Parallel Backpropagation for Shared-Feature Visualization

Alexander Lappe[1,2] **Anna Bognár**[3] **Ghazaleh Ghamkhari Nejad**[3]
Albert Mukovskiy[1] **Lucas Martini**[1,2] **Martin A. Giese**[1] **Rufin Vogels**[3]
[1]Hertie Institute, University Clinics Tübingen [2]IMPRS-IS [3] KU Leuven
alexander.lappe@uni-tuebingen.de

Abstract

High-level visual brain regions contain subareas in which neurons appear to respond more strongly to examples of a particular semantic category, like faces or bodies, rather than objects. However, recent work has shown that while this finding holds on average, some out-of-category stimuli also activate neurons in these regions. This may be due to visual features common among the preferred class also being present in other images. Here, we propose a deep-learning-based approach for visualizing these features. For each neuron, we identify relevant visual features driving its selectivity by modelling responses to images based on latent activations of a deep neural network. Given an out-of-category image which strongly activates the neuron, our method first identifies a reference image from the preferred category yielding a similar feature activation pattern. We then backpropagate latent activations of both images to the pixel level, while enhancing the identified shared dimensions and attenuating non-shared features. The procedure highlights image regions containing shared features driving responses of the model neuron. We apply the algorithm to novel recordings from body-selective regions in macaque IT cortex in order to understand why some images of objects excite these neurons. Visualizations reveal object parts which resemble parts of a macaque body, shedding light on neural preference of these objects.

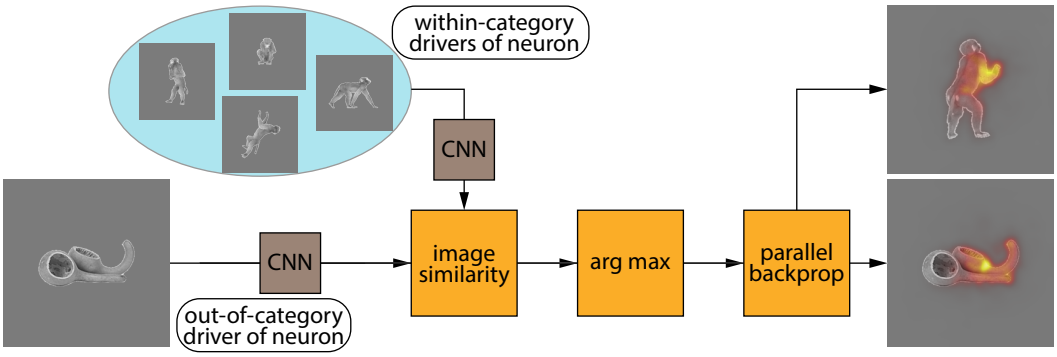

Figure 1: **Why does this object image activate IT neurons that are selective for bodies?** Our goal is to visually explain responses of category-selective neurons to outside-of-category (ooc.) stimuli. We start with an ooc. stimulus (object), that strongly activates a neuron which primarily fires for a specific category (bodies). We compute latent CNN activations for the image, as well as for a set of within-category reference images. A neuron-specific similarity metric operating on the latent activations finds the reference image most similar to the ooc. stimulus. The proposed parallel backpropagation method then highlights the shared features driving the neural response.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Background. The primate visual system has evolved to process a highly diverse set of tasks and stimuli. In higher visual areas, specialized subregions exist, in which neurons preferentially discharge in response to images stemming from a particular semantic class. These areas are often hypothesized to underlie specific computations that may only be relevant for a specific semantic concept, but it is not entirely clear why they emerge. The most prominent category-selective brain regions consist of 'face-cells' ('face patches'), which on average fire more strongly when stimulated with faces than objects [1, 2], as well as body-selective regions which have been studied to a lesser extent [3, 4, 5]. For face-selective cells, it has been shown that object images also elicit responses, albeit more sparsely than face images [6, 7]. Going further, [6] showed that models trained solely on non-face images could predict responses to face images in macaque face cells. Therefore, it has been argued that face-selective cells are not driven by the semantic concept of a face, but instead respond to visual features that are more common in face than object images. Hence, out-of-category (ooc.) images may still activate otherwise category-selective cells, as long as relevant features are apparent in the image. As of yet, characterization of these features has remained relatively rough, largely confined to the finding that face cells tend to prefer round objects and body cells tend to prefer spiky objects [8]. Other work suggests that only a small fraction of response variance in face cells can be explained by simple shape features like roundness [6]. While these global tuning properties are useful to understand average preferences of patches, we therefore argue that more fine-grained feature characterizations are necessary to understand responses at the single-image and single-neuron level.

Contribution. To address this gap, we propose a deep-neural-network-based method for visualizing features of an ooc. stimulus, which are responsible for eliciting high responses from an otherwise category-selective neuron. Our approach relies on analyzing the similarity of the image to a selected within-category image which displays similar features, as shown in Fig. 1. Specifically, after finding a within-category image with similar, neuron-specific features, we use gradient methods to highlight features extracted from a vision model which are shared between the two images and drive the neural response. This helps to visually answer the question why a feature detector that is preferably activated by within-class images, would also respond to the ooc. image in question. The procedure is compatible with any backbone visualization method that returns a saliency map for hidden units of a convolutional neural network. Further, it is entirely class-agnostic, and can therefore be used for any selective brain region, as well as for studying internal behaviour of artificial vision systems. In this work, we show results for a novel set of multi-unit recordings from body-selective regions in macaque superior temporal sulcus. Body cells constitute a particularly interesting problem, as different body poses produce vast variability among body images, suggesting a rich set of features driving these neurons. We summarize our contributions as follows:

- We present a novel method for visualizing features driving responses of category-selective neurons to out-of-category images, shedding light on why these neurons do not exclusively respond to within-category images,
- We present results from multi-unit recordings of body-selective neurons in macaque IT cortex, demonstrating that these neurons encode overlapping visual features for bodies and objects,
- We apply the proposed visualization method to the data, discussing why some non-body objects activate these neurons.

2 Related work

Category-selective visual brain areas. The largest part of the body of work studying category selectivity in the brain is targeted towards face patches [2, 1]. Several papers have come to the conclusion that face-selective neurons respond to visual features correlated with faces, rather than the semantic concept of a face. [7] generated maximally exciting images for face cells, which activated the neurons strongly but were not rated as face-like by human participants. [9] showed that face cells also respond selectively to pareidolia images, which are objects eliciting perception of a face in humans. Neurons still responded after the images were scrambled, destroying perception of a face, indicating that responses were driven by low-level features. Recently, [6] successfully predicted responses of face cells to face images, after training a linear readout model using solely non-face

images. [8] proposed a unified view of IT cortex organization, arguing that selectivity was based on a small set of principal axes of image space. Reducing the dimension allows the authors to state that face cells prefer objects with high scores on principal components corresponding to 'stubby' and 'animate', whereas body-selective cells respond to 'spiky' and 'animate' objects. Recent work [10] challenges the hypothesis of shared coding principles between faces and non-faces in face cells, showing that computational mechanisms in these brain regions are far from being well understood. Our neurophysiological recordings add additional evidence to this debate. Highly relevant to our work is that of [11], which visualized body cell responses to bodies and some objects by showing fragments of a highly activating image to determine most important image regions. This approach showed that a large proportion of cells respond to local body fragments. The main advantages of our computational approach are its high-throughput, allowing to visualize a large number of cells and objects simultaneously, and enhancing interpretability of features by analyzing them in the context of the preferred semantic class.

Attribution methods. A substantial body of work exists on attributing behaviour of deep computer vision models to specific image regions. The most common goal is to understand why classification models output the observed class label, given an input image, meaning that attribution methods are often applied to the very last network layer. Several families of approaches have been proposed for tackling this problem: The first relies on image perturbations like occlusion [12, 13], the second is based on an analysis of latent activations [14, 15], and the third computes gradients of class activations w.r.t. pixel intensities [16, 17, 18]. More recent work [19, 20] has shown that some of these methods rely too strongly on the input image itself, being independent of the network to varying degrees. Even though our reweighting scheme is compatible with any attribution method that is computable for latent activations, we therefore rely on vanilla backpropagation in this work, which has been shown to not be biased towards reconstructing the input image. We provide results for integrated gradients [18] in the appendix. Further, relevant to our work is that of [21] which computes saliency maps for image similarity. The approach is similar to that of Grad-CAM, as it relies on analyzing feature activations before and after a global pooling layer at the end of the network. For that reason, the latter method is mainly suitable for visualizing global similarity as judged by a network trained to do so, rather than more local features relevant for biological neurons along the visual hierarchy in the primate brain.

3 Methods

We propose an approach for visual explanations of neural responses to out-of-category (ooc.) stimuli in category-selective visual brain areas. The categories currently of most interest in the high-level vision literature are faces and bodies, but the method can be applied to any semantic category. We leverage differentiable neuron models to explain high neural firing rates in response to an ooc. image by analysing visual similarity to an image from the preferred category which also drives the neuron. Formally, let $x_{\text{out}} \in \mathbb{R}^{3\times h\times w}$ denote an ooc. image yielding high activity from a recorded category-selective neuron. The overall goal of this work is to determine the visual features of x_{out} driving the neural response. We approach this problem in three steps:

1. We learn a linear readout vector w on top of a pretrained CNN $f(\cdot)$ to predict neural responses to within-category stimuli $x_{\text{in},1}, \ldots, x_{\text{in},N}$. The learned vector then incorporates information about which features apparent in within-category images are relevant for the neural response.
2. We employ a neuron-specific similarity metric based on the learned readout to find a within-category image x_{in} with similar visual features as x_{out}. Visual inspection of this reference image on its own can yield insights on why x_{out} activates the neuron.
3. We backpropagate gradients of CNN activations to the pixels of both images and reweight them to highlight features that are
   (a) present in x_{out},
   (b) present in x_{in},
   (c) highly relevant for the neural response.

   Explicitly highlighting shared features helps identify specific image regions responsible for the images' neuron-specific similarity.

3.1 Modelling neural responses

In recent years, a pretrained CNN backbone combined with a trainable linear readout module has become the gold standard in modelling the stimulus-driven variance in neural responses to images [22, 23]. In these models, an image $x \in \mathbb{R}^{3\times h\times w}$ is fed through a CNN $f(\cdot)$ up to a predetermined layer to obtain a latent representation $a \in \mathbb{R}^c$. The predicted response for a neuron is then computed as

$$\hat{y} = \langle a, w \rangle, \tag{1}$$

where $w \in \mathbb{R}^c$ is a learned weight vector, and $\langle \cdot, \cdot \rangle$ denotes the standard dot product. The vector w encapsulates information about which visual dimensions the neuron is tuned to, as a large $w^{(i)}$ implies that increased activity in feature i will strongly increase the predicted neural response. Conversely, $w^{(i)} = 0$ implies no change in predicted activity if feature i is manipulated and thus such features can be ignored when trying to understand a neuron's response pattern (assuming a perfect model fit). Therefore, training a model to predict neural responses to within-category stimuli reveals which visual features drive responses within the category. If the model then generalizes to ooc. stimuli, we can infer that ooc. neural responses are driven by features present in both data distributions.

3.2 Neuron-specific image similarity

If an image x_{out} strongly drives a neuron, we propose to interpret the relevant image features by studying a visually *similar* image from the category that the neuron is selective to. To quantify similarity, we adapt the common method of computing the cosine similarity of latent activations computed using a CNN [24]. Since we have additional information on which latent dimensions drive the neuron, we weight feature i by the corresponding weight $w^{(i)}$. Formally, for images x_1, x_2 and their corresponding latent activations $a_1 := f(x_1)$, $a_2 := f(x_2)$, we define the neuron-specific image similarity as

$$s(x_1, x_2) := \frac{\langle a_1 \odot w, a_2 \odot w \rangle}{||a_1 \odot w||_2 ||a_2 \odot w||_2}, \tag{2}$$

where $\odot$ denotes the Hadamard product. Since w is often sparse, this formulation will effectively ignore most dimensions and focus solely on those that are highly relevant for predicting the neural response. For the downstream procedure of generating a visual explanation for the neural response to the image x_{out}, we simply select the most similar image from our within-category dataset, i.e.

$$x_{\text{in}} = \arg\max\{s(x_{\text{out}}, x) : x \in D_c\}, \tag{3}$$

where D_c denotes the set of within-category images.

3.3 Parallel backpropagation for shared-feature visualization

Even though the images x_{out} and x_{in} ideally have a high similarity metric, the similar features are not always clearly identifiable. In some cases, the features might not be easily interpretable, or in other cases, several features might be shared upon visual inspection, such that it remains unclear whether all of them or a subset contribute to neural activity.

In order to further highlight shared features contributing to model activity, we compute weighted gradients of the latent model activations w.r.t. to image pixels [16]. This method was originally devised for classification models to highlight those image pixels that most strongly influence a given class probability. We build on this work to visualize shared features between images by introducing a simple reweighting procedure. Specifically, we strengthen the influence of features that are shared, while attenuating features that are specific to only one of the images. For the sake of brevity, we illustrate the procedure for the gradients of x_{out} only, as the process is analogous for x_{in}. First of all, note that the gradient of the predicted neural response $\hat{y}_{\text{out}}$ is given by

$$\frac{\partial \hat{y}_{\text{out}}}{\partial x_{\text{out}}} = \frac{\partial}{\partial x_{\text{out}}} \langle a_{\text{out}}, w \rangle = \sum_i w^{(i)} \frac{\partial}{\partial x_{\text{out}}} a^{(i)}_{\text{out}}. \tag{4}$$

Due to the sum rule of calculus, the gradient of the predicted response is given by a weighted sum of the gradients of the latent features, with $w^{(i)}$ acting as weight for feature i. Hence, features that are highly relevant for model predictions will dominate the gradient. In turn, those pixels for which an

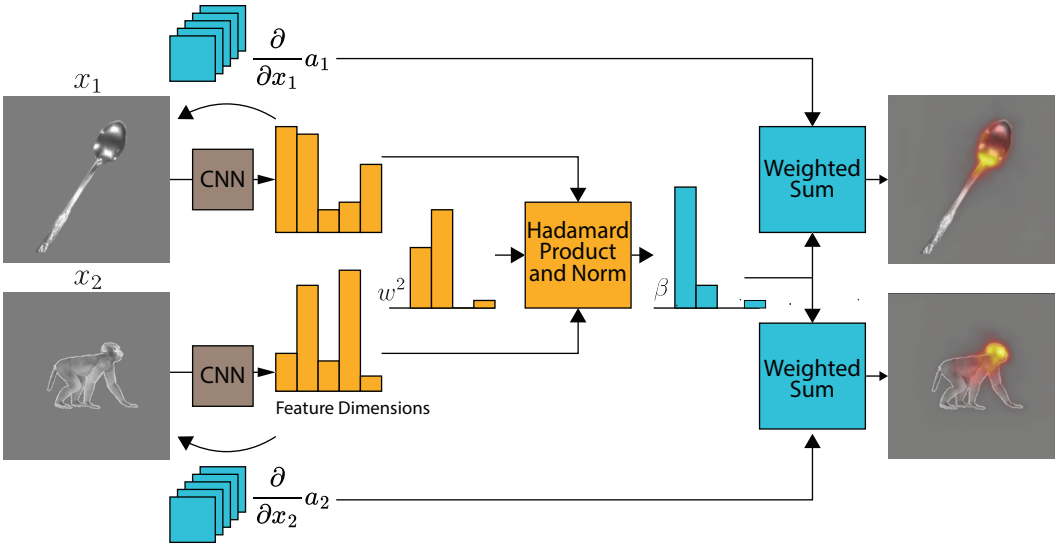

Figure 2: Sketch of the parallel backpropagation method. A pre-trained CNN cut off at a predetermined layer computes latent feature activations. These are backpropagated to obtain the Jacobians of the two activation vectors w.r.t. to the respective images. We then calculate the normalized Hadamard product of the activation vectors and the element-wise square of the learned linear readout vector for the considered neuron. The pixel saliency map is then computed as the sum of gradients of each feature, weighted by the feature's entry in the Hadamard product.

infinitesimal increase would strongly enhance one of the highly relevant features will be assigned a high intensity when plotting the gradient over the image.

For our purposes, this gradient map is not satisfactory. First of all, it does not take the other image into account at all, and thus fails to leverage the images' similarity structure. Further, since the gradient weights do not carry information on whether a feature is present in the image in the first place, it is theoretically possible to assign high intensity to pixels that would strongly increase features with low activity. To remedy these deficiencies, we propose an adjusted pixel saliency map I based on replacing the weights for the features in (4).

Before reweighting the features, we first smooth and normalize each gradient to have unit norm, i.e. $||\frac{\partial}{\partial x_{\text{out}}} a_{\text{out}}^{(i)}||_2 = 1$. Smoothing alleviates pixel-level noise and allows the user to determine *regions* of pixels with high contribution. It has been shown that a post-processing smoothing step substantially improves the quality of gradient-based attributions [20]. Normalization ensures that the pixel saliency map is not determined solely by the gradient magnitude, as some features may be more sensitive to pixel perturbation than others. More importantly, it allows us to later bound the norm of the saliency map from above, which substantially improves interpretability. Subsequently, we weight each feature by its contribution to the neuron-specific similarity metric $s(\cdot, \cdot)$ given in (2). Formally, we define the weight for feature i as

$$\beta^{(i)} := \frac{a_{\text{in}}^{(i)} w^{(i)} a_{\text{out}}^{(i)} w^{(i)}}{||a_{\text{in}} \odot w||_2 ||a_{\text{out}} \odot w||_2}, \tag{5}$$

and finally the pixel saliency maps

$$I(x_{\text{out}}; x_{\text{in}}, w) := \sum_i \beta^{(i)} \frac{\partial}{\partial x_{\text{out}}} a_{\text{out}}^{(i)}, \qquad I(x_{\text{in}}; x_{\text{out}}, w) := \sum_i \beta^{(i)} \frac{\partial}{\partial x_{\text{in}}} a_{\text{in}}^{(i)}. \tag{6}$$

The revised weight vector β imposes high weights only on those features that are highly activated in the feature vectors of both images (captured by $a_1^{(i)}$ and $a_2^{(i)}$), and are also relevant for the neuron's activity (captured by $(w^{(i)})^2$). Dividing by the product of norms allows us to bound the intensity of

the saliency maps by writing

$$||I(x_{\text{out}}; x_{\text{in}}, w)||_2 = ||\sum_i \beta^{(i)} \frac{\partial}{\partial x_{\text{out}}} a_{\text{out}}^{(i)}||_2 \leq \sum_i ||\beta^{(i)} \frac{\partial}{\partial x_{\text{out}}} a_{\text{out}}^{(i)}||_2$$
$$= \sum_i \beta^{(i)} \underbrace{||\frac{\partial}{\partial x_{\text{out}}} a_{\text{out}}^{(i)}||_2}_{=1} = \sum_i \beta^{(i)} = s(x_{\text{in}}, x_{\text{out}}). \tag{7}$$

Note that this inequality only holds if $a_{\text{in}}^{(i)} \geq 0$ and $a_{\text{out}}^{(i)} \geq 0$ for all $i = 1, \ldots, c$, as this is needed to ensure $\beta^{(i)} > 0$. However, this constraint is usually satisfied as latent activation are commonly fed through a ReLU layer before extraction. The inequality yields that the total intensity of the pixel saliency map (for either image) as given by its L_2 norm is bounded by the neuron-specific similarity. Thus if the images are dissimilar, the saliency maps will have low intensity. Finally, note here that the reweighting procedure is agnostic to the way the salience map for each feature is generated. Any (backpropagation) pixel-attribution method may be used, as long as each latent feature is assigned one saliency map with unity norm.

Running these steps successively for one recorded neuron yields as output one ooc. image and one within-category image, along with one saliency map per image. To study shared features, we display the images side-by-side with the saliency maps overlayed, as seen in Figs 4 and 5.

4 Experimental setup

Model architecture. For experiments, we use a Resnet-50 [25], which was adversarially trained on ImageNet [26, 27], as our CNN backbone. This architecture has shown good fits to neural data in several studies [28, 29, 30]. To predict the response to an image, we feed it through the Resnet up to the last ReLU of layer 4.1. Subsequently, a Gaussian readout [31, 28] reduces the dimension of the Resnet activations from $c \times h_a \times w_a$ to c by selecting a learned spatial location in the latent feature map from which to extract the final features. The receptive fields of the extracted features cover the entire foreground of the images. Finally, a fully-connected layer maps these features to neural responses.

Stimuli. For gathering neural data, we utilized two sets of images. The first consisted of 475 images of a macaque avatar on a gray background (see Fig. 1, 2). The images were subsampled from a set of 720 images, in which the avatar appeared in 45 unique poses, extracted from nine different behavioural classes, each shown from 16 viewpoints [32]. The second set comprised 6,857 objects from varying categories shown on the same gray background and was combined from the OpenImages dataset [33], as well as several smaller ones [34, 35]. To test for body selectivity, we used an additional set of 2068 control body images including a variety of species. These images were only used to test for category-selectivity of the recorded cells. All other references to body images in the text refer to the original image set showing the monkey avatar. All stimuli were centered with respect to the fixation point. They were shown to the monkey at a resolution of 280x280, and were resized to 224 for the Resnet.

Neural data collection. We recorded multi-unit activity (MUA) from and surrounding two fMRI-defined body category-selective patches [5] in the macaque superior temporal sulcus (STS), using 16-channel Plexon V probes, while the subjects performed a fixation task. Since body patch neurons typically discharge sparsely for non-body objects, we employed online stimulus selection: In a first phase, we recorded responses to the set of 475 monkey avatar images. We then trained a model for each channel to predict neural responses to these images. Subsequently, we predicted neural responses to our set of object images, and for each neuron we selected the highest and lowest predicted activator, as well as the object most similar to the top-activating body image, according to $s(\cdot, \cdot)$. We followed the same procedure for the control body images. Finally, in a second experimental phase we recorded responses to these novel images as well as a subset of 75 of the original body images, to test recording stability. We included recording channels in the analysis if the test/retest stability was higher than .60, as measured by the correlation between responses to body images before and after model fitting. Further, we tested each channel for body-category-selectivity by comparing the median response to the selected objects and the selected control bodies using a Mann-Whitney U-test. Both the animal

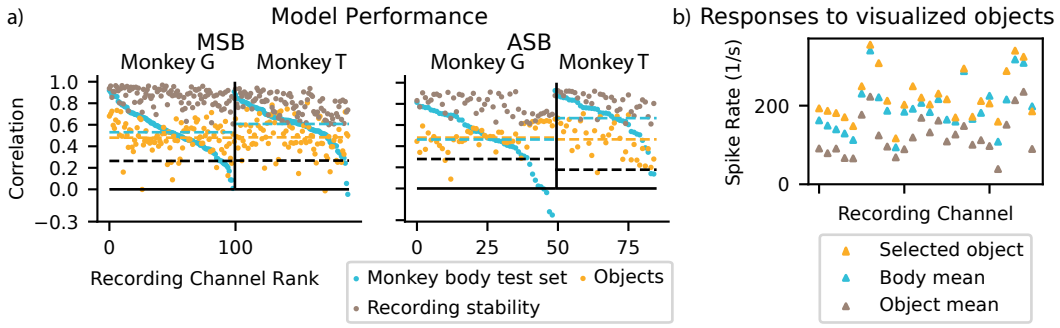

Figure 3: a) Correlation between predicted and recorded neural responses for held-out monkey body images (blue) and object images (orange). Significant positive correlation for object images demonstrates that the visual features predictive of responses to body images are also predictive of responses to objects. Brown dots show the correlation between responses to the same stimuli in the first and second recording phase. Dashed lines show the average across channels (colored) and the .05 confidence interval for the correlation coefficient under the null hypothesis that $\rho = 0$ (black). b) Neural responses to the objects for which features are visualized in Figs. 4 and 5. Responses to visualized objects are higher than mean responses to objects and bodies in the vast majority of cases.

care and experimental procedures adhere to regional (Flanders) and European guidelines and have been approved by the Animal Ethical Committee of Leuven under the protocol number P182/2019. Further experimental details are given in the appendix.

Fitting the neuron model. We split the monkey image set into a training/validation/test split consisting of 400/50/25 images. The parameters of the Gaussian readout location, as well as the linear weights, were trained simultaneously using the Adam optimizer [36], to minimize the mean squared error between recorded and predicted responses. We augmented the training data by incorporating silhouettes of the monkey, for which the tuning of body patch cells in mid-STS has been shown to be largely preserved [37]. Further, we penalized the readout-weights using $L_{1/2}$ regression to sparsify the vector while still allowing for some large weights. We set the learning rate to 10^{-4} and the weight of the regularization to 0.1 After training for 2500 epochs, we selected the model with lowest loss on the validation set. Importantly, the model was not trained to predict responses to non-body images, meaning that it must utilize the same visual features to predict bodies and objects. Models and training runs, as well as the visualization procedure were implemented in PyTorch [38]. To compute the Jacobian of latent features for visualization, we utilized the corresponding functionality of PyTorch's autograd. All experiments were run on a single Nvidia RTX 2080Ti.

5 Results

5.1 Model generalizes from bodies to objects

After training the models on body images, we first test how well they generalize to images of objects. Model performance in terms of correlation between predicted and recorded neural response is shown in Fig. 3 a). $93.7/95.3$ % of channels in the posterior/anterior region exhibit a significant positive correlation, suggesting an at least partially class-agnostic feature preference. Interestingly, strong performance on the held-out monkey images does not seem to be a necessary condition for strong performance on objects. This effect may be caused by body images being highly similar due to fine-grained sampling of poses and viewpoints, compared to high feature variance among objects.

5.2 Feature visualization

Shared features between objects and bodies. Having found that there exists a set of visual features driving neural responses to both objects and bodies, we apply the proposed visualization method to characterize these features. Results are displayed in Figs. 4 (posterior region MSB) and 5 (anterior region ASB). We show results from multi-unit sites for which model performance on the out-of-category images was high ($r > 0.4$), selecting a subset representing a wide range of

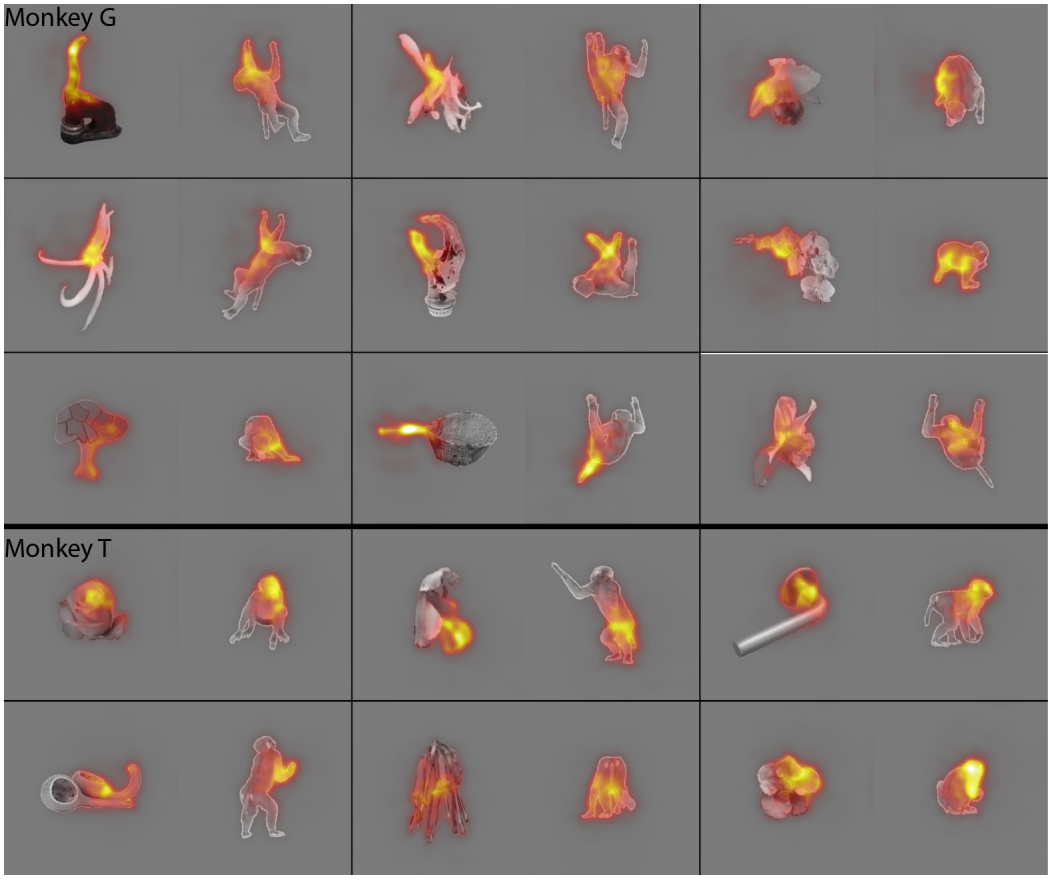

Figure 4: Results obtained by applying the proposed method to multi-unit recordings from body-selective cells in macaque STS (posterior region). Each subplot corresponds to one recording channel. The object on the left is among the most highly activating objects for the channel. The image on the right corresponds to the most similar preferred body.

highlighted features. For neighbouring recording sites, visualizations are often similar, likely due to similar feature preferences. For each channel, we display the image pair with highest neuron-specific similarity among the top-5 activating objects and top-15 activating bodies. Fig. 3 b) demonstrates that the visualized objects activate the corresponding channels more strongly than the average body/object image in most cases.

The method discovers a variety of shared features between highly activating bodies and objects. Most of them correspond to parts of the body rather than the entire body, which is aligned with previous findings for neurons from MSB.[11]. In fact, specific object parts seems to bear resemblance to specific body parts in the model's latent space. For example, extended structures appear to activate the same latent dimensions as arms/shoulders, so a model that relies on arms/shoulders to explain responses to body images also predicts strong responses to other extended structures. We observe objects driving the model due to similarity to limbs, tails, heads, torsos as well as more diffuse features which are more difficult to interpret. Interestingly, while a a lot of the observed features could be characterized as 'spiky', we also find neurons which are activated by stubby objects. The corresponding bodies show crouched poses without protruding extremities. This demonstrates that at the single-image and multi-unit level, tuning properties are more fine-grained than previously suggested [8]. In some cases, we observe the same object with different highlighted features in different recording channels, indicating that objects may have multiple features that activate different neurons. An example of this can be seen in Fig. 5, where the highlighted features of the same image correspond to a leg and a torso of the effective body images of two different channels. Additional results are given in the appendix.

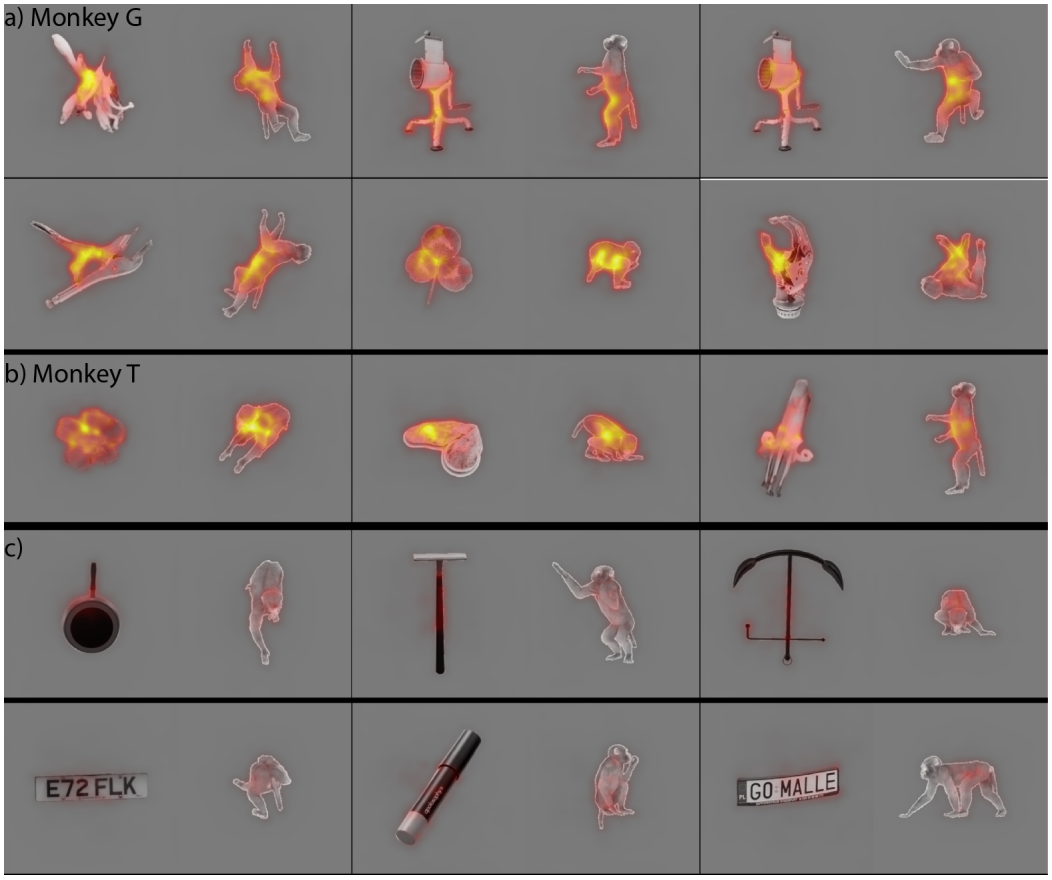

Figure 5: a)/b) Results for ASB region. c) Example results for objects lacking a highly similar body image among the channel's top drivers, and for weakly activating objects. Top row: strongly driving objects. Bottom row: weakly driving objects. Low similarity is properly reflected by low intensity of the saliency map.

Objects without shared features. We find that for some object images eliciting a high neural response, the visualization method yields no shared features with any of the top-activating body images (Fig. 5 c)). This could be either due to neurons preferring features not present in the original body image set, or due to tuning properties not captured by the model. These cases demonstrate the utility of the bound for the norm of the saliency maps given in (7), as the lack of image similarity is clearly reflected by the visualization.

6 Discussion

Limitations. The proposed approach is made possible through the use of a differentiable neuron model, which means that the visualization quality depends on the ability of the model to capture neural tuning properties. As models of visual processing further improve in the future, we predict that visualization quality will improve accordingly. Further, the visualizations will reflect idiosyncrasies of the underlying attribution method used for generating saliency maps for latent features. Since this backbone can be chosen freely, advances on the topic of attribution methods will also improve our visualizations.

Conclusion. We presented a method for visualizing selectivity of class-selective visual feature detectors when confronted with out-of-class images. Further, we showed that body-selective neurons encode bodies and objects using an at least partially shared feature set. We visualized these features, providing an explanation for why some objects activate body-selective neurons. Future work could involve using the same method for other category-selective areas, like face patches. Additionally, one

could test these visualizations by presenting highlighted fragments to the subject in a closed-loop fashion.

Acknowledgments and Disclosure of Funding

The authors thank Rajani Raman and Prerana Kumar for valuable discussions about this project. Further, the authors thank C. Fransen, I. Puttemans, A. Hermans, W. Depuydt, C. Ulens, S. Verstraeten, S. T. Riyahi, J. Helin, and M. De Paep for technical and administrative support. AL, AB, GKN, AM, LM, MG and RV are supported by ERC-SyG 856495. MG is supported by HFSP RGP0036/2016, BMBF FKZ 01GQ1704. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Alexander Lappe and Lucas Martini. Parts of the stimulus images are due to courtesy of Michael J. Tarr, Carnegie Mellon University, http://www.tarrlab.org/.

Contributions. AL, AB, MG and RV conceptualized the study. AL developed and implemented the visualization method. AB, GGN and RV collected the neural data. AL and AB analyzed the data. AM, AB, LM and AL contributed to stimulus generation. AL wrote the initial draft of the manuscript, and all authors contributed to the final version. MG and RV supervised the project.

References

[1] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, 17(11):4302–4311, June 1997. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.17-11-04302.1997.

[2] Doris Y. Tsao, Winrich A. Freiwald, Roger B. H. Tootell, and Margaret S. Livingstone. A Cortical Region Consisting Entirely of Face-Selective Cells. *Science*, 311(5761):670–674, February 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1119983.

[3] Paul E. Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A Cortical Area Selective for Visual Processing of the Human Body. *Science*, 293(5539):2470–2473, September 2001. doi: 10.1126/science.1063414.

[4] Serguei V. Astafiev, Christine M. Stanley, Gordon L. Shulman, and Maurizio Corbetta. Extrastriate body area in human occipital cortex responds to the performance of motor actions. *Nature Neuroscience*, 7(5): 542–548, May 2004. ISSN 1097-6256. doi: 10.1038/nn1241.

[5] Rufin Vogels. More Than the Face: Representations of Bodies in the Inferior Temporal Cortex. *Annual Review of Vision Science*, 8(1):383–405, 2022. doi: 10.1146/annurev-vision-100720-113429.

[6] Kasper Vinken, Jacob S. Prince, Talia Konkle, and Margaret S. Livingstone. The neural code for "face cells" is not face-specific. *Science Advances*, 9(35):eadg1736, September 2023. ISSN 2375-2548. doi: 10.1126/sciadv.adg1736.

[7] Alexandra Bardon, Will Xiao, Carlos R. Ponce, Margaret S. Livingstone, and Gabriel Kreiman. Face neurons encode nonsemantic features. *Proceedings of the National Academy of Sciences*, 119(16): e2118705119, April 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2118705119.

[8] Pinglei Bao, Liang She, Mason McGill, and Doris Y. Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, July 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2350-5.

[9] Saloni Sharma, Kasper Vinken, and Margaret S. Livingstone. When the whole is only the parts: Non-holistic object parts predominate face-cell responses to illusory faces. *bioRxiv*, 2023.

[10] Yuelin Shi, Dasheng Bi, Janis K. Hesse, Frank F. Lanfranchi, Shi Chen, and Doris Y. Tsao. Rapid, concerted switching of the neural code in inferotemporal cortex. *bioRxiv*, pages 2023–12, 2023.

[11] Ivo D. Popivanov, Philippe G. Schyns, and Rufin Vogels. Stimulus features coded by single neurons of a macaque body category selective patch. *Proceedings of the National Academy of Sciences of the United States of America*, 113(17):E2450–2459, April 2016. ISSN 1091-6490. doi: 10.1073/pnas.1520371113.

[12] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks, November 2013.

[13] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models, September 2018.

[14] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization, December 2015.

[15] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7.

[16] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014.

[17] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net, April 2015.

[18] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017.

[19] Julius Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Neural Information Processing Systems*, October 2018.

[20] Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards Better Understanding Attribution Methods. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10213–10222, June 2022. doi: 10.1109/CVPR52688.2022.00998.

[21] Abby Stylianou, Richard Souvenir, and Robert Pless. Visualizing Deep Similarity Networks, January 2019.

[22] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

[23] Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, 15(4):1–27, April 2019. doi: 10.1371/journal.pcbi.1006897.

[24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. https://arxiv.org/abs/2002.05709v3, February 2020.

[25] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

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

[27] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust ImageNet models transfer better?, 2020.

[28] Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization. *bioRxiv : the preprint server for biology*, 2023. doi: 10.1101/2023.05.12.540591.

[29] Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago A. Cadena, Kelli Restivo, George Denfield, Andreas S. Tolias, and Fabian H. Sinz. Towards robust vision by multi-task learning on monkey visual cortex, 2021.

[30] Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H. McDermott. Model metamers illuminate divergences between biological and artificial neural networks. *bioRxiv : the preprint server for biology*, 2023. doi: 10.1101/2022.05.19.492678.

[31] 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, Alexander S. Ecker, and Fabian H. Sinz. Generalization in data-driven models of primary visual cortex. *bioRxiv : the preprint server for biology*, 2020. doi: 10.1101/2020.10.05.326256.

[32] Anna Bognar, Albert Mukovskiy, Ghazal Ghamkhari Nejad, Nick Taubert, Michael Stettler, Rajani Raman, Martin Giese, and Rufin Vogels. Simultaneous recordings from posterior and anterior body-responsive regions in the macaque Superior Temporal Sulcus. *Journal of Vision*, 23(9):5403, August 2023. ISSN 1534-7362. doi: 10.1167/jov.23.9.5403.

[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, July 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-020-01316-z.

[34] Timothy F. Brady, Talia Konkle, George A. Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14325–14329, September 2008. ISSN 1091-6490. doi: 10.1073/pnas.0803390105.

[35] Sophie Lebrecht, Moshe Bar, Lisa Feldman Barrett, and Michael J. Tarr. Micro-Valences: Perceiving Affective Valence in Everyday Objects. *Frontiers in Psychology*, 3:107, April 2012. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00107.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017.

[37] Ivo D. Popivanov, Jan Jastorff, Wim Vanduffel, and Rufin Vogels. Tolerance of Macaque Middle STS Body Patch Neurons to Shape-preserving Stimulus Transformations. *Journal of Cognitive Neuroscience*, 27(5):1001–1016, May 2015. ISSN 0898-929X. doi: 10.1162/jocn_a_00762.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, December 2019.

[39] Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):2134, 2018.

[40] A. Bognár, R. Raman, N. Taubert, Y. Zafirova, B. Li, M. Giese, B. De Gelder, and R. Vogels. The contribution of dynamics to macaque body and face patch responses. *NeuroImage*, 269:119907, 2023. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2023.119907. URL `https://www.sciencedirect.com/science/article/pii/S1053811923000551`.

A Appendix / supplemental material

A.1 Additional Results

A.1.1 Synthetic data

To test our method on a wider set of semantic categories, we generate six synthetic, category-selective neurons. To do so, we gather a small set of within-category images $x_{\text{in},1}, \dots, x_{\text{in},N}$ and a set of out-of-category images $x_{\text{out},1}, \dots, x_{\text{out},M}$ from the stimulus sets used for the electrophysiological experiments. Feeding these through the CNN and sampling a spatial location using a Gaussian readout with random location preference yields activation matrices $A_{\text{in}} \in \mathbb{R}^{N \times c}$ and $A_{\text{out}} \in \mathbb{R}^{M \times c}$. A synthetic neuron with readout weights $w \in \mathbb{R}^c$, that on average prefers the within-category images can then easily be found by solving

$$\arg\max_w (A_{\text{in}} w)^\intercal A_{\text{in}} w - (A_{\text{out}} w)^\intercal A_{\text{out}} w = \arg\max_w w^\intercal (A_{\text{in}}^\intercal A_{\text{in}} - A_{\text{out}}^\intercal A_{\text{out}}) w.$$

This formulation is useful since the Rayleigh quotient is solved by setting w to be the first eigenvalue of $A_{\text{in}}^\intercal A_{\text{in}} - A_{\text{out}}^\intercal A_{\text{out}}$ [39]. Of course, the Resnet also contains category-selective neurons that do not need to be constructed artificially. However, we aim to make the experiment as similar to neural recordings as possible, where model neurons are usually given as a linear readout of latent activations.

Figure 6: Results of applying the parallel backpropagation method to six synthetic, category-selective neurons. Each row corresponds to one neuron, with the preferred category indicated above.

We observe that the visualizations clearly mark features that are common among within-class images, showing why the ooc. stimuli drive the otherwise category-selective neurons.

A.1.2 Integrated Gradients

Our visualization procedure is based on a weighted sum of a tensor containing saliency maps for all latent features. As discussed in the main text, the method is agnostic towards how the individual saliency maps are generated. For the main experiments, we use the vanilla gradient method which results in reweighting the Jacobian of the latent features. Here, we experiment with using a different saliency backbone, namely Integrated Gradients [18]. For an image x, Integrated Gradients approximates the path integral of the gradients along the straightline path from an image of zeros to x. Originally developed for visualizing scalar outputs, we adapt the formula from [18] to the multi-dimensional case to yield

$$\text{IntegratedGrads}(x) = x \odot \frac{1}{m} \sum_{k=1}^{m} Jf\left(\frac{k}{m} x\right),$$

where $Jf(x)$ denotes the Jacobian of the CNN features w.r.t. x. Since [19] found that this formulation is heavily influenced by the element-wise multiplication with the input x, we omit this step in the computation. Results shown in 7 demonstrate that the visualizations are almost indistinguishable from those computed using vanilla backpropagation.

A.2 Additional details for neural data collection

Experiment details. We recorded neuronal responses during passive fixation task using a 2x2 degree fixation window (EyeLink 1000 infrared eye tracker sampling at 1000 Hz). Stimuli were shown gamma-corrected on a

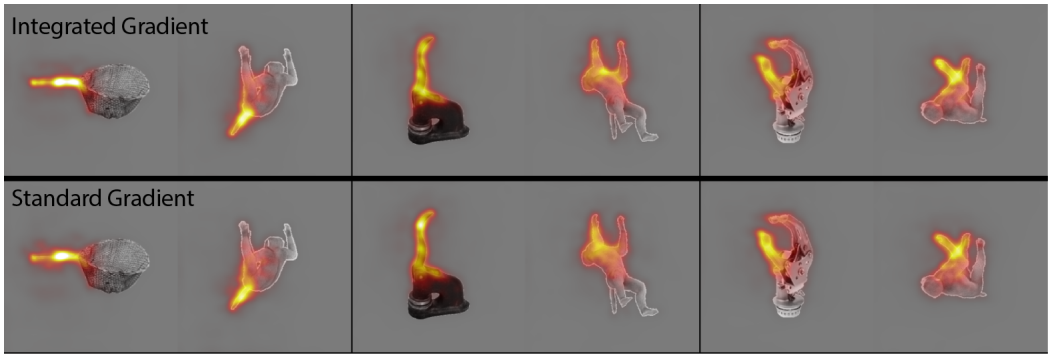

Figure 7: Results for computing the Jacobian using the Integrated Gradient method (top row). Bottom row shows the results for the standard gradient computation from the main text for reference.

22.5-inch ViewPixx monitor at a distance of about 57 cm, with a resolution of 1920x1080 and a refresh rate of 120 Hz. A set of 475 stimuli of a monkey avatar in various poses were presented 8 times in a pseudorandom sequence for 200 ms on a gray background, with a 250 ms interstimulus interval. During the interstimulus interval, only a fixation dot was present and monkeys could receive a brief juice reward . Stimulus onset and offset were indicated by a photodiode, detecting luminance changes synced with the stimuli, in the display corner (invisible for the monkeys). Control of stimulus presentation, event timing, and juice delivery was managed by an in-house Digital Signal Processing-based computer system, which also monitored the photodiode signal and tracked eye positions. Neuronal data was collected from and surrounding two fMRI-defined body-selective patches in the ventral superior temporal sulcus (STS) using 16-site linear electrodes (Plexon V probe) with Open Ephys acquisition board and software (sampling rate: 30000 Hz, filtered between 500-5000 Hz). Multi-unit activities were extracted using Plexon Offline Sorter, after applying a high-pass Butterworth filter with a cutoff frequency of 250 Hz. From the stimulus event synchronized continuous data, 550 ms trials were extracted, comprising 200 ms prestimulus and 150 ms poststimulus periods. Responsive (at least for one stimulus showing stronger than 5 spikes/s net responses (baseline: -75 to +25 ms, response: +50 to +250 ms)), and body-selective MUAs ($p<0.01$ Kruskal-Wallis test), with a split-half reliability >0.5 (Spearman Brown corrected) were selected. The responses of these MUA sites were used to predict neural responses to our set of object and body images, and for each neuron we selected the highest and lowest predicted activator, as well as the object most similar to the top-activating avatar stimuli, according to $s(\cdot, \cdot)$. Finally, in a second experimental phase we recorded responses to these object and body images as well as a subset of the original monkey avatar stimuli, to test recording stability (same experimental design as in the first phase). For all neurons considered in this work, we tested for body-selectivity by comparing the median response to bodies to the median response to objects using a Mann-Whitney U-test, considering only channels for which the test detected a significant difference.

Animals and husbandry. Two male 7 years old rhesus monkeys (Macaca mulatta), weighing 9.2 and 11 kg, respectively, contributed to this study. The animals were housed in enclosures at the KU Leuven Medical School and experienced a natural day-night cycle. Each monkey shared its enclosure with at least one other cage companion. On weekdays, dry food was provided ad libitum, and the monkeys obtained water, or other fluids, during experiments until they were satiated. During weekends, the animals received water along with a mixture of fruits and vegetables. The animals had continuous access to toys and other forms of enrichment. After fMRI scanning, we implanted a custom-made plastic recording chamber, allowing a dorsal approach to temporal body patches. In each animal, the location of the recording chamber was guided by the fMRI body localizer described in [40]. Surgery was performed using standard aseptic procedures and under full anesthesia.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim to present a method for shared-feature visualization and then apply it to novel recordings from macaque IT. The former is presented in the methods section, the latter in the results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations of the paper in a dedicated subsection at the end of the manuscript.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The derivation of the visualization method is is based on simple, self-contained mathematical arguments. Necessary conditions for the upper bound of the saliency map are clearly stated.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all major experimental details in 4. This goes for both neurophysiological as well as in-silico experiments. Fine details are given in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and data necessary to reproduce all results given in the paper are included in the submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The main paper contains all hyperparameters necessary to qualitatively reproduce the results. Finer details can be found in the attached code base.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: This question is applicable to model performance in terms of correlations on held-out data. We provide the appropriate confidence intervals in Fig. 3. Computation of these intervals, as well as significance for single-channel correlations is part of the attached code base. The code base also includes the tests for body-selectivity of recorded neurons.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We state that all experiments were run on a single Nvidia RTX 2080Ti.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: After review of the code of ethics, we believe that our work does not violate any of the statements. Additionally, we provide information about animal experiments in the appendix.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: At this point in time, we cannot identify any harmful impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We believe that misuse of our work is unlikely at this point in time, and therefore do not provide any safeguards.

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: To the best of out knowledge, we have cited all creators of assets that were used in this work. For the creator of some of the stimuli, we have incorporated a statement in the acknowledgements per his request.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We answered NA because we do not publish a new model architecture or data set. However, we will make the code to run the visualization method public upon publication of the manuscript.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not use human participants in this work. We provide statements regarding the use of non-human primates in the appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We provide statements about the role of non-human primates in the appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.