinded from retinal degenerative diseases.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility technologies**; **Empirical studies in visualization**; *Usability testing*.

KEYWORDS

retinal implant, visually impaired, scene simplification, deep learning, simulated prosthetic vision, vision augmentation

*Both authors contributed equally to this research.

1 INTRODUCTION

Retinal degenerative diseases such as retinitis pigmentosa (RP) and age-related macular degeneration (ARMD) lead to a gradual loss of photoreceptors in the eye that may cause profound visual impairment in more than 10 million people worldwide. Analogous to cochlear implants, retinal neuroprostheses (also known as the *bionic eye*, Fig. 1A) aim to restore vision to these individuals by electrically stimulating surviving retinal cells to evoke neuronal responses that are interpreted by the brain as visual percepts (*phosphenes*). Existing devices generally provide an improved ability to localize high-contrast objects, navigate, and perform basic orientation tasks [2]. Future neural interfaces will likely enable applications such as controlling complex robotic devices, extending memory, or augmenting natural senses with artificial inputs [14].

However, despite recent progress in the field, there are still several limitations affecting the possibility to provide useful vision in daily life [8]. Interactions between the device electronics and the underlying neurophysiology of the retina have been shown to lead to distortions that can severely limit the quality of the generated visual experience [7, 15]. Other challenges include how to improve visual acuity, enlarge the field-of-view, and reduce a complex visual scene to its most salient components through image processing.

Rather than aiming to restore "natural" vision, there is potential merit in borrowing computer vision algorithms as preprocessing techniques to maximize the usefulness of bionic vision. Whereas edge enhancement and contrast maximization are already routinely employed by current devices, relatively little work has explored the extraction of high-level scene information.

To address these challenges, we make three contributions:

 i. We adopt state-of-the-art computer vision algorithms to explore deep learning–based scene simplification as a preprocessing strategy for bionic vision.
 ii. Importantly, we use an established and psychophysically validated computational model of bionic vision to generate realistic predictions of simulated prosthetic vision (SPV).
iii. We systematically evaluate the ability of these algorithms to support scene understanding with a user study focused on a novel dataset of natural outdoor scenes.

2 BACKGROUND

Retinal implants are currently the only FDA-approved technology to treat blinding degenerative diseases such as RP and ARMD. Most current devices acquire visual input via an external camera and perform edge extraction or contrast enhancement via an external video processing unit (VPU), before sending the signal through wireless coils to a microstimulator implanted in the eye or the brain (see Fig. 1A). This device receives the information, decodes it and stimulates the visual system with electrical current, ideally resulting in artificial vision. Two devices are already approved for commercial use: Argus II (60 electrodes, Second Sight Medical Products, Inc., [26]) and Alpha-IMS (1500 electrodes, Retina Implant AG, [36]). In addition, PRIMA (378 electrodes, Pixium Vision, [24]) has started clinical trials, with others to follow shortly [3, 13].

However, a major outstanding challenge in the use of these devices is translating electrode stimulation into a code that the brain can understand. A common misconception is that each electrode

in the grid can be thought of as a 'pixel' in an image [10, 11, 25, 30, 33], and most retinal implants linearly translate the grayscale value of a pixel in each video frame to a current amplitude of the corresponding electrode in the array [26]. To generate a complex visual experience, the assumption then is that one simply needs to turn on the right combination of pixels.

In contrast, a growing body of evidence suggests that individual electrodes do not lead to the perception of isolated, focal spots of light [7, 12, 15]. Although consistent over time, phosphenes vary drastically across subjects and electrodes [7, 27] and often fail to assemble into more complex percepts [31, 41]. Consequently, retinal implant users do not see a perceptually intelligible world [12].

A recent study demonstrated that the shape of a phosphene generated by a retinal implant depends on the retinal location of the stimulating electrode [7]. Because retinal ganglion cells (RGCs) send their axons on highly stereotyped pathways to the optic nerve, an electrode that stimulates nearby axonal fibers would be expected to antidromically activate RGC bodies located peripheral to the point of stimulation, leading to percepts that appear elongated in the direction of the underlying nerve fiber bundle (NFB) trajectory (Fig. 2, *right*). Ref. [7] used a simulated map of NFBs in each patient's retina to accurately predict phosphene shape, by assuming that an axon's sensitivity to electrical stimulation:

 i. decays exponentially with decay constant ρ as a function of distance from the stimulation site,
 ii. decays exponentially with decay constant λ as a function of distance from the cell body, measured as axon path length.

As can be seen in Fig. 2 (*left*), electrodes near the horizontal meridian are predicted to elicit circular percepts, while other electrodes are predicted to produce elongated percepts that will differ in angle based on whether they fall above or below the horizontal meridian. In addition, the values of ρ and λ dictate the size and elongation of elicited phosphenes, respectively, which may drastically affect visual outcomes. Understanding the qualitative
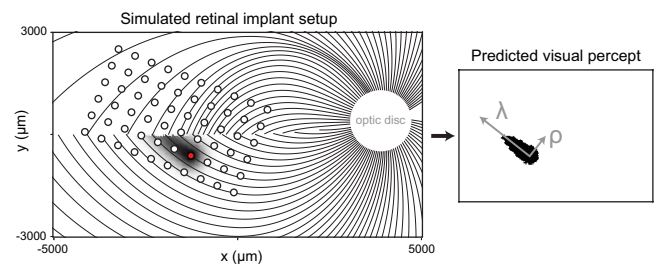
Figure 2: A simulated map of retinal NFBs (*left*) can account for visual percepts (*right*) elicited by retinal implants (reprinted with permission from [6]). *Left*: Electrical stimulation (red circle) of a NFB (black lines) could activate retinal ganglion cell bodies peripheral to the point of stimulation, leading to tissue activation (black shaded region) elongated along the NFB trajectory away from the optic disc (white circle). *Right*: The resulting visual percept appears elongated as well; its shape can be described by two parameters, λ (spatial extent along the NFB trajectory) and ρ (spatial extent perpendicular to the NFB). See Ref. [6] for more information.

experience associated with retinal implants and finding ways to improve is therefore indispensable to the development of visual neuroprostheses and related vision augmentation technologies.

3 RELATED WORK

Most retinal implants are equipped with an external VPU that is capable of applying simple image processing techniques to the video feed in real time. In the near future, these techniques may include deep learning–based algorithms aimed at improving a patient's scene understanding.

Based on this premise, researchers have developed various image optimization strategies, and assessed their performance by having sighted observers (i.e., *virtual patients*) conduct daily visual tasks under simulated prosthetic vision (SPV) [1, 9, 11, 21, 28, 37]. Simulation allows a wide range of computer vision systems to be developed and tested without requiring implanted devices.

SPV studies suggest that one benefit of image processing may be to provide an importance mapping that can aid scene understanding; that is, to enhance certain image features or regions of interest, at the expense of discarding less important or distracting information [1, 9, 17, 33]. This limited compensation may be significant to retinal prosthesis patients carrying out visual tasks in daily life.

One of the most commonly explored strategies is to highlight visually salient information in the scene. In biologically-inspired models, visual saliency is often defined as a *bottom-up* process that highlights regions whose low-level visual attributes (e.g., color, contrast, motion) may differ from their immediate surroundings. Early work used this approach to build a visual saliency map whose salient regions coincided with the regions gazed at by human subjects when looking at images [29]. More recent research showed that saliency was able to improve eye-hand coordination [20], obstacle avoidance [35], object detection [40], and object recognition [21, 38]. However, saliency prediction improved markedly with the advent of deep learning models, which are commonly trained on human eye movement data to predict an observer's fixation locations while freely-viewing a set of images. The current state-of-the-art in saliency prediction is DeepGaze II [18], a probabilistic model that uses transfer learning from VGG-19 pre-trained on the SALICON dataset. DeepGaze has yet to be applied to the field of bionic vision.

Current retinal prostheses are implanted in only one eye, and thus are unable to convey binocular depth cues. Previous work has therefore explored the possibility of obtaining depth information through additional peripherals, such as an RGB-D sensor, and studied behavioral performance of virtual patients typically navigating an obstacle course under SPV. For example, Ref. [30] used depth cues to generate a simplified representation of the ground to indicate the free space within which virtual patients could safely walk around. Depth cues were also shown to help avoid nearby obstacles that are notoriously hard to detect with other computer vision algorithms, such as branches hanging from a tree [23]. Ref. [28] used depth to increase the contrast of object boundaries and showed that this method reduced the number of collisions with ground obstacles. In addition, retinal prosthesis patients were shown to benefit from distance information provided by a thermal sensor when trying to avoid nearby obstacles and people [32]. However, recent advances in deep learning enable the estimation of relative depth from single monocular images, thereby eliminating the need of external depth sensors and peripherals. One of the most promising deep neural networks is monodepth2 [16], which uses a self-supervised method to estimate per-pixel monocular depth. Deep learning–based depth estimation has yet to be applied to the field of bionic vision.

Finally, recent advances in semantic segmentation have found application in bionic vision to simplify the representation of both outdoor scenes [17] and indoor scenes [33]. The latter study combined semantic and structural image segmentation to build a schematic representation of indoor environments, which was then shown to improve object and room identification in a SPV task [33].

However, a common limitation of all the above studies is that their prosthetic vision simulation assumed that phosphenes are small, isolated, independent light sources. It is therefore unclear how their findings would translate to real retinal prosthesis patients, whose phosphenes are large, elongated, and often fail to assemble into more complex percepts [7, 12, 31, 41]. In addition, since the above algorithms were developed in isolation and tested on different behavioral tasks, a side-by-side comparison of their ability to aid scene understanding is still lacking.

To address these challenges, we used a neurobiologically inspired computational model of bionic vision to generate realistic predictions of SPV, and applied it to several state-of-the-art computer vision algorithms that might be used to aid scene understanding. To allow for a fair comparison between algorithms, we asked virtual patients to make perceptual judgments about natural outdoor scenes, and assessed their performance using objective metrics as we systematically varied a number of model parameters.

4 METHODS

Following the workflow outlined in Fig. 1B, we created SPV videos of various outdoor scenes captured by a head-mounted camera (Section 4.1). We first processed the raw videos with one of four scene simplification strategies based on state-of-the-art computer vision algorithms (Section 4.2). We then fed the preprocessed videos into a prosthetic vision simulator to simulate the artificial vision likely to be experienced by different retinal prosthesis patients (Section 4.3). Example frames of the resulting SPV videos can be seen in Fig. 3. Finally, we conducted a user study to evaluate how well the resulting SPV videos could support scene understanding in a variety of outdoor scenarios (Section 4.4).

4.1 Visual Stimuli

Stimuli consisted of 16 first-person videos (each 5 s long) recorded on the University of California, Santa Barbara (UCSB) campus using head-mounted Tobii Pro Glasses 2. All videos were recorded outdoors in broad daylight, and were aimed at capturing scenarios that are relevant for orientation and mobility of a retinal prosthesis patient (e.g., walking on a sidewalk, crossing a street, strolling through a park). The final dataset was carefully assembled so as to uniformly cover a variety of conditions. That is, four videos did not include any people or cars; four videos had one or more person present; four videos had one or more cars present; and four videos had both people and cars present.

The raw dataset is publicly available as part of the Supplementary Material (see Section 7).
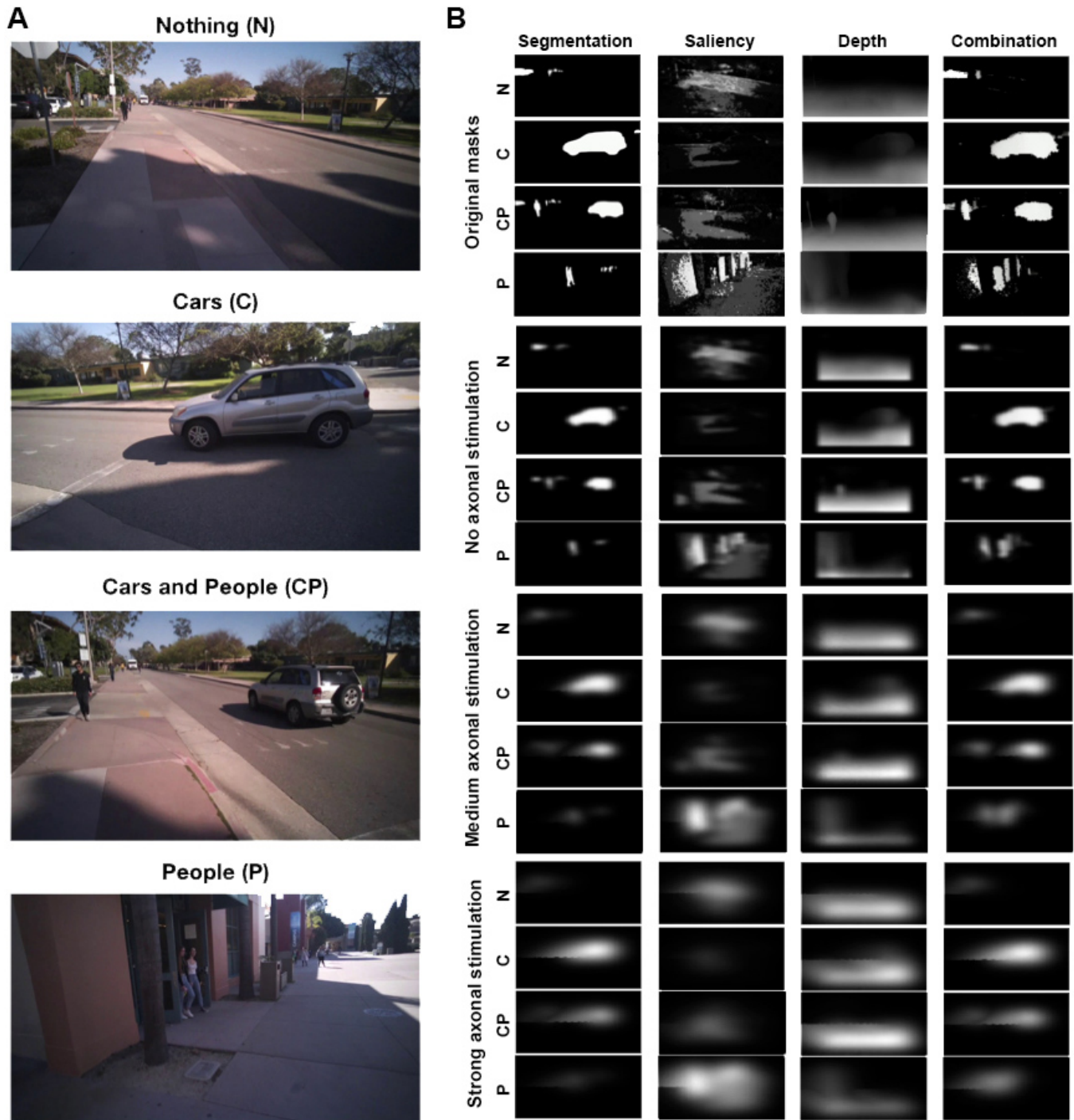
Figure 3: Example frames of the simulated prosthetic vision (SPV) dataset. *A)* **Example frames containing: neither people nor cars (N), cars only (C), both cars and people (CP), and people only (P).** *B)* **Same example frames after being processed with different scene simplification strategies (columns) and SPV of a 32×32 electrode array with different phosphene sizes and elongations (rows). Simulations are shown for the original masks (no SPV), small phosphenes with no axonal stimulation ($\rho = 100\,\mu m, \lambda = 0\,\mu m$), medium-sized phosphenes with intermediate axonal stimulation ($\rho = 300\,\mu m, \lambda = 1000\,\mu m$), and large phosphenes with strong axonal stimulation ($\rho = 500\,\mu m, \lambda = 5000\,\mu m$). Phosphene size and elongation drastically affect SPV quality, but previous work often ignored these parameters in their predictions.**

4.2 Scene Simplification Strategies

Stimuli were processed by four different scene simplification strategies, adapted from state-of-the-art computer vision algorithms.

4.2.1 Highlighting Visually Salient Information. We used DeepGaze II [18] to highlight visually salient information in the recorded videos. DeepGaze II produced a saliency map that assigned an importance value $\in [0, 1]$ to each pixel in a given video frame. Importance values were then linearly mapped to stimulus amplitudes applied to the simulated retinal implant.

4.2.2 Substituting Depth for Intensity. We used a self-supervised monocular depth estimation model called monodepth2 [16] (specifically the pre-trained `mono+stereo_640x192 model`) to predict a per-pixel relative depth map from each frame in the videos. We first sorted the depth values of all pixels in a frame and removed all depth values above the 80th percentile (where the 0th and 100th percentile referred to the nearest and farthest pixels to the viewer, respectively). We then applied an exponential decay on the depth values such that the closest pixels had grayscale value 180 and the farthest pixels had grayscale value 0.

4.2.3 Object Segmentation. To segment objects of interest from background clutter, we used a combination of the scene parsing algorithm from the MIT Scene Parsing Benchmark [43, 44] and an object segmentation algorithm called detectron2 [42]. Given that all the stimuli were outdoor scenes, we obtained the detected object binary masks that were labeled as a person, bicycle, car, or bus for each video frame. If there was no object detected in the scene, then we only represented the main structural edges from the scene-parsing algorithm. The scene-parsing algorithm sometimes produces more than 50 parsed regions from the scene. In order to produce less clustered output, we only preserve the regions labeled as roads or sidewalks. For the parsed regions, we then extracted the structural edges for better visualization in the end. The resulting binary masks were then linearly mapped to stimulus amplitudes applied to the simulated retinal implant.

4.2.4 Combining Saliency, Depth, and Segmentation. Recognizing the complementary qualities of the three algorithms described above, we wondered whether a combination of saliency, depth, and object segmentation could further improve scene understanding. While segmentation excels at highlighting objects of interests, it might miss regions of interest that do not have a clear semantic label (which would be highlighted by the more bottom-up–driven saliency detector) or nearby obstacles (which would be highlighted by the depth algorithm). To arrive at a binary mask of salient objects, we thresholded the saliency map to retain only the $10\,\%$ most salient pixels and combined it with the object segmentation map using a logical OR. We then scaled the grayscale value of each pixel in the new binary mask with a quadratic function of depth, similar to the above: $y = -\frac{45}{16}\left(\frac{8}{d_{\max}-d_{\min}}x - \frac{16}{d_{\max}-d_{\min}}\right)^2 + 180$.

4.3 Simulated Prosthetic Vision

The preprocessed videos were then used as input stimuli to the pulse2percept simulator [5], which provides an open-source implementation of Ref. [7] (among others). The simulator takes a downscaled version of the preprocessed image, and interprets the grayscale value of each pixel in a video frame as a current amplitude delivered to the simulated retinal implant. However, pulse2percept describes the output of SPV not as a pixelated image, but determines the shape of each phosphene based on the retinal location of the simulated implant as well as model parameters ρ and λ (see Section 2). As can be seen in (Fig. 2, *left*), electrodes near the horizontal meridian were thus predicted to elicit circular percepts, while other electrodes were predicted to produce elongated percepts that differed in angle based on whether they fell above or below the horizontal meridian.

Importantly, ρ and λ seem to vary drastically across patients [7]. Although the reason for this is not fully understood, it is clear that the choice of these parameter values may drastically affect the quality of the generated visual experience. To cover a broad range of potential visual outcomes, we thus simulated nine different conditions with $\rho = \{100, 300, 500\}$ and $\lambda = \{0, 1000, 5000\}$.

To study the effect that the number of electrodes in a retinal implant has on scene understanding, we simulated three different retinal implants consisting of 8×8, 16×16, and 32×32 electrodes arranged on a rectangular grid. These sizes roughly correspond to existing and near-future retinal implants.

4.4 Virtual Patients

We recruited 45 sighted undergraduate students (ages 18–21; 31 females, 14 males) from the student pool at UCSB to act as virtual patients in our experiments. Subjects were asked to watch SPV videos depicting various outdoor scenes and indicate whether they believe people and/or cars to be present in the scene. We were primarily interested in investigating their perceptual performance as a function of the four different scene simplification strategies, three retinal implant resolutions, and nine combinations of model parameters ρ and λ. All experiments were performed under a protocol approved by the university's Institutional Review Board.

4.4.1 Experimental Setup and Apparatus. The experiment was set up using a recent online platform called SimplePhy [19]. All subjects completed the experiment online using a personal laptop or computer.

We used a between-subjects design where each subject was randomly assigned to one of the nine model parameter conditions ($\rho \in \{100, 300, 500\} \times \lambda \in \{0, 1000, 5000\}$). Each condition was completed by five different subjects. Within each condition, each subject completed all 16 videos with the four scene simplification strategies (depth, saliency, segmentation, combination) and three electrode grid resolutions (8×8, 16×16, 32×32). Therefore, each subject completed 192 trials ($16 \times 4 \times 3$) in total, which took about 45–60 minutes to finish.

4.4.2 Experimental Task and Procedure. Subjects underwent a short online practice session consisting of 8 practice trials, where they were shown original videos from the head-mounted camera alongside their corresponding SPV videos. An example trial is shown in Fig. 4. Note that the video sequences used in the practice session did not appear in the actual experiment. After each video, a new screen appeared ('response screen' in Fig. 4) on which subjects indicated whether they believed the scene contained any people or cars. Subjects also indicated their confidence on a five-level Likert scale

Press the Space Bar to Start
Press the Enter Key to Replay

Response screen

Trial 1

There is one or more cars in the video

True False

How confident are you? 1 = not confident at all, 5 = completely confident

1 2 3 4 5

There is one or more people in the video

True False

How confident are you? 1 = not confident at all, 5 = completely confident

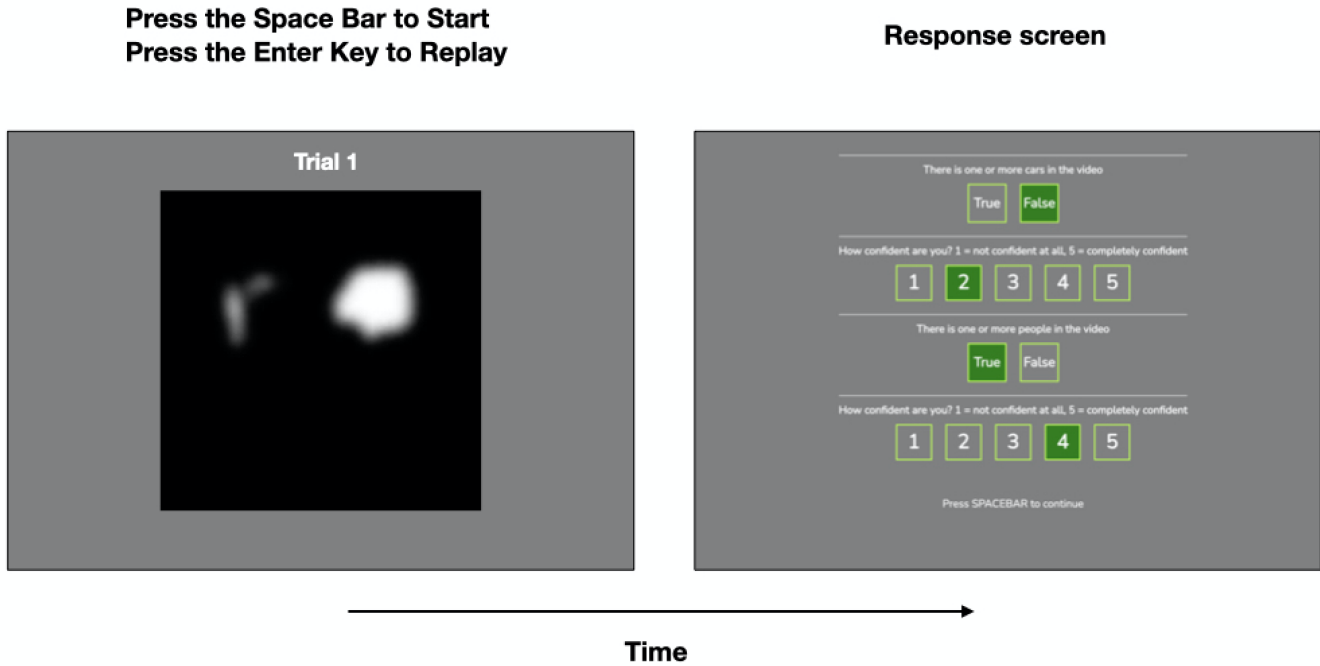1 2 3 4 5

Press SPACEBAR to continue

Time

Figure 4: Example trial conducted using the SimplePhy online platform [19]. After watching a five-second long video clip of a simulated prosthetic vision (SPV) outdoor scene, participants had to indicate whether they believe cars and people to be present in the scene. Participants also indicated their confidence on a five-level Likert scale (1 = not confident at all, 5 = completely confident).

(1 = not confident at all, 5 = completely confident). Detecting cars and people is an essential component for orientation & mobility. Increasing a patient's ability to detect and recognize moving objects may prevent them from dangerous situations in real-life scenarios.

4.4.3 Evaluating performance. Perceptual performance was assessed using the sensitivity index (d', "d-prime"), which is a dimensionless statistic from signal detection theory that can be used to measure a participant's perceptual sensitivity [34]:

$$d' = Z(\text{hit rate}) - Z(\text{false discovery rate}), \quad (1)$$

where the function $Z(p)$, with $p \in [0, 1]$, is the inverse of the cumulative distribution function of the Gaussian distribution. Here, the hit rate was calculated from the number of trials in which a participant correctly identified people or cars to be present, and the false-discovery rate (FDR) was calculated from the number of trials in which a participant indicated to see either people or cars, although none of them were present. A higher d' indicates better ability to discriminate between trials in which a target is present (signal) and trials in which a target is absent (noise). $d' = 0$ indicates that a participant is performing at chance levels.

We used bootstrapping to test for statistical significance. Ten thousand bootstrap re-samples were used to estimate within-subject and between-subject differences. All p values were corrected using FDR to control the probability of incorrectly rejecting the null hypotheses when doing multiple comparisons [22].

For the sake of completion, perceptual performance was also evaluated on four common statistical indicators: accuracy (number of correct predictions), precision (number of correct predictions divided by the number of all trials containing either people or cars), recall (number of correct predictions divided by the number of all trials that should have been identified as containing either people or cars), and the F1 score (harmonic mean of the precision and recall). Note that some of these are part of d'.

5 RESULTS

5.1 Effect of Scene Simplification Strategy on Perceptual Performance

Fig. 5 shows the perceptual performance of virtual patients as a function of the different scene simplification strategies. Subjects performed best using the segmentation algorithm ($d' : \mu = 1.13, \sigma = 1.01$). Performance based on saliency (d': $\mu = 0.07$, standard deviation $\sigma = 0.66, p < 0.001$), depth ($d' : \mu = 0.29, \sigma = 0.66, p < 0.001$), and combination ($d' : \mu = 1.01, \sigma = 0.91, p < 0.05$) was significantly worse. Saliency performed worse, followed by depth ($p < 0.01$) and the combination algorithm ($p < 0.001$). Interestingly, the combination algorithm was not able to benefit from the complementary information contributed by the individual saliency, depth, and segmentation algorithms.

These findings are further corroborated by other objective measures such as accuracy, precision, and recall (see Table 1) that reveal
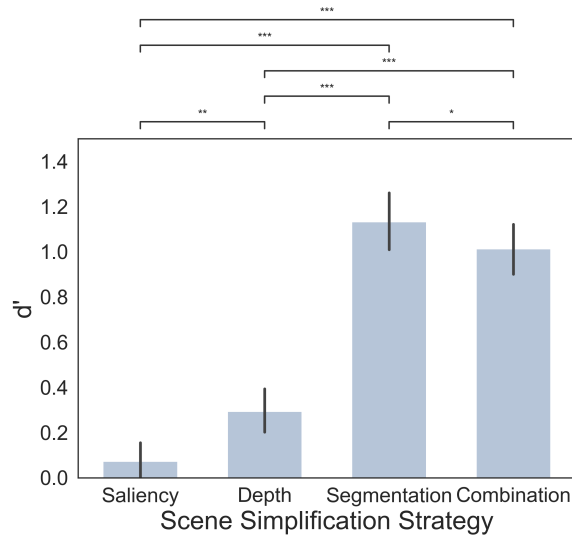
Figure 5: Effect of Scene Simplification Strategy. The p values are corrected using false-discovery rate (FDR), *<.05, **<.01, ***<.001. The error bars show the 95% confidence interval.

Condition	Accuracy	Precision	Recall	F1
Saliency	0.51	0.53	0.46	0.46
Depth	0.54	0.56	0.56	0.53
Segmentation	**0.68**	**0.73**	**0.63**	**0.68**
Combination	0.66	0.72	0.62	0.67

Table 1: Virtual patient's ability to identify people and cars in outdoor scenes using different scene simplification strategies (bold: best overall).

object segmentation as the most beneficial scene simplification strategy.

Subjects' confidence ratings in the segmentation condition ($\mu = 3.02, \sigma = 1.10$) and combination condition ($\mu = 2.96, \sigma = 1.08$) were both significantly higher than the those in the saliency condition ($\mu = 2.65, \sigma = 1.12$) and the depth condition ($\mu = 2.68 \sigma = 1.07$; all $p < 0.001$). No difference between saliency and depth condition was found ($p = 0.09$).

5.2 Effect of Phosphene Size and Elongation on Perceptual Performance

Fig. 6 shows the perceptual performance of virtual patients as a function of phosphene size (ρ) and elongation (λ). As expected, smaller phosphenes led to better perceptual performance, with $\rho = 100\,\mu m$ scoring significantly better ($d' : \mu = 0.81, \sigma = 1.02$) than $\rho = 300\,\mu m$ ($d' : \mu = 0.6, \sigma = 0.89, p = 0.03$) and $\rho = 500\,\mu m$ ($d' : \mu = 0.52, \sigma = 0.96, p = 0.02$). No significant difference in d' was found between the conditions with $\rho = 300\,\mu m$ and $\rho = 500\,\mu m$ ($p = 0.28$).
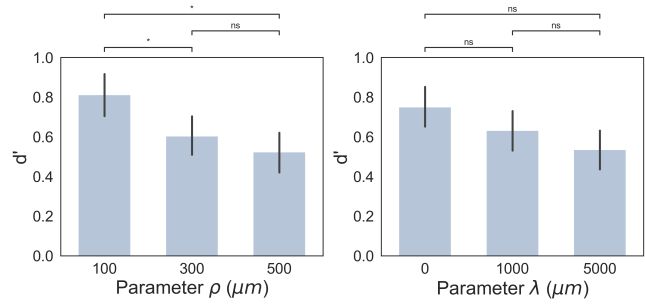

Figure 6: Sensitivity index (d') as a function of phosphene width (ρ) and length (λ). The p values were corrected using false-discovery rate (FDR): * $p < .05$, ** $p < .01$, *** $p < .001$. The error bars show the 95% confidence interval.
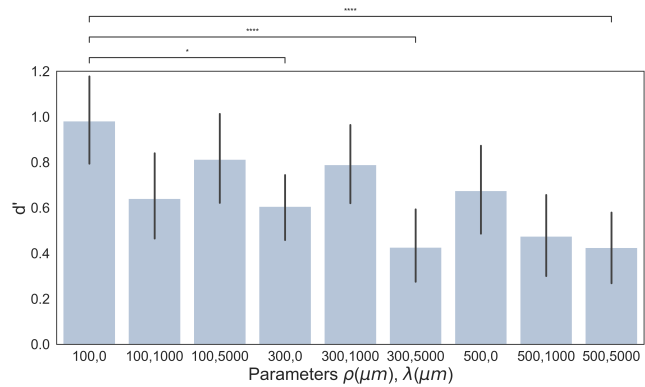

Figure 7: Sensitivity index (d') for each tested combination of phosphene width (ρ) and length (λ).The p values are corrected using false-discovery rate (FDR), *<.05, **<.01, ***<.001. The error bars show the 95% confidence interval.

ρ (μm)	λ (μm)	Accuracy	Precision	Recall	F1
100	0	***0.64***	***0.69***	0.53	0.56
	1000	0.59	0.58	***0.70***	***0.62***
	5000	0.62	0.68	0.63	0.60
300	0	0.59	0.61	**0.62**	**0.58**
	1000	**0.62**	**0.63**	0.58	**0.58**
	5000	0.57	0.58	0.56	0.55
500	0	**0.60**	**0.60**	0.63	**0.59**
	1000	0.58	0.56	**0.66**	**0.59**
	5000	0.57	0.55	**0.66**	**0.59**

Table 2: Virtual patient's ability to identify people and cars in outdoor scenes as a function of phosphene size (ρ) and elongation (λ; bold: best in box, italics: best overall).

A similar trend was evident with respect to phosphene elongation. Here, $\lambda = 0$ indicated circular phosphenes, similar to the SPV studies described in Section 3, and led to similar performance as $\rho = 100\,\mu m$ ($d' : \mu = 0.75, \sigma = 0.99$). And while there was a trend evident indicating that more elongated phosphenes may lead to
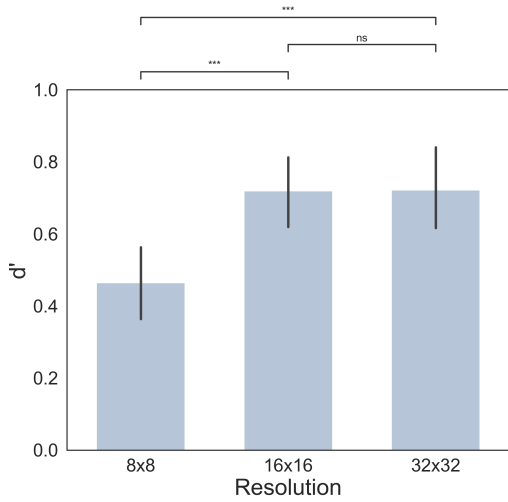
Figure 8: Sensitivity index (d') as a function of electrode grid size. The p values are corrected using false-discovery rate (FDR), *<.05, **<.01, ***<.001. The error bars show the 95% confidence interval.

Resolution	Accuracy	Precision	Recall	F1
8×8	0.57	0.59	0.57	0.54
16×16	**0.61**	**0.62**	0.64	**0.61**
32×32	**0.61**	**0.62**	**0.65**	**0.61**

Table 3: Virtual patient's ability to identify people and cars in outdoor scenes as a function of electrode grid size (bold: best overall).

poorer perceptual performance, this trend did not reach statistical significance ($\lambda = 1000\,\mu\text{m}, d' : \mu = 0.63, \sigma = 0.98, p > 0.05$; $\lambda = 5000\,\mu\text{m}, d' : \mu = 0.53, \sigma = 0.91, p > 0.05$).

This trend could later be confirmed by investigating d' across all nine model parameter conditions (see Fig. 7). Here we found a clear decreasing trend in d' as phosphene size(ρ) and phosphene elongation(λ) increased. However, notice that d' was positive in all conditions, indicating that subjects performed better than chance even when phosphenes were unusually large and elongated.

Similar patterns were found in all the other behavioral performance measurements (see Table 2). Overall, the highest accuracy, precision, recall, and F1 scores (italics) were achieved with the smallest tested phosphene size ($\rho = 100\,\mu\text{m}$), but not necessarily with the shortest phosphene length.

Unfortunately, the subjects' confidence ratings across conditions with different phosphene sizes (ρ) and phosphene elongations (λ) did not show any significant difference ($\rho = 100 : \mu = 2.74, \sigma = 1.19; \rho = 300 : \mu = 2.95, \sigma = 1.07; \rho = 500 : \mu = 2.80, \sigma = 1.04; \lambda = 0 : \mu = 2.60, \sigma = 1.14; \lambda = 1000 : \mu = 3.18, \sigma = 1.14; \lambda = 5000 : \mu = 2.71, \sigma = .93$, all $p > .05$).

5.3 Effect of Electrode Grid Size on Perceptual Performance

Fig. 8 shows the perceptual performance of virtual patients as a function of electrode grid size. As expected, performance improved as the number of electrodes in the array was increased from 8×8 ($d' : \mu = 0.47, \sigma = 0.87$) to 16×16 ($d' : \mu = 0.72, \sigma = 0.93, p < 0.001$). However, further increasing the number of electrodes to 32×32 did not measurably affect performance ($p = 0.37$).

This finding is again corroborated by accuracy, precision, recall, and F1 scores (Table 3), indicating virtually identical performance for 16×16 and 32×32.

Again, no significant difference in confidence ratings was found for different electrode array resolution ($8 \times 8 : \mu = 2.74, \sigma = 1.12; 16 \times 16 : \mu = 2.85, \sigma = 1.09; 32 \times 32 : \mu = 2.89, \sigma = 1.10$, all $p > 0.05$).

6 DISCUSSION

6.1 Object Segmentation May Support Scene Understanding

The present study provides the first side-by-side comparison of several deep learning–based scene simplification strategies for bionic vision. Considering a number of relevant implant configurations in combination with a psychophysically validated model of SPV, we identified object segmentation as the most promising image processing strategy to support outdoor scene understanding of virtual patients (see Fig. 5 and Table 1). This finding is consistent with recent studies indicating the potential utility of semantic segmentation for bionic vision [17, 33].

Object segmentation compared favorably with two other scene simplification strategies: based on visual saliency and monocular depth estimation. Whereas the saliency model struggled with the lighting conditions of the outdoor data set (often highlighting regions of increased contrast, and falling victim to shadows), the depth model often failed to highlight nearby obstacles. However, these models may prove their value in less structured test environments, where performance is less focused on semantic labeling and more concerned with the decluttering of complex scenes or the avoidance of nearby obstacles.

6.2 Increased Phosphene Size Impedes Perceptual Performance

To the best of our knowledge, this study is also the first to study SPV with a neurobiologically inspired, psychophysically validated model of phosphene shape [7]. Whereas previous studies assumed that phosphenes are isolated, focal spots of light [10, 11, 25, 30, 33], here we systematically evaluated perceptual performance across a wide range of common phosphene sizes (ρ) and elongation (λ). As expected, the best performance was achieved with small, circular phosphenes ($\rho = 100\,\mu\text{m}, \lambda = 0$; see Fig. 6), and increasing phosphene size and elongation negatively affected performance (Fig. 7). This finding suggests that future studies of SPV should take into account realistic phosphene shape when making predictions and drawing conclusions.

However, it is worth mentioning that the sensitivity index (d') remained positive in all tested conditions, indicating that subjects

performed better than chance even when phosphenes were unusually large and elongated. This result suggests that all tested scene simplification strategies enabled the virtual patients to perform above chance levels, no matter how degraded the SPV quality.

6.3 Increasing the Number of Electrodes Does Not Necessarily Improve Performance

As expected, perceptual performance improved as the size of the electrode grid was increased from 8×8 to 16×16. However, further increasing the number of electrodes to 32×32 did not measurably affect performance. This result is consistent with previous literature suggesting that number of electrodes is not the limiting factor in retinal implants [4, 8].

6.4 Limitations and Future Work

Although the present results demonstrate the utility of deep learning–based scene simplification for bionic vision, there are a number of limitations that should be addressed in future work.

First, in an effort to focus on scenes important for orientation and mobility, we limited our dataset to outdoor scenes. However, it would also be valuable to evaluate the performance of different scene simplification strategies on indoor scenarios. Because indoor scenes have different layouts and types of objects, the algorithms studied here might have different performances compared to outdoor scenes. For example, the saliency model might perform better in highlighting salient regions without the interference of light and shadow contrasts.

Second, to keep the perceptual judgments amenable to quantitative performance measures, we limited the current study to a simple detection task involving common semantic object categories (i.e., people and cars). This might explain the superior performance of the semantic segmentation algorithm, which operates with semantic labels. In contrast, the depth and saliency algorithms might prove more valuable when applied to open-ended navigation tasks. In the future, we plan to conduct such SPV studies in immersive virtual reality (VR) to gain more comprehensive insight into the behavioral performance of virtual patients.

Third, the present study should be understood as a first step towards the ultimate goal of creating a retinal implant supported by deep learning–based image preprocessing. Such a device would require all processing to happen in real time at the edge. One solution could come in the form of low-power, low-latency neuromorphic hardware coupled with an event-based vision sensor. Future iterations of this work may include end-to-end training of scene simplification strategies fitted to a specific implant technology or even an individual patient. Overall this work has the potential to drastically improve the utility of prosthetic vision for people blinded from retinal degenerative diseases.

7 DATA AVAILABILITY

All raw video sequences (original and preprocessed) are available on the Open Science Framework (https://osf.io/s2udz). SPV models were based on the pulse2percept Python package [5]. Code used to implement the scene simplification strategies is available on GitHub (https://github.com/bionicvisionlab/2021-han-scene-simplification, v0.1).

REFERENCES

[1] W. I. Al-Atabany, T. Tong, and P. A. Degenaar. 2010. Improved content aware scene retargeting for retinitis pigmentosa patients. *Biomed Eng Online* 9 (Sept. 2010), 52. https://doi.org/10.1186/1475-925X-9-52

[2] Lauren N. Ayton, Nick Barnes, Gislin Dagnelie, Takashi Fujikado, Georges Goetz, Ralf Hornig, Bryan W. Jones, Mahiul M. K. Muqit, Daniel L. Rathbun, Katarina Stingl, James D. Weiland, and Matthew A. Petoe. 2020. An update on retinal prostheses. *Clinical Neurophysiology* 131, 6 (June 2020), 1383–1398. https://doi.org/10.1016/j.clinph.2019.11.029

[3] Lauren N. Ayton, Peter J. Blamey, Robyn H. Guymer, Chi D. Luu, David A. X. Nayagam, Nicholas C. Sinclair, Mohit N. Shivdasani, Jonathan Yeoh, Mark F. McCombe, Robert J. Briggs, Nicholas L. Opie, Joel Villalobos, Peter N. Dimitrov, Mary Varsamidis, Matthew A. Petoe, Chris D. McCarthy, Janine G. Walker, Nick Barnes, Anthony N. Burkitt, Chris E. Williams, Robert K. Shepherd, Penelope J. Allen, and for the Bionic Vision Australia Research Consortium. 2014. First-in-Human Trial of a Novel Suprachoroidal Retinal Prosthesis. *PLOS ONE* 9, 12 (Dec. 2014), e115239. https://doi.org/10.1371/journal.pone.0115239 Publisher: Public Library of Science.

[4] Matthew R. Behrend, Ashish K. Ahuja, Mark S. Humayun, Robert H. Chow, and James D. Weiland. 2011. Resolution of the Epiretinal Prosthesis is not Limited by Electrode Size. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 19, 4 (Aug. 2011), 436–442. https://doi.org/10.1109/TNSRE.2011.2140132

[5] M. Beyeler, G. M. Boynton, I. Fine, and A. Rokem. 2017. pulse2percept: A Python-based simulation framework for bionic vision. In *Proceedings of the 16th Science in Python Conference*, K. Huff, D. Lippa, D. Niederhut, and M. Pacer (Eds.). 81–88. https://doi.org/10.25080/shinma-7f4c6e7-00c

[6] Michael Beyeler, Geoffrey M. Boynton, Ione Fine, and Ariel Rokem. 2019. Model-Based Recommendations for Optimal Surgical Placement of Epiretinal Implants. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (Lecture Notes in Computer Science)*, Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (Eds.). Springer International Publishing, 394–402. https://doi.org/10.1007/978-3-030-32254-0_44

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

[8] M. Beyeler, A. Rokem, G. M. Boynton, and I. Fine. 2017. Learning to see again: biological constraints on cortical plasticity and the implications for sight restoration technologies. *J Neural Eng* 14, 5 (June 2017), 051003. https://doi.org/10.1088/1741-2552/aa795e

[9] Justin R. Boyle, Anthony J. Maeder, and Wageeh W. Boles. 2008. Region-of-interest processing for electronic visual prostheses. *Journal of Electronic Imaging* 17, 1 (Jan. 2008), 013002. https://doi.org/10.1117/1.2841708 Publisher: International Society for Optics and Photonics.

[10] S. C. Chen, G. J. Suaning, J. W. Morley, and N. H. Lovell. 2009. Simulating prosthetic vision: I. Visual models of phosphenes. *Vision Research* 49, 12 (June 2009), 1493–506.

[11] G. Dagnelie, P. Keane, V. Narla, L. Yang, J. Weiland, and M. Humayun. 2007. Real and virtual mobility performance in simulated prosthetic vision. *J Neural Eng* 4, 1 (March 2007), S92–101. https://doi.org/10.1088/1741-2560/4/1/S11

[12] Cordelia Erickson-Davis and Helma Korzybska. 2020. What do blind people "see" with retinal prostheses? Observations and qualitative reports of epiretinal implant users. *bioRxiv* (Feb. 2020), 2020.02.03.932905. https://doi.org/10.1101/2020.02.03.932905

[13] Laura Ferlauto, Marta Jole Ildelfonsa Airaghi Leccardi, Naïg Aurelia Ludmilla Chenais, Samuel Charles Antoine Gilliéron, Paola Vagni, Michele Bevilacqua, Thomas J. Wolfensberger, Kevin Sivula, and Diego Ghezzi. 2018. Design and validation of a foldable and photovoltaic wide-field epiretinal prosthesis. *Nature Communications* 9, 1 (March 2018), 1–15. https://doi.org/10.1038/s41467-018-03386-7

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

[15] I. Fine and G. M. Boynton. 2015. Pulse trains to percepts: the challenge of creating a perceptually intelligible world with sight recovery technologies. *Philos Trans R*

Soc Lond B Biol Sci 370, 1677 (Sept. 2015), 20140208. https://doi.org/10.1098/rstb .2014.0208

[16] C. Godard, O. M. Aodha, M. Firman, and G. Brostow. 2019. Digging Into Self-Supervised Monocular Depth Estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 3827–3837. https://doi.org/10.1109/ICCV.2019. 00393 ISSN: 2380-7504.

[17] Lachlan Horne, Jose Alvarez, Chris McCarthy, Mathieu Salzmann, and Nick Barnes. 2016. Semantic labeling for prosthetic vision. *Computer Vision and Image Understanding* 149 (Aug. 2016), 113–125. https://doi.org/10.1016/j.cviu.2016.02.0 15

[18] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv:1610.01563 [cs, q-bio, stat]* (Oct. 2016). http://arxiv.org/abs/1610.01563 arXiv: 1610.01563.

[19] Lago, Miguel. 2021. SimplePhy: An open-source tool for quick online perception experiments. *Behavior Research Methods* (2021). https://doi.org/10.3758/s13428-020-01515-z

[20] Heng Li, Tingting Han, Jing Wang, Zhuofan Lu, Xiaofei Cao, Yao Chen, Liming Li, Chuanqing Zhou, and Xinyu Chai. 2017. A real-time image optimization strategy based on global saliency detection for artificial retinal prostheses. *Information Sciences* 415, Supplement C (Nov. 2017), 1–18. https://doi.org/10.1016/j.ins.2017 .06.014

[21] Heng Li, Xiaofan Su, Jing Wang, Han Kan, Tingting Han, Yajie Zeng, and Xinyu Chai. 2018. Image processing strategies based on saliency segmentation for object recognition under simulated prosthetic vision. *Artificial Intelligence in Medicine* 84 (Jan. 2018), 64–78. https://doi.org/10.1016/j.artmed.2017.11.001

[22] Junning Li, Yonggang Shi, and Arthur W. Toga. 2015. Controlling False Discovery Rate in Signal Space for Transformation-Invariant Thresholding of Statistical Maps. *Information processing in medical imaging : proceedings of the ... conference* 9123 (July 2015), 125–136. https://doi.org/10.1007/978-3-319-19992-4_10

[23] P. Lieby, N. Barnes, C. McCarthy, Nianjun Liu, H. Dennett, J. G. Walker, V. Botea, and A. F. Scott. 2011. Substituting depth for intensity and real-time phosphene rendering: Visual navigation under low vision conditions. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, Boston, MA, 8017–8020. https://doi.org/10.1109/IEMBS.2011.6091977

[24] H. Lorach, G. Goetz, R. Smith, X. Lei, Y. Mandel, T. Kamins, K. Mathieson, P. Huie, J. Harris, A. Sher, and D. Palanker. 2015. Photovoltaic restoration of sight with high visual acuity. *Nat Med* 21, 5 (May 2015), 476–82. https://doi.org/10.1038/nm.3851

[25] Wen Lik Dennis Lui, Damien Browne, Lindsay Kleeman, Tom Drummond, and Wai Ho Li. 2011. Transformative reality: Augmented reality for visual prostheses. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. 253–254. https://doi.org/10.1109/ISMAR.2011.6092402 ISSN: null.

[26] Y. H. Luo and L. da Cruz. 2016. The Argus((R)) II Retinal Prosthesis System. *Prog Retin Eye Res* 50 (Jan. 2016), 89–107. https://doi.org/10.1016/j.preteyeres.2015.09 .003

[27] Yvonne H.-L. Luo, Joe Jiangjian Zhong, Monica Clemo, and Lyndon da Cruz. 2016. Long-term Repeatability and Reproducibility of Phosphene Characteristics in Chronically Implanted Argus II Retinal Prosthesis Subjects. *American Journal of Ophthalmology* 170 (Oct. 2016), 100–109. https://doi.org/10.1016/j.ajo.2016.07.021 Publisher: Elsevier.

[28] Chris McCarthy, Janine G. Walker, Paulette Lieby, Adele Scott, and Nick Barnes. 2014. Mobility and low contrast trip hazard avoidance using augmented depth. *Journal of Neural Engineering* 12, 1 (Nov. 2014), 016003. https://doi.org/10.1088/ 1741-2560/12/1/016003 Publisher: IOP Publishing.

[29] N. Parikh, L. Itti, and J. Weiland. 2010. Saliency-based image processing for retinal prostheses. *Journal of Neural Engineering* 7, 1 (Jan. 2010), 016006. https: //doi.org/10.1088/1741-2560/7/1/016006 Publisher: IOP Publishing.

[30] Alejandro Perez-Yus, Jesus Bermudez-Cameo, Gonzalo Lopez-Nicolas, and Jose J. Guerrero. 2017. Depth and Motion Cues With Phosphene Patterns for Prosthetic Vision. 1516–1525. http://openaccess.thecvf.com/content_ICCV_2017_worksho ps/w22/html/Perez-Yus_Depth_and_Motion_ICCV_2017_paper.html

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

[32] Roksana Sadeghi, Arathy Kartha, Michael P. Barry, Paul Gibson, Avi Caspi, Arup Roy, and Gislin Dagnelie. 2019. Thermal and Distance image filtering improve independent mobility in Argus II retinal implant. *Journal of Vision* 19, 15 (Dec. 2019), 23–23. https://doi.org/10.1167/19.15.23 Publisher: The Association for Research in Vision and Ophthalmology.

[33] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Josechu J. Guerrero. 2019. Indoor Scenes Understanding for Visual Prosthesis with Fully Convolutional Networks. In *VISIGRAPP*. https://doi.org/10.5220/0007257602180225

[34] A. J. Simpson and M. J. Fitter. 1973. What is the best index of detectability? *Psychological Bulletin* 80, 6 (1973), 481–488. https://doi.org/10.1037/h0035203 Place: US Publisher: American Psychological Association.

[35] A. Stacey, Y. Li, and N. Barnes. 2011. A salient information processing system for bionic eye with application to obstacle avoidance. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 5116–5119.

[36] K. Stingl, K. U. Bartz-Schmidt, D. Besch, A. Braun, A. Bruckmann, F. Gekeler, U. Greppmaier, S. Hipp, G. Hortdorfer, C. Kernstock, A. Koitschev, A. Kusnyerik, H. Sachs, A. Schatz, K. T. Stingl, T. Peters, B. Wilhelm, and E. Zrenner. 2013. Artificial vision with wirelessly powered subretinal electronic implant alpha-IMS. *Proc Biol Sci* 280, 1757 (April 2013), 20130077. https://doi.org/10.1098/rspb.2013.0077

[37] Victor Vergnieux, Marc J.-M. Macé, and Christophe Jouffrais. 2017. Simplification of Visual Rendering in Simulated Prosthetic Vision Facilitates Navigation. *Artificial Organs* 41, 9 (Sept. 2017), 852–861. https://doi.org/10.1111/aor.12868 Publisher: John Wiley & Sons, Ltd.

[38] Jing Wang, Heng Li, Weizhen Fu, Yao Chen, Liming Li, Qing Lyu, Tingting Han, and Xinyu Chai. 2016. Image Processing Strategies Based on a Visual Saliency Model for Object Recognition Under Simulated Prosthetic Vision. *Artificial Organs* 40, 1 (2016), 94–100. https://doi.org/10.1111/aor.12498 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/aor.12498.

[39] J. D. Weiland and M. S. Humayun. 2005. A biomimetic retinal stimulating array. *IEEE Engineering in Medicine and Biology Magazine* 24, 5 (Sept. 2005), 14–21. https://doi.org/10.1109/MEMB.2005.1511496 Conference Name: IEEE Engineering in Medicine and Biology Magazine.

[40] J. D. Weiland, N. Parikh, V. Pradeep, and G. Medioni. 2012. Smart image processing system for retinal prosthesis. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 300–303. https://doi.org/10.1109/ EMBC.2012.6345928 ISSN: 1558-4615.

[41] R. G. H. Wilke, G. Khalili Moghadam, N. H. Lovell, G. J. Suaning, and S. Dokos. 2011. Electric crosstalk impairs spatial resolution of multi-electrode arrays in retinal implants. *Journal of Neural Engineering* 8, 4 (June 2011), 046016. https: //doi.org/10.1088/1741-2560/8/4/046016

[42] Wu, Yuxin, Kirillov, Alexander, Massa, Francisco, Lo, Wan-Yen, and Girshick, Ross. 2019. Detectron2. https://github.com/facebookresearch/detectron2

[43] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2016. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442* (2016).

[44] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.