Deep Auto-encoder with Neural Response

Xuming Ran\textsuperscript{1,}, Jie Zhang\textsuperscript{2,3,}, Ziyuan Ye\textsuperscript{1,}, Haiyan Wu\textsuperscript{4,}, Qi Xu\textsuperscript{5,}, Huihui Zhou\textsuperscript{6,∗} and Quanying Liu\textsuperscript{1,∗}

\textsuperscript{1}Southern University of Science and Technology, China
\textsuperscript{2}Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
\textsuperscript{3}University of the Chinese Academy of Sciences, China
\textsuperscript{4}Centre for Cognitive and Brain Sciences and Department of Psychology, University of Macau
\textsuperscript{5}School of Artificial Intelligence, Dalian University of Technology, China
\textsuperscript{6}Pengcheng Laboratory, China

\section*{Abstract}
Artificial intelligence and neuroscience are deeply interactive. Artificial neural networks (ANNs) have been a versatile tool to study the neural representation in the ventral visual stream, and the knowledge in neuroscience in return inspires ANN models to improve performance in the task. However, how to merge these two directions into a unified model has less studied. Here, we propose a hybrid model, called deep auto-encoder with the neural response (DAE-NR), which incorporates the information from the visual cortex into ANNs to achieve better image reconstruction and higher neural representation similarity between biological and artificial neurons. Specifically, the same visual stimuli (i.e., natural images) are input to both the mice brain and DAE-NR. The DAE-NR jointly learns to map a specific layer of the encoder network to the biological neural responses in the ventral visual stream by a mapping function and to reconstruct the visual input by the decoder. Our experiments demonstrate that if and only if with the joint learning, DAE-NRs can (i) improve the performance of image reconstruction and (ii) increase the representational similarity between biological neurons and artificial neurons. The DAE-NR offers a new perspective on the integration of computer vision and visual neuroscience.

\section{1 Introduction}
Computer visions are interested in how to add information from biological neurons into artificial neural networks (ANN). Such biology-inspired AE models may help improve performance in image reconstruction tasks, as well as bring biological interpretability [8, 31, 30]. The question is how to integrate such information into AEs.

On the other hand, computational neuroscience is interested in building models that map stimuli to neural responses. However, it is difficult for traditional models to express the nonlinear characteristics between stimulus and neural response. As a powerful tool, deep learning can be used to build computational neuroscience models and uncover the relationship between stimuli and neural spikes [18]. Although biological and artificial neural networks may have fundamentally differences in computing and learning [22], both are realized by interconnected neurons: the former by biological neurons; the latter by artificial neurons. Some previous work focused on exploring the similarities between the information representation of biological neurons and artificial neurons using end-to-end ANNs, suggesting that artificial neurons in different layers of ANNs share similar representations with biological neurons in the corresponding brain regions along the ventral visual pathway [6, 41, 36, 2]. The question is how to build ANNs with the highest resemblance to biological neural representation.

In this paper, we aim to tackle these two questions at one piece. Specifically, we propose a simple biologically interpretable deep autoencoder model, which jointly learns to map a specific layer of the encoder network to the biological neural responses by a mapping function and to reconstruct the visual input by the decoder. As a result, the representations in its encoder have a higher similarity with the real neural responses, and the decoder can achieve better reconstruction performance. Our contributions can be summarized as:

- We propose a novel model called Deep Autoencoder with Neural Response (DAE-NR). The model can simultaneously learn to predict neural responses and to reconstruct the visual stimuli. (Sec.3).
- DAE-NR can improve the image reconstruction quality with the help of a Poisson loss on the predicted neural activity, compared to the traditional DAE models (Sec.4.2).

\textsuperscript{∗}Corresponding author:liuqy@sustech.edu.cn to Q.L. and zhouhh@pcl.ac.cn to H.Z.
Figure 1: The illustration of the model of (a) the standard deep auto-encoder (DAE) for images reconstruction; (b) the convolutional neural network with factorized readout (CNN-FR) for prediction of neuron responses; (c) the DAE with the neuron response (DAE-NR) for images reconstruction and predictions of neuron responses. $s$ is the biological neural response, the prediction of biological neural response is represented as $\hat{s}$, and $h_i (i \in \{1, 2, 3, 4\})$ is the feature of the $i$th convolutional layer.

- DAE-NR provides the higher resemblance between artificial neurons and biological neurons, compared to the end-to-end computational neuroscience model without the image reconstruction task (Sec.4.3).

2 Related work

Image reconstruction via autoencoders: A big breakthrough of image reconstruction was introduced in [13] which equips an autoencoder with a stack of restricted Boltzmann machines such that the AE can go deeper and go beyond the limited layers. The denoising autoencoder [33] further improves the robustness of the autoencoder by adding noise into the raw inputs and reconstructing the raw inputs. Convolutional autoencoders [23] ensemble the convolutional neural network (CNN) to model the encoder and decoder [21, 11], adopting the merits of convolutional layers in image feature extraction. More variants of autoencoders for image reconstruction can be found in a recent survey paper [11]. However, the variants of autoencoder for image reconstruction all suffer the same problem: that is, the parameters have high learning degree of freedom. In other words, the parameters learn only under the guidance of gradient by error backpropagation. Thus, some meaningful constraints on parameter space would favor the learning.

Neural similarity in computational neuroscience: There are many models in computational neuroscience have been proposed to study the relationship between stimuli and the corresponding neural spike responses, which is called neural spike encoding. The goal of these models is to increase the similarity between predicted and true neural responses. Historically, many efforts have been made to find the tuning curve of the specific feature of visual stimuli, such as the orientation of bar, which could predict the neural responses [14, 15, 4, 7, 26]. This is feasible for some neurons in the primary visual cortex, but not for all neurons. Nonlinear methods provide a more general way to predict neural responses, for instance, energy models [14], the linear-nonlinear model (LN) and its extended version LNLN [25]. Traditional machine learning methods, such as generalized linear models (GLMs) [40], the multi-layer percep-
Figure 2: The reconstructed images with neurons in Region 3. From top to bottom, each row displays the original images (a), the images reconstructed by DAE (b), DAE-NR$_1$ (c), DAE-NR$_2$ (d), DAE-NR$_3$ (e), DAE-NR$_4$ (f), respectively.

Figure 3: The number of significant neurons and insignificant neurons of region 3 in the image reconstruction experiments. The threshold for significance is $p \leq 0.05$.

3 Method

In this section, we introduce the notations of variables and spaces used in the model. The visual stimulus (i.e., natural images) and the corresponding neural responses (i.e., neural spikes in V1 region) are denoted as $x \in \mathbb{R}^{N \times P \times P \times C}$ and $s \in \mathbb{R}^{N \times M}$, respectively. The features of stimuli in the $i$th convolutional layer are denoted as $h_i \in \mathbb{R}^{N \times K \times K \times F}$, where $N$, $P$, $C$, $K$, $F$ and $M$ indicate the sample size ($N$), the image resolution ($P$), the image channel ($C$), the kernel size of feature of stimuli ($K$), the number of kernels ($F$), and the number of V1 neuron ($M$), respectively. The space of visual stimuli, the neural response, and the feature of stimuli in the $i$th convolutional layer are denoted by $\mathcal{X}$, $\mathcal{S}$, and $\mathcal{H}_i$.

We then introduce the background of deep auto-encoders (DAE) in Sec. 3.1, convolutional neural networks with a factorized readout layer (CNN-FR) in sec. 3.2, and our proposed method (i.e., DAE-NR) in Sec. 3.3.

3.1 DAE

The standard DAE consists of an encoder, the latent space, and a decoder, which has been demonstrated as a considerable model for image reconstruction. In our work, rather than embedding to the latent space, we shift the encoder and decoder to $i$th layer of DAE. In this way, the architecture of our...
Table 1: The quantitative results of image reconstruction with all neurons in the region 1, 2, and 3, respectively.

| Region 1 | Region 2 | Region 3 |
|----------|----------|----------|
| Model    | MSE↓     | PSNR↑   | SSIM↑   | MSE↓     | PSNR↑   | SSIM↑   | MSE↓     | PSNR↑   | SSIM↑   |
| DAE      | 0.022  | 23.709 | 0.771   | 0.024  | 23.338 | 0.754   | 0.081  | 17.039 | 0.561   |
| DAE-NR₁  | 0.021  | 23.829 | 0.776   | 0.023  | 23.932 | 0.753   | 0.044  | 19.751 | 0.763   |
| DAE-NR₂  | 0.021  | 23.779 | 0.775   | 0.023  | 23.440 | 0.759   | 0.043  | 19.819 | 0.764   |
| DAE-NR₃  | 0.021  | 23.778 | 0.775   | 0.024  | 23.330 | 0.755   | 0.043  | 19.789 | 0.761   |
| DAE-NR₄  | 0.022  | 23.721 | 0.773   | 0.023  | 23.491 | 0.760   | 0.059  | 18.462 | 0.668   |

Figure 4: The number of significant neurons (orange color) and insignificant neurons (green color) in region 3 in the image reconstruction experiments. The threshold for significance is \( p \leq 0.05 \). The first row are the baseline models (CNN-FRs) and the second row are our models (DAE-NRs).

DAE turns to be: (i) an encoder \( f₁ : \mathcal{X} \rightarrow \mathcal{H}_i \), to embed the input to neural representation in \( i \)th layer; (ii) a decoder \( f₂ : \mathcal{H}_i \rightarrow \hat{\mathcal{X}} \), to reconstruct the input based on neural representations in \( i \)th layer. We formally describe them as follows:

The encoder:
\[
\mathbf{h}_i = f₁(\phi_1, \mathbf{x}),
\]

The decoder:
\[
\hat{\mathbf{x}} = f₂(\phi_2, \mathbf{h}_i),
\]

where the \( \hat{\mathbf{x}} \) is reconstructed from the original image \( \mathbf{x} \). Both the encoder and the decoder are realized by convolutional neural networks with parameters \( \phi_1 \) and \( \phi_2 \), respectively. The goal of DAE is to reconstruct the image. Thus the the loss function of DAE is formulated as a \( L² \) norm:

\[
\mathcal{L}(\phi_1, \phi_2) = \|\mathbf{x} - \hat{\mathbf{x}}\|²/₂.
\]

3.2 CNN-FR

The CNN-FR consists of two parts, the convolutional layers as the encoder and the factorized readout layer [18, 2, 3, 45]. The convolutional layers convolve the image with a number of kernels followed by batch normalization, resulting in multiple feature maps. The readout layer pools the output of the convolutional layer (i.e., \( \mathbf{h}_i \)) by applying a sparse mask on each neuron. Let us denote that \( \mathbf{h}_i \) lies in the feature space \( \mathcal{H}_i \) and the neural responses in the space \( \mathcal{S} \). The mapping function \( f₃ : \mathcal{H}_i \rightarrow \hat{\mathcal{S}} \) is

\[
\hat{\mathbf{s}} = f₃(\mathbf{h}_i, \theta_s, \theta_d) = \left[ \sum (\theta_s \cdot \mathbf{h}_i) \right] \ast \theta_d + b,
\]

where \( \theta_s \) is the spatial mask, \( \theta_d \) is the weights sum of all features \( \mathbf{h}_i \), and \( b \) is the bias.

We use the Poisson loss to optimize the representation similarity between artificial neurons (i.e., \( \hat{\mathbf{s}} \)) and biological neurons neurons (i.e., \( \mathbf{s} \)) in the mapping function as Eq.(4),

\[
\mathcal{L}(\theta_s, \theta_d) = \sum (f₃(\mathbf{h}_i, \theta_s, \theta_d) - s \log f₃(\mathbf{h}_i, \theta_s, \theta_d)) = \sum (f₃(\mathbf{h}_i, \theta_s, \theta_d) - s \log f₃(\mathbf{h}_i, \theta_s, \theta_d)) \cdot \left( f₃(\mathbf{h}_i, \theta_s, \theta_d) \right).
\]

Previous studies have shown that the activity of V1 neurons are sparse in response to natural stimuli, and the activity of neural population with higher sparseness exhibits greater discriminability between natural stimuli [34, 39, 9, 42]. Likewise, [44] have reported that the resemblance between the representation of biological neurons and artificial neurons in higher convolutional layers exist only under the sparsity constraint on the CNN, regardless of any other factors (e.g., model structure, training algorithm, receptive field size, and property of training stimuli). In our study, the neural similarity of V1 neurons is brought to a specific layer of the encoder.
of DAEs ($f_i$, $i \in \{1, 4\}$) with the sparsity constraints on artificial neurons of that layer.

### 3.3 DAE-NR

The DAE-NR combines the function of DAE and CNN-FR. It consists of three parts, including an encoder $f_1: \mathcal{X} \to \mathcal{H}_1$, a decoder $f_2: \mathcal{H}_4 \to \mathcal{X}$, and a mapping function $f_3: \mathcal{H}_2 \to \mathcal{S}$. The loss function of DAE-NR explicitly considers both the image reconstruction task and the neural representation similarity task, as defined in Eq. (5).

$$L(\phi_1, \phi_2, \theta_s, \theta_d) = \alpha * L(\phi_1, \phi_2) + \beta * L(\theta_s, \theta_d),$$

where $\alpha$ and $\beta$ are the hyperparameters to trade-off the image reconstruction task and the neural representation similarity task. Intuitively, the larger $\alpha$ favors the reconstruction task, while the larger $\beta$ biases toward the neural representation similarity task.

## 4 Experiments and Results

### 4.1 Experimental settings

**Dataset:** We conduct experiments on a publicly available dataset with the gray-color images as visual stimuli and the corresponding neural responses. The dataset of neural data from [1]. The neural responses in the dataset are recorded in three regions in primary visual cortex (V1) of sedated mice visually stimulated with the natural images (See Appendix Fig. 5). The number of neurons over the three brain regions are shown in Appendix Table 4.

**Network architectures:** The size of the DAE part in DAE-NR is 48C23-48C13-48C23-48C13-48DC23-48DC13-1DC23 (kCnm and kDCnm indicates $k$ filters in the convolutional and deconvolutional layer, $n$ stride and $m$ filters,), and the latent variable size is 100. Each convolutional layer is followed by a batch normalization with an ELU activation function, while the activation function in the final layer for image reconstruction is tanh. The CNN-FR part in DAE-NR shares the four different convolutional layers ($h_1$, $h_2$, $h_3$, and $h_4$) of the encoder of DAE with the convolutional layer parts, and keeps the factorized readout layer part. So we have the baseline DAE model for image reconstruction and CNN-FR with $h_i$, ($CNN-FR_i, i \in \{1, 2, 3, 4\}$) for neural representations similarity. The CNN-FR part in DAE-NR can be readout from four different convolutional layers (i.e., $h_i$ of the encoder of DAE), so there are four variants of DAE-NR, i.e., DAE-NR$_i$, $i \in \{1, 2, 3, 4\}$ representing the CNN-FR extracted from the different convolutional layers $h_i$.

**Training procedures:** We preprocess the images (e.g., reshape the size of natural image to 32 $\times$ 32 $\times$ 1 and normalize the intensity of image to [-1,1]), and then input them to the model. The model is trained with an initial learning rate of 0.001 and the early stopping strategy is applied based on a separated validation set. If the error in validation set does not improve by 1000 steps, we return the best parameter set, reduce the learning rate by two, and train the second time. We stop the training procedure if the error does not improve in the second time.

**Tasks:** There are two tasks in the experiment: 1) the image reconstruction (IR) task, and 2) the neural representation similarity (NRS) task. The settings for hyperparameters are as listed in Appendix Table 5&6 for two tasks, respectively. In the IR task, we use the mean squared error (MSE$\downarrow$), structural similarity (SSIM$\uparrow$) [38], and peak signal-to-noise ratio (PSNR$\uparrow$) [38] as metrics to quantify the image reconstruction performance$^1$. We compare the DAE-NR$_1$, DAE-NR$_2$, DAE-NR$_3$ and DAE-NR$_4$ with the baseline model (i.e., the standard DAE). In the NRS task, we use Pearson correlation (CC$\uparrow$) as a metric to quantitatively evaluate models. We implement the traditional end-to-end computational neuroscience model using CNN-FR$_i$ ($i \in \{1, 2, 3, 4\}$) as baseline models.

### 4.2 Effects of neural responses in DAE-NR on image reconstruction

Ten examples of reconstructed images by DAE-NR with neural responses in brain region 3 are presented in Fig. 2. It is obvious that DAE-NR models, no matter which layer the neural response are mapped to, can better reconstruct the original images, compared with the standard DAE. The results of the brain region 1 and 2 are illustrated in Appendix Fig. 6

$^1$The up arrow $\uparrow$ indicates that the higher the value, the better, and the down arrow $\downarrow$ indicates that the lower the better.
and Fig. 7. The quantitative comparisons of image reconstruction performance are listed in the Table 1. It shows that DAE-NR1, DAE-NR2, and DAE-NR3 achieve the best performance when the network gets information from the neurons in the region 1, 2, and 3, respectively.

Fig. 2 and Table 1 together suggest that the information from biological neurons could help DAE-NR for image reconstruction. Further questions are under what circumstances and to what extent DAE-NR is beneficial from the neural response. Our hypothesis is that only the biological neurons with the high representation similarity to the artificial neurons are informative and helpful for DAE-NR.

To test our hypothesis, we identify the biological neurons that significantly correlate with the artificial neurons in DAE-NR model, as well as the insignificant neurons in each brain region. Fig. 3 shows the number of significant neurons and insignificant neurons in region 3 in the IR task. The number for region 1&2 is in Appendix Fig. 8. Then, we train the DAE-NRs with the information from these two groups of neurons in the regions 3 for IR task, separately. The quantitative results of three metrics are shown in Table 2, confirming that the significant neurons largely improve the IR performance, while the insignificant neurons jeopardize DAE-NR, resulting even worse IR results than the standard DAE (see results of Regions 3 in Table 1). These experiments verify our hypothesis, suggesting that the information from significant neurons is capable of guiding the DAE-NR to better reconstruct images.

4.3 Effects of image reconstruction in DAE-NR on neural representation similarity

Here we test the effects of image reconstruction in DAE-NR on neural representation similarity. Pearson correlation (CC↑) between the neural representation of artificial neurons and biological neurons in region 1, region 2 and region 3 are shown in the Table 3. Our models (i.e., DAE-NR), obtain larger CC↑ in all three regions, compared to the baseline models without the image reconstruction loss (i.e., CNN-FR). The results imply that the image reconstruction in our model is beneficial to the neural representation similarity. This phenomenon seems counter-intuitive as there is a tradeoff between the IR loss and the NSR loss (Eq. (5)). We hypothesize that the reason is rooted on the feature learning for image reconstruction in DAE. With the image reconstruction loss, more biological neurons would share similarity with artificial neurons. Indeed, our experimental results confirm the hypothesis (Fig. 4), suggesting that the image reconstruction loss can help DAE-NR to improve neural representation similarity between artificial and biological neurons.

5 Discussion and Conclusion

In this study, we proposed the DAE-NRs, a hybrid model that integrates the neural response into deep autoencoder models. Inspired by the CNN with a factorized readout layer in computational neuroscience [18, 2], we used a Possion loss (i.e., θs, θt) to bring neural information to a specific layer of DAE, resulting better image reconstruction performance (Fig. 2 & Table 1). In return, the IR task contributes to the feature learning in DAE and leads to a higher neural representation similarity between biological neurons and artificial neurons (Table 3 & Fig. 4). Our work provides a bridge between the DAEs for the image reconstruction and the computational neuroscience models for neural representation similarity.

Besides the IR task and NRS task in this work, DAE-NR enables many other potential applications in future. For instance, DAE-NR provides a more natural way to synthesize images that maximize neural activity or control neural population, compared with the method proposed by [2]. DAE-NR can serve as a data engine for biological experiments and synthesizes neural responses under visual stimuli. Moreover, it is important to investigate the generalizability of DAE-NR on variants of autoencoder [17, 32], on other stimuli (e.g., sound, face), as well as on computational neuroscience models [24, 16] in additional tasks, e.g., classification, generation and detection [27]. Although we only tested DAE-NR with mice neural data, the pipeline of DAE-NR can be easily extended to the neurons of primates (e.g. monkey [43] or human).

Beyond DAE-NRs, there are many future directions that combine artificial intelligence and brain intelligence that are worthy of investigation.

Acknowledgment

This work was funded in part by the National Natural Science Foundation of China (62001205), Guangdong Natural Science Foundation Joint Fund (2019A1515110138), Shenzhen Science and Technology Innovation Committee (20200925155957004, KCXFZ2020122117340001), Shenzhen-Hong Kong-Macao Science and Technology Innovation Project (SGDX20201109280100), Shenzhen Key Laboratory of Smart Healthcare Engineering (ZDSYS2020081114403009). The authors declare no conflicts of interest.

References

[1] J. Antolík, S. Hofer, J. Bednar, and T. Mrsic-Flogel. Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. PLoS Computational Biology, 12, 2016.

[2] P. Bashivan, K. Kar, and J. DiCarlo. Neural population control via deep image synthesis. Science, 364, 2019.

[3] S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. Tolias, M. Bethge, and A. S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. PLoS Computational Biology, 15, 2019.

[4] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? The Journal of Neuroscience, 25:10577–10597, 2005.

[5] G. P. Das, P. J. Vance, D. Kerr, S. A. Coleman, T. M. McGinnity, and J. K. Liu. Computational modelling of salamander retinal ganglion cells using machine learning approaches. Neurocomputing, 325:101–112, 2019.
[6] J. DiCarlo, D. Zoccolan, and N. Rust. How does the brain solve visual object recognition? *Neuron*, 73:415–434, 2012.

[7] U. C. Dräger. Receptive fields of single cells and topology in mouse visual cortex. *Journal of Comparative Neurology*, 160, 1975.

[8] C. Federer, H. Xū, A. Fyshe, and J. Zylberberg. Improved object recognition using neural networks trained to mimic the brain’s statistical properties. *Neural networks: the official journal of the International Neural Network Society*, 131:103–114, 2020.

[9] E. Froudarakis, P. Berens, A. S. Ecker, R. J. Cotton, F. H. Sinz, D. Yatsenko, P. Saggau, M. Bethge, and A. S. Tolias. Population code in mouse v1 facilitates read-out of natural scenes through increased sparseness. *Nature neuroscience*, 17:851–857, 2014.

[10] K. Fukushima, S. Miyake, and T. Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13:826–834, 1983.

[11] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[12] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[13] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[14] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[15] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195, 1968.

[16] W. F. Kindel, E. D. Christensen, and J. Zylberberg. Using deep learning to reveal the neural code for images in primary visual cortex. *ArXiv*, abs/1706.06208, 2017.

[17] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

[18] D. A. Klindt, A. S. Ecker, T. Euler, and M. Bethge. Neural system identification for large populations separating what and where. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3509–3519, 2017.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84–90, 2012.

[20] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1989.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] T. Macpherson, A. Churchland, T. Sejnowski, J. DiCarlo, Y. Kamitani, H. Takahashi, and T. Hikida. Natural and artificial intelligence: A brief introduction to the interplay between ai and neuroscience research. *Neural Networks*, 2021.

[23] J. Masci, U. Meier, D. Cireșan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, pages 52–59. Springer, 2011.

[24] L. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. Baccus. Deep learning models of the retinal response to natural scenes. *Advances in neural information processing systems (NeurIPS)*, 29:1369–1377, 2016.

[25] A. F. Meyer, R. S. Williamson, J. F. Linden, and M. Sahani. Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation. *Frontiers in systems neuroscience*, 10:109, 2017.

[26] C. M. Niell, M. P. Stryker, and W. M. Keck. Highly selective receptive fields in mouse visual cortex. *The Journal of Neuroscience*, 28:7520 – 7536, 2008.

[27] X. Ran, M. Xu, L. Mei, Q. Xu, and Q. Liu. Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation. *Neural Networks*, 2021.

[28] S. Ravishankar, J. C. Ye, and J. A. Fessler. Image reconstruction: From sparsity to data-adaptive methods and machine learning. *Proceedings of the IEEE*, 108:86–109, 2020.

[29] R. J. Rowekamp and T. Sharpee. Cross-orientation suppression in visual area v2. *Nature Communications*, 8, 2017.

[30] S. Safarani, A. Nix, K. F. Willeke, S. A. Cadena, K. Restivo, G. Denfield, A. S. Tolias, and F. H. Sinz. Towards robust vision by multi-task learning on monkey visual cortex. In *Advances in neural information processing systems (NeurIPS)*, 2021.

[31] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, K. Schmidt, D. Yamins, and J. J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018.

[32] A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[33] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.
[34] W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287 5456:1273–6, 2000.

[35] B. Vintch, J. Movshon, and E. P. Simoncelli. A convolutional subunit model for neuronal responses in macaque v1. *The Journal of Neuroscience*, 35:14829 – 14841, 2015.

[36] E. Y. Walker, F. H. Sinz, E. Cobos, T. Muhammad, E. Froudarakis, P. G. Fahey, A. S. Ecker, J. Reimer, X. Pitkow, and A. S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065, 2019.

[37] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler. Image reconstruction is a new frontier of machine learning. *IEEE transactions on medical imaging*, 37(6):1289–1296, 2018.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[39] M. Weliky, J. Fiser, R. H. Hunt, and D. N. Wagner. Coding of natural scenes in primary visual cortex. *Neuron*, 37:703–718, 2003.

[40] B. Willmore, R. Prenger, M. C. Wu, and J. Gallant. The berkeley wavelet transform: A biologically inspired orthogonal wavelet transform. *Neural Computation*, 20:1537–1564, 2008.

[41] D. Yamins and J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19:356–365, 2016.

[42] T. Yoshida and K. Ohki. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, 11, 2020.

[43] J. Zhang, X. Zhu, S. Wang, H. Esteky, Y. Tian, R. Desimone, and H. Zhou. Visual attention in the fovea and the periphery during visual search. *bioRxiv*, 2021.

[44] C. Zhuang, Y. Wang, D. Yamins, and X. Hu. Deep learning predicts correlation between a functional signature of higher visual areas and sparse firing of neurons. *Frontiers in Computational Neuroscience*, 11, 2017.

[45] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. DiCarlo, and D. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, 118, 2021.

A Experimental settings

In this section, we introduce the details of the settings in our experiments. Firstly, the Table 4 is the number of neural datasets from three brain regions and the Fig. 5 illustrates the training image of the neural stimuli. Secondly, the Table 5 and Table 6 are the settings of the hyperparameters for image reconstruction and neural similarity experiments, respectively.

B Additional experiments

In this section, we provide results from additional experiments. The Fig. 6 and Fig. 7 display the results of image reconstruction using the neural responses on the region 1 and 2. The Fig. 8 shows the number of significant neurons and insignificant neurons of region 1 and 2 in the image reconstruction experiments, respectively.
Table 4: The neural dataset containing neural responses in three brain regions under visual stimulation [1]. The number in bracket is the repeated times.

| Region  | Number of train images | Number of test images | Number of neurons |
|---------|------------------------|-----------------------|-------------------|
| Region 1 | 1800 (1)               | 50 (10)               | 103               |
| Region 2 | 1260 (1)               | 50 (8)                | 55                |
| Region 3 | 1800 (1)               | 50 (12)               | 102               |

Figure 5: The examples of training stimuli in each brain region.

Table 5: The settings of the hyperparameters for image reconstruction experiments.

| Models  | Region 1  | Region 2  | Region 3  |
|---------|-----------|-----------|-----------|
| α : β   | α : β     | α : β     | α : β     |
| DAE-NR1 | 1e-0:1e-5 | 1e-0:1e-3 | 1e-0:1e-4 |
| DAE-NR2 | 1e-0:1e-5 | 1e-0:1e-4 | 1e-0:1e-5 |
| DAE-NR3 | 1e-0:1e-4 | 1e-0:1e-3 | 1e-0:1e-5 |
| DAE-NR4 | 1e-0:1e-5 | 1e-0:1e-4 | 1e-0:1e-4 |

Table 6: The settings the hyperparameters for neural similarity experiments.

| Models  | Region 1  | Region 2  | Region 3  |
|---------|-----------|-----------|-----------|
| α : β   | α : β     | α : β     | α : β     |
| DAE-NR1 | 1e-0:9e-1 | 7e-1:1e-0 | 8e-1:1e-0 |
| DAE-NR2 | 7e-1:1e-0 | 1e-0:1e-2 | 4e-1:1e-0 |
| DAE-NR3 | 1e-0:6e-1 | 1e-4:1e-0 | 1e-0:4e-1 |
| DAE-NR4 | 1e-0:9e-1 | 1e-0:1e-2 | 1e-0:9e-1 |
Figure 6: The reconstructed images with neurons in Region 1. From top to bottom, each row displays the original images (a), the images reconstructed by DAE (b), DAE-NR$_1$ (c), DAE-NR$_2$ (d), DAE-NR$_3$ (e), DAE-NR$_4$ (f), respectively.

Figure 7: The reconstructed images with neurons in Region 2. From top to bottom, each row displays the original images (a), the images reconstructed by DAE (b), DAE-NR$_1$ (c), DAE-NR$_2$ (d), DAE-NR$_3$ (e), DAE-NR$_4$ (f), respectively.
Figure 8: The number of significant neurons and insignificant neurons in region 1 and 2 in the image reconstruction experiments. The threshold for significance is $p \leq 0.05$. 