Extended Abstract Track

Beyond Pixels: A Differentiable Pipeline for Probing Neuronal Selectivity in 3D

Editors: List of editors' names

Abstract

Visual perception relies on inference of 3D scene properties such as shape, pose, and lighting. To understand how visual sensory neurons enable robust perception, it is crucial to characterize their selectivity to such physically interpretable factors. However, current approaches mainly operate on 2D pixels, making it difficult to isolate selectivity for physical scene properties. To address this limitation, we introduce a differentiable rendering pipeline that optimizes deformable meshes to obtain MEIs directly in 3D. The method parameterizes mesh deformations with radial basis functions and learns offsets and scales that maximize neuronal responses while enforcing geometric regularity. Applied to models of monkey area V4, our approach enables probing neuronal selectivity to interpretable 3D factors such as pose and lighting. This approach bridges inverse graphics with systems neuroscience, offering a way to probe neural selectivity with physically grounded, 3D stimuli beyond conventional pixel-based methods.

Keywords: Visual cortex, Macaque V4, Neural selectivity, Differentiable rendering, Inverse graphics, 3D scene representation, Maximally Exciting Inputs (MEIs)

1. Introduction

Primate visual sensory neurons are selective to a wide range of features, from simple edges and luminance to complex attributes such as shape, color, texture, and lighting conditions (Hubel and Wiesel, 1968; Hegdé and Van Essen, 2000; Pasupathy, 2006; Arcizet et al., 2009; Kim et al., 2019). Area V4, for example, contains neurons tuned to curvature, material properties, and 3D shape cues, reflecting the growing complexity of representations along the ventral stream (Pasupathy et al., 2020; Roe et al., 2012; Pasupathy et al., 2019; Srinath et al., 2021). Recent work has shown that optimized input images—such as Maximally Exciting Inputs (MEIs)—can reveal aspects of neuronal selectivity and invariances (Bashivan et al., 2019; Walker et al., 2019; Ding et al., 2023; Franke et al., 2022; Bashiri et al., 2025). However, these approaches operate in 2D pixel space, entangling multiple visual factors and making it difficult to isolate interpretable 3D properties. To study whether and how neurons show selectivity to physically interpretable factors, we would ideally synthesize the underlying 3D scene factors directly and observe the resulting neuronal responses.

Here, we introduce a differentiable rendering-based pipeline to obtain 3D Maximally Exciting MEIs (3D-MEI): an approach that optimizes directly in 3D object space (meshes, textures, poses) through a differentiable renderer. This grounds selectivity in physically realizable structure and enables systematic tests of tuning and invariances—such as tolerance to viewpoint, lighting, or material—that pixel-based MEIs do no capture explicitly. Testing our approach on modeled V4 neurons of the macaque visual cortex, we ❶ show that our pipeline can generate physically meaningful objects that resemble the pixel based MEI and strongly drive the neuronal response, ❷ demonstrate the potential of our approach to study neuron encoding to complex scene parameters such as lighting conditions.

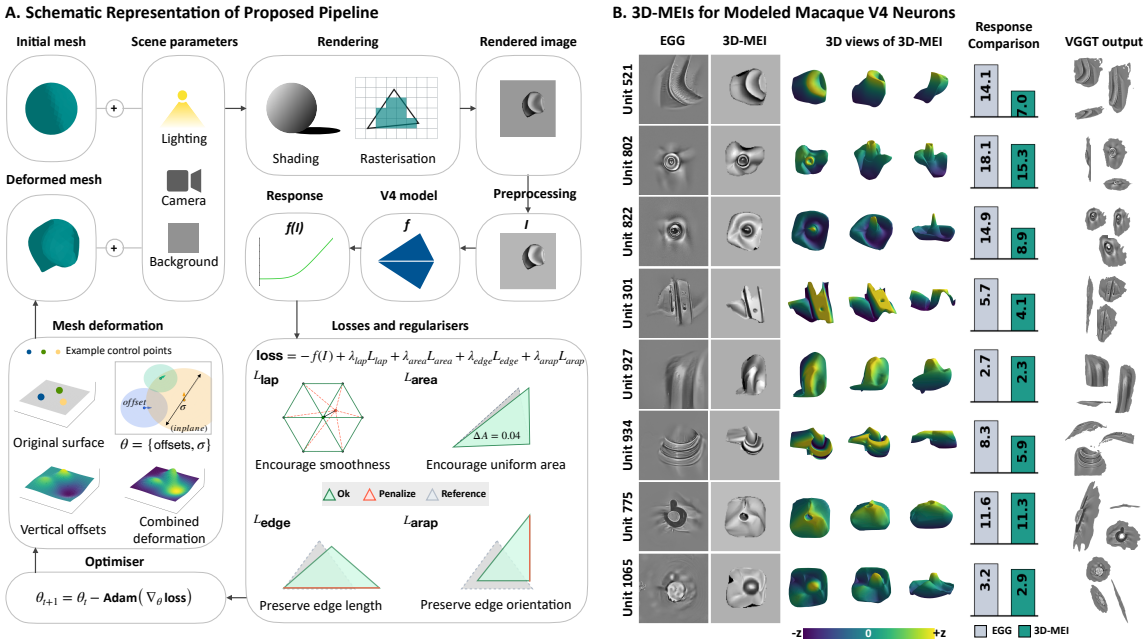

Figure 1: **A.** Schematic overview of the differentiable rendering pipeline. Starting from an initial mesh and scene parameters (lighting, camera, background), the pipeline iteratively deforms the mesh using radial basis function offsets and optimizes scene parameters to maximize model responses. Losses and regularizers enforce smoothness, uniform surface areas, edge-length preservation, and locally rigid deformations. **B.** 3D Maximally exciting images (3D-MEIs) generated for multiple selected model macaque V4 neurons. For each selected neuron (unit), we show the pixel based EGG MEI and the 3D-MEI, as well as their 3D views. Right: Bar plots comparing activation values for EGG and 3D-MEI. A 2D-to-3D model (VGGT) applied to EGG MEIs fails to generate meaningful 3D structures.

2. Methods

Our pipeline (Fig. 1A) operates mainly in 3D object space, deforming an initial mesh to optimize its shape with respect to a neuron's response. The neuron's response is obtained using an image-computable response-predictive model. Given a response-predictive (encoding) model f, a differentiable renderer R, and a parameterized 3D mesh $M = (V, F)$ with vertices V and faces F, the pipeline optimizes V such that the rendered projection $I = R(M)$ maximizes the neural model output $f(I)$. For f, we use deep encoding model from Pierzchlewicz et al. (2023), which contains a task driven core with Gaussian readout (see Appendix C for details), and use PyTorch3D (Ravi et al., 2020) as differentiable renderer R. Meshes are initialized as either a sheet or a sphere, and are placed at the origin while the camera is fixed overhead at a height z at (0, 0, z). A fixed lighting is provided by a point light source positioned near the camera (see Fig. 3D) and the scene is rendered against a uniform gray background. I is normalized to match training mean and std, and

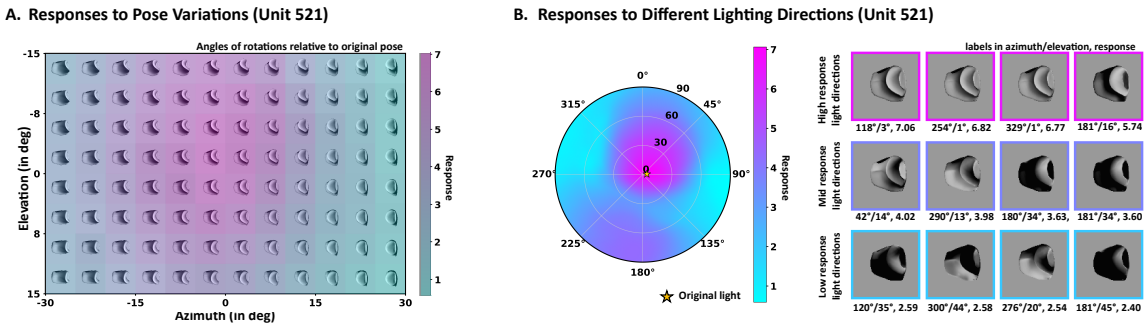

Figure 2: Response profile of Unit 521 to variations in 3D pose and lighting direction. **A.** Heatmap of model responses to pose variations (azimuth × elevation) relative to the original view, showing tolerance to moderate rotations. **B.** Light-direction tuning of the same unit. Responses are shown for selected light positions on the hemisphere, grouped into high, mid, and low response sets (labels indicate azimuth/elevation in degrees and response values).

scaled to have a fixed norm of 25 which was used by Pierzchlewicz et al. (2023). The current pipeline does not optimize surface texture and color of objects.

Mesh Deformation via RBF Kernels To discourage uncontrolled mesh deformations, we use a set of K radial basis function (RBF) kernels (Buhmann, 2003) centered at control points c_k distributed uniformly over the mesh. Each RBF has a learnable scale σ_k and offset δ_k, which control the scale and the direction of the displacement, respectively. The deformation Δv at any vertex v is then expressed as a weighted sum over all kernel offsets $\Delta v = \sum_{k=1}^{K} \delta_k \cdot \exp\left(-\|v - c_k\|^2/(2\sigma_k^2)\right)$. Here, we use all vertices as control points.

Loss Functions and regularization Our goal is to maximize the neuronal response. However, unconstrained optimization can yield degenerate meshes. To prevent this, we incorporate a set of geometric regularizers. Specifically, we use ❶ Laplacian smoothing ($\mathcal{L}_{lap}$) (Sorkine et al., 2004) to encourage smooth surface curvature, ❷ uniform edge length loss ($\mathcal{L}_{edge}$) to penalize large variation in edge lengths of the triangles in the mesh, ❸ triangle area loss ($\mathcal{L}_{area}$) to encourage uniform triangle areas to avoid collapse, and finally, ❹ As-Rigid-As-Possible Loss ($\mathcal{L}_{arap}$) adapted from Sorkine and Alexa (2007) to maintain local surface geometry by penalizing non-rigid deformations. Details on the loss function and regularizers can be found in Appendix B. The total loss and optimization objective is $\mathcal{L} = -f(I) + \lambda_{lap} \cdot \mathcal{L}_{lap} + \lambda_{edge} \cdot \mathcal{L}_{edge} + \lambda_{area} \cdot \mathcal{L}_{area} + \lambda_{arap} \cdot \mathcal{L}_{arap}$, which we optimize using the Adam optimizer (Kingma and Ba, 2014). λ_* coefficients were found by a grid search.

3. Experiments and Results

First, we verified that our pipeline can recognizably reconstruct known 3D shapes (Fig. 3A), as well as models of simple and complex cells whose optimal stimuli are pre-defined Gabor patterns (Fig. 3C). Our pipeline successfully recovers the underlying Gabor structures. For

a complex cell, which is phase invariant, the optimization also yields different phase variants of the Gabor under different initialization seeds, which shows the potential of the pipeline to capture invariances. Next, we apply our pipeline to a set of V4 neurons modeled via deep encoding model (Pierzchlewicz et al., 2023). We manually selected neurons based on their visual appearance of their pixel optimized MEIs and picked neurons that exhibited geometric, shape-like features.

The synthesized 3D-MEIs closely resemble pixel based MEIs synthesized with energy guided diffusion (EGG, Pierzchlewicz et al., 2023, Fig. 1B). Compared to pixel MEIs, 3D-MEIs elicit lower activation in the model neurons (Fig. 1B right). We attribute this to the fact that the current 3D-MEIs do not contain textures, whereas pixel optimized MEIs have texture information. This can limit the maximum activation value achievable with the pipeline since V4 neurons are believed to encode both texture and shape (Kim et al., 2019; Pasupathy et al., 2020). While our approach can be extended to optimize texture and color, we chose to focus on shape optimization for now. We also compared our method against a 2D-to-3D model such as VGGT (Wang et al., 2025) that is trained to "lift" images of objects to 3D point clouds, by applying it to pixel optimized MEIs. However, the resulting geometry lacked meaningful 3D structure (see Fig. 1B).

Exploring Tuning Properties via Scene Manipulation Once the mesh is optimized, the scene can be systematically manipulated along physically meaningful axes, enabling interpretable exploration of neuronal tuning properties in 3D space. Importantly, these tuning properties cannot easily be explored in images space, since pixel manipulations do not directly capture physical transformations.

To showcase this possibility, we explore tuning of an example neuron along different poses and lighting directions separately, varying one at a time while keeping the other scene parameters fixed. To assess tuning along pose variations, we systematically rotated the object around azimuth and elevation axes and recorded the model responses (Fig. 2A). To assess tuning along lighting directions, we move the point light source along the forward facing half dome and record the model responses for each location of point light (Fig. 2B polar heatmap). For the selected neuron, we observe a strong near-frontal preference, skewed toward the upper-right quadrant of the dome. We also visualize the rendered scenes in three categories: high response, mid response and low response. The curved geometry—a tuning feature in area V4—is visible in almost all light directions but the neuron only prefers certain shadings (Fig. 2B right side) consistent with a shape-light-direction combination that resemble this units preferred stimulus.

Summary

We present a proof-of-concept showing that differentiable rendering can probe neural representations in ways inaccessible to pixel-based MEIs. By operating directly in 3D space, our approach enables the dissection of invariances and sensitivities in relation to physically grounded factors such as shape, curvature, and illumination. Our pipeline opens up a new way to probe the inner workings of visual representations, and disentangle tuning of neurons to geometry from other visual features such as texture. We believe that by optimizing physically plausible 3D stimuli, we can move beyond the limitations of pixel space to understand biological vision.

References

Fabrice Arcizet, Christophe Jouffrais, and Pascal Girard. Coding of shape from shading in area v4 of the macaque monkey. *BMC neuroscience*, 10(1):140, 2009.

Mohammad Bashiri, Luca Baroni, Ján Antolík, and Fabian H Sinz. Learning and aligning single-neuron invariance manifolds in visual cortex. In *The Thirteenth International Conference on Learning Representations*, 2025.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

Martin D. Buhmann. *Radial Basis Functions: Theory and Implementations*. Cambridge University Press, 2003.

Zhiwei Ding, Dat T Tran, Kayla Ponder, Erick Cobos, Zhuokun Ding, Paul G Fahey, Eric Wang, Taliah Muhammad, Jiakun Fu, Santiago A Cadena, Stelios Papadopoulos, Saumil Patel, Katrin Franke, Jacob Reimer, Fabian H Sinz, Alexander S Ecker, Xaq Pitkow, and Andreas S Tolias. Bipartite invariance in mouse primary visual cortex. *bioRxiv*, page 2023.03.15.532836, March 2023.

Katrin Franke, Konstantin F Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H Sinz, and Andreas S Tolias. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, October 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jay Hegdé and David C Van Essen. Selectivity for complex shapes in primate visual area v2. *The Journal of Neuroscience*, 20(5):RC61, 2000.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

Taekjun Kim, Wyeth Bair, and Anitha Pasupathy. Neural coding for shape and texture in macaque area v4. *Journal of Neuroscience*, 39(24):4760–4774, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay K Jagadish, Eric Wang, Edgar Y Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S Tolias, et al. Generalization in data-driven models of primary visual cortex. *BioRxiv*, pages 2020–10, 2020.

Anitha Pasupathy. Neural basis of shape representation in the primate brain. *Progress in brain research*, 154:293–313, 2006.

Anitha Pasupathy, Taekjun Kim, and Dina V Popovkina. Object shape and surface properties are jointly encoded in mid-level ventral visual cortex. *Current opinion in neurobiology*, 58:199–208, 2019.

Anitha Pasupathy, Dina V Popovkina, and Taekjun Kim. Visual functions of primate area v4. *Annual review of vision science*, 6(1):363–385, 2020.

Pawel Pierzchlewicz, Konstantin Willeke, Arne Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, et al. Energy guided diffusion for generating neurally exciting images. *Advances in Neural Information Processing Systems*, 36:32574–32601, 2023.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Thomas Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Pytorch3d: An open-source library for 3d deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

Anna W Roe, Leonardo Chelazzi, Charles E Connor, Bevil R Conway, Ichiro Fujita, Jack L Gallant, Haidong Lu, and Wim Vanduffel. Toward a unified theory of visual area v4. *Neuron*, 74(1):12–29, 2012.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. *Symposium on Geometry Processing*, 2007.

Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and Hans-Peter Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 175–184, 2004.

Ramanujan Srinath, Alexandriya Emonds, Qingyang Wang, Augusto A Lempel, Erika Dunn-Weiss, Charles E Connor, and Kristina J Nielsen. Early emergence of solid shape coding in natural and deep network vision. *Current Biology*, 31(1):51–65, 2021.

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065, 2019.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

Appendix A. Supplementary Results

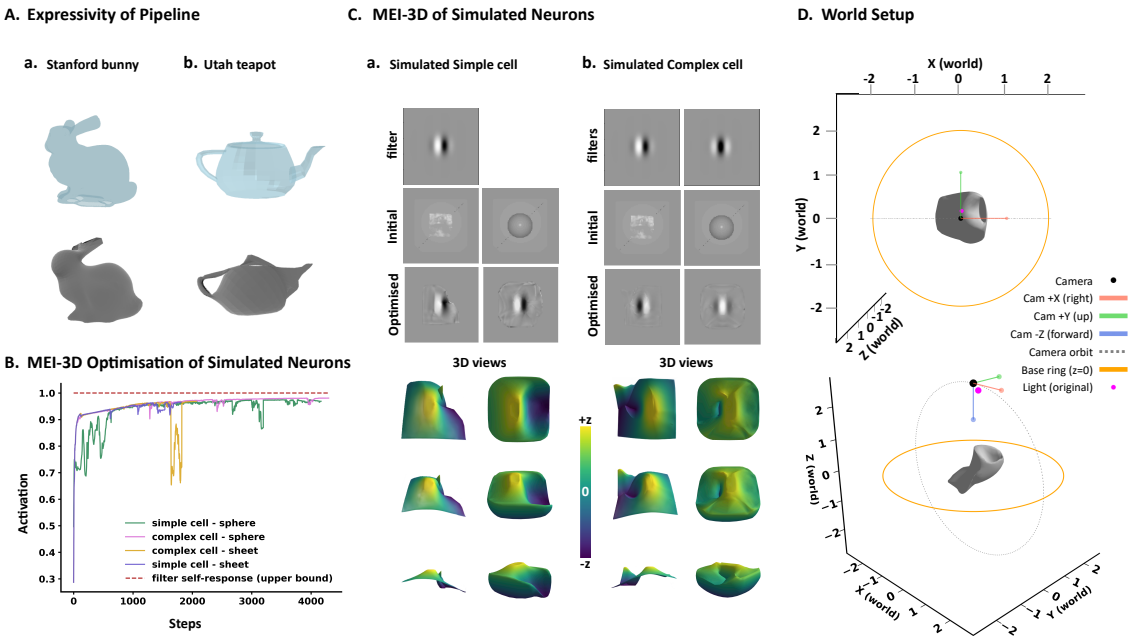

Figure 3: Validation of the pipeline on simulated data. **A.** RBF-based mesh deformation produces complex targets such as the Stanford Bunny (from a sphere) and Utah Teapot (from a torus) by minimizing Chamfer loss (Ravi et al., 2020) between the input and target meshes. **B.** Optimization curves for simulated simple and complex cells with "sheet" and "sphere" initializations approach to the upper bound of 1. **C.** 3D-MEI results for a simulated simple cell **a.** and complex cell **b.**. Outputs recover the ground-truth filters; for the phase-invariant complex cell, different initializations yield Gabor-like patterns in distinct phases (example shown). **D.** World setup of the differentiable renderer R, showing world and camera axes. The base ring (z=0) marks the base of half-dome used in light-direction tuning.

Appendix B. Details on Loss Function and Regularizers

Given a response-predictive model f, a differentiable renderer R, and a parameterized 3D mesh $M = (V, F)$ with vertices V, (where $V = \{v_i\}_{i=1}^{|V|}$, $v_i \in \mathbb{R}^3$) and faces F, (where $F = \{f_t = (i, j, k)\}_{t=1}^{|F|}$, $f_t \in \{1, \ldots, |V|\}^3$), the pipeline optimizes V such that the rendered projection $I = R(M)$ maximizes the model output $f(I)$. The total loss and optimization objective of our pipeline is

$$\mathcal{L} = -f(I) + \lambda_{lap} \cdot \mathcal{L}_{lap} + \lambda_{edge} \cdot \mathcal{L}_{edge} + \lambda_{area} \cdot \mathcal{L}_{area} + \lambda_{arap} \cdot \mathcal{L}_{arap}$$

where, $\mathcal{L}_{lap}$ is the Laplacian smoothing loss, $\mathcal{L}_{edge}$ is uniform edge length loss, $\mathcal{L}_{area}$ is triangle area loss and $\mathcal{L}_{arap}$ is the simplified and adapted version of the classical As Rigid as Possible loss.

We use Laplacian smoothing loss $\mathcal{L}_{lap}$ (Sorkine et al., 2004) to enforce local smoothness of the surface by penalizing deviations of each vertex v_i from the mean position of its 1-ring neighbors $\mathcal{N}(i)$.

$$\mathcal{L}_{\text{lap}} = \frac{1}{|V|} \sum_{i=1}^{|V|} \left\| v_i - \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} v_j \right\|^2 \tag{1}$$

We define uniform edge length loss $\mathcal{L}_{\text{edge}}$ to discourage edges from becoming disproportionately long or short compared to others. Without it, the mesh could stretch, shear, or collapse. By penalizing variance in edge lengths, the mesh maintains spatial regularity and stable deformations.

$$\mathcal{L}_{\text{edge}} = \frac{1}{|E|} \sum_{(i,j) \in E} \Big(\|v_i - v_j\| - \bar{\ell} \Big)^2, \quad \bar{\ell} = \frac{1}{|E|} \sum_{(i,j) \in E} \|v_i - v_j\| \tag{2}$$

where, E is the edge set extracted from F.

Similarly, triangle area loss $\mathcal{L}_{\text{area}}$ enforces uniformity in the areas of all the triangular faces. As the vertices move, some triangles can become skinny or even degenerate. By penalizing the variance in the area of triangles, $\mathcal{L}_{\text{area}}$ promotes evenly sized triangles, preserving mesh quality and preventing local collapse.

$$\mathcal{L}_{\text{area}} = \frac{1}{|F|} \sum_{t \in F} \big(A_t - \bar{A} \big)^2, \quad A_t = \tfrac{1}{2} \left\| (v_j - v_i) \times (v_k - v_i) \right\|, \quad \bar{A} = \frac{1}{|F|} \sum_{t \in F} A_t \tag{3}$$

While uniform edge length loss and triangle area losses regularize local distances and face sizes, they do not directly constrain changes in edge orientation. To address this, we use a *direction-preserving* variant of the As-Rigid-As-Possible (ARAP) energy (Sorkine and Alexa, 2007) that penalizes deviations of each deformed edge from its original direction. Specifically, each edge vector $v_i - v_j$ is projected onto its initial unit vector $\widehat{e}_{ij}^{(0)} = \widehat{v_i^{(0)} - v_j^{(0)}}$, and we penalize the residual orthogonal component:

$$\mathcal{L}_{\text{arap}} = \frac{1}{|E|} \sum_{(i,j) \in E} \left\| (v_i - v_j) - \text{proj}_{\widehat{e}_{ij}^{(0)}} (v_i - v_j) \right\|^2 \tag{4}$$

This preserves the original orientation of the edges while ignoring changes in magnitude.

Appendix C. Response predictive V4 model

We use the "Gaussian model" defined in Pierzchlewicz et al. (2023) as the image computable encoding model for macaque V4 neurons. The model contains a pre-trained robust ResNet50

($L_2, \varepsilon = 0.1$) (He et al., 2016; Salman et al., 2020) core and a Gaussian readout (Lurz et al., 2020). The model uses 3 layers with 1024 channels, resulting in a 1024 dimension feature space, followed by batch normalization(Ioffe and Szegedy, 2015) and ReLU non-linearity. The EGG based pixel optimized MEIs shown in Fig 1 is also generated from this model.