Robust Decoding of Rich Dynamical Visual Scenes With Retinal Spikes

Yu, Zhaofei; Bu, Tong; Zhang, Yijun; Jia, Shanshan; Huang, Tiejun; Liu, Jian K

DOI:
10.1109/TNNLS.2024.3351120

Robust Decoding of Rich Dynamical Visual Scenes with Retinal Spikes

Zhaofei Yu, Tong Bu, Yijun Zhang, Shanshan Jia, Tiejun Huang, Jian K. Liu

***Abstract*—Sensory information transmitted to the brain activates neurons to create a series of coping behaviors. Understanding the mechanisms of neural computation and reverse engineering the brain to build intelligent machines requires establishing a robust relationship between stimuli and neural responses. Neural decoding aims to reconstruct the original stimuli that trigger neural responses. With the recent upsurge of artificial intelligence, neural decoding provides an insightful perspective for designing novel algorithms of brain-machine interface. For humans, vision is the dominant contributor to the interaction between the external environment and the brain. In this study, utilizing the retinal neural spike data collected over multi trials with visual stimuli of two movies with different levels of scene complexity, we used a neural network decoder to quantify the decoded visual stimuli with six different metrics for image quality assessment establishing comprehensive inspection of decoding. With the detailed and systematical study of the effect and single and multiple trials of data, different noise in spikes, and blurred images, our results provide an in-depth investigation of decoding dynamical visual scenes using retinal spikes. These results provide insights into the neural coding of visual scenes and services as a guideline for designing next-generation decoding algorithms of neuroprosthesis and other devices of brain-machine interface.**

***Index Terms*—Neural decoding, Neural spikes, Image reconstruction, Deep learning, Visual scenes, Video,**

I. Introduction

IN daily life, the sensory information in the external environment is transmitted to the brain, which activates neurons to create a series of coping behaviors. Describing the relationship between stimuli and neural responses is a critical problem for both understanding the mechanisms of the brain and reverse engineering the brain to build intelligent devices and machines, particularly neuroprosthesis [1]. Neural encoding, predicting neural responses to stimuli, has been a focus of visual science of the retina in the last several decades [2]. In contrast, neural decoding, i.e., understanding how the brain detects, interprets, and responds to external stimuli, is less studied [3]. For humans, vision is the dominant contributor to the interaction between the external environment and the brain. With the recent upsurge of artificial intelligence, neural decoding provides a promising perspective for designing novel algorithms to make better neuroprosthesis in the context of brain-machine interface [1], [3].

For the neural decoding problem, there have mainly been two subdivided targeted questions: stimulus classification, or pixel-by-pixel reconstruction images. Pixel-by-pixel reconstruction is more challenging. There has been a focus on visual reconstruction from neural signals. The functional magnetic resonance image activity of the visual cortex has been widely used in this field [4], [5], [6], [7], [8], [9], [10], [11], while most recently, fine neural signals including neural spikes [12], [13], [14], [15], [16], [17], [18], [19], [20], and calcium imaging signals [21], [22], [23] have also been studied on.

The visual pathway starts from the retina, where the light energy is transferred into the neuronal signal, goes through lateral geniculate nucleus (LGN) and terminates in the visual cortex. The retina plays an important role in the whole visual system, in which all visual information is represented by spikes of a population of the retinal ganglion cells (RGCs) and then transmitted to the downstream regions. As the retina does not receive feedback from the higher part, the RGC population can be thought of as a computational device to process visual information holistically. Much effort has been made in the research of the encoding mechanism of RGCs, and various neural mechanisms of the retinal visual computation have been discovered according to its neurons, and neural circuitry [24], [2], [25], [26], [27], [28]. Besides, numerous encoding models have been developed based on different properties of neurons and neural circuits in the retina [29], [30], [31], [32], [33], [34], [35].

From the perspective of decoding visual stimuli, most decoding methods over RGC depended on linear methods because of the interpretability and computational efficiency. On the other hand, linear decoding methods are usually derived from the decoding scenes of spatially uniform white noise stimuli [18], [15], [36]. These types of decoding methods are only capable of decoding the coarse structure of natural images, hardly recovering fine natural scenes [36], [18].

For more accurate decoding of complex visual scenes, nonlinear methods have been introduced as the backbone of the recent decoders. Optimal Bayesian methods have been employed to decode white noise effectively, but not as well when applied on large neural populations [4], [5], [31]. Some works incorporated Bayesian inference for decoding natural images [4], while the demand of key prior information and computationally expensive process for accurate prior of natural

This work was supported in part by the National Natural Science Foundation of China under Grant 62176003, 62088102, and in part by Royal Society Newton Advanced Fellowship of UK under Grant NAF-R1-191082.

Z. Yu, T. BU, S. Jia, and T. Huang are with the Institute for Artificial Intelligence, Peking University, Beijing 100190, China, and also with the National Engineering Research Center of Visual Technology, Peking University, Beijing 100190, China. (e-mail: yuzf12@pku.edu.cn, putong30@pku.edu.cn, jiashsh@stu.pku.edu.cn, tjhuang@pku.edu.cn)

Y. Zhang is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the National Engineering Research Center of Visual Technology, Peking University, Beijing 100190, China. (e-mail: yijzhang@sjtu.edu.cn)

J. Liu is with the School of Computer Science, University of Birmingham, Birmingham, B15 2TT, United Kingdom. (e-mail: j.liu.22@bham.ac.uk)

scenes have made these methods not so efficient.

In recent years, deep learning techniques [37], [38], [] Although some of these decoding methods ameliorated the image quality of visual stimuli [12], [14], [17], [39], [40], results above were attained with limited experimental physiological data or artificial spike data. Recent studies used a scale of one thousand macaque RGCs for decoding static natural images [41], whereas, for dynamical visual scenes, it is still in demand for efficient decoding methods in which large populations of neurons and complex natural scene stimuli would be leveraged. In particular, efficient decoding using a single trial of neural responses is more needed, given that no prior work has paid attention to noise corruption in neural decoding.

In this work, utilizing the RGC spike data collected over multi trials with visual stimuli of two movies with different levels of scene complexity, we extended our recent model [17], [42] and proposed an end-to-end deep learning model inspired by the U-net used in image segmentation. To quantify the decoded visual stimuli, six different metrics for image quality assessment are used for a comprehensive inspection of decoding results. For our two movie stimulus RGC datasets, the salamander & tiger datasets, we first studied the sparse encoding property of RGC through decoding with a different number of cells chosen from the datasets. Surprisingly, the reconstruction performance is not significantly better when the cell number increases, also depending on the assessment metrics used and the specific movie stimuli decoded. Then, we studied the characteristics of cross-trial spike trains whereby observing the decoding performance of models trained with single-trial or multi-trial spike data. By adding various types of artificial noise into real neuronal responses, we studied the noise immunity of our neural decoding model. Besides, we also compared the similarity of the decoding results with frequency low-pass filtered visual stimuli. Meanwhile, the comparison between the frequency low-pass filtered results, and those of Gaussian smoothed low-pass is used to establish the relationship between the neural decoder and the Gaussian low-pass filter with different sizes of Gaussian kernels. Our results provide an in-depth study of decoding dynamic visual scenes using neural spikes considering the effect of biased assessment metrics, the robustness of a single trial, the immunity of noisy spikes, and the similarity of low-pass filtering. These results shed insight into neural decoding and services as a guideline for designing next-generation decoding algorithms of brain-machine interfaces.

II. Methods

A. Decoding model

We built a deep neural network to decode RGC responses. The holistic decoding process consists of two stages: the signal converter part, which samples neural signals to pixels, and the U-Net part, which is typically referred to as the auto-encoder. The input spike signal is an array consisting of M vectors, and each length is N, where M is the number of stimuli and N is the number of all recorded RGCs. To map every cell response to every pixel in reconstruction frames, we first employed a multi-layer perceptron to transform the input spike signal into a vector with the same size as the target visual stimuli. In our experiments, the down-sampled stimulus frames with 90×90 pixels were taken as reconstruction objectives, which are ensured to be covered within the object salamander's retina receptive fields. Following an auto-encoder analogous to U-Net, the transformed vector is used as input in the U-Net part, where the reconstructed visual stimuli are generated. In the first half of the autoencoder, the input vector is convolved and down-sampled to completely extract the signal features. This step is inspired by the concept of a full convolutional network, which is typically used for image segmentation. The network begins upsampling and convolution in the second half and eventually recovers a reconstructed frame of the same size as the visual stimulus. It should be noticed that the 'skip connection' structure has been added to the network structure of the autoencoder component. Skip connections across network layers of different depths can combine low-level (shallow network) and high-level (deep network) features, which can both be captured in visual stimuli, during the reconstruction process (see Fig 1). As a result, under the circumstance of small datasets, more specifics and precise placements can be deciphered. Between the different layers of the network structure, three skip connections were used. Additionally, we used the batch normalization layer behind the activation layer after the convolution process to obtain smooth gradient propagation. Also, a spatial dropout layer was added to the U-net component to minimize overfitting. We employed the back-propagation algorithm to perform end-to-end training with the mean square error (MSE) as the objective function in order to optimize our network. Our decoder can decode visual stimuli directly from the neural responses due to the end-to-end characteristic, eliminating the need for intermediary processing.

The presented frame sizes of salamander and tiger datasets are the same, i.e., 360×360 pixels. We used the down-sampled 90×90-pixel stimuli as the reconstruction objective. The structure of the decoding models for salamander and tiger datasets are the same, while the input spike signals are different. For the signal converter part, the input shape was set as the neuron response array size in the respective dataset. The middle dense layer contains 512 neurons, and operations for batch normalization, activation, and dropout are performed in that order. The output shape of the signal converter is set as the target reconstructed frame size, i.e., 8100 (90×90). The entire information processing can be divided into two parts for the U-Net autoencoder component. We used convolution and down-sampling in the first stage to process and reduce the size of the input to the target size. The kernel sizes of four layers in the first stage are (90,90,64), (30,30,128), (15,15,256), (15,15,512). The convolution kernels are (2,2) for layers with (15,15) size and (3,3) for other layers. Another four layers in the second stage correspond to the reversed structure order in their respective first stages. The second stage's function is to upsample the down-sampled frames to the target reconstructed size. We can see from the above-mentioned structure change that the stride operation size in our model is (2,2) for down and up-sampling. The downsampling operation in the first

stage is realized by the MaxPooling2D function in TensorFlow, while the upsampling operation in the second stage is realized by the Upsampling2D function. The MaxPooling2D function in TensorFlow implements the down-sampling operation in the first stage whereas the Upsampling2D function in the second stage implements the up-sampling process. The activation function in our overall model framework is ReLU. Additionally, a SpatialDropout2D layer was added behind the batch normalization layer in the middle layers (those with the smallest kernel size) to reduce overfitting.

We trained our decoding model with the Adam method and batch size of 64 to update network parameters, while the learning rate of the decoding process is 0.0005. The training epochs are controlled by the early-stop mechanism, i.e., when the validation error (MSE in the test set) is not decreasing, the model parameter update will be stopped. The model was implemented with Keras deep learning library, Tensorflow as backend, employed on Nvidia v100 super graphics card. The Learning rate-customized Adam method was used to train the model [43].

B. Datasets

Our decoder was used to reconstruct natural visual scenes as part of natural movies, specifically, the tiger and salamander movies. The salamander movie consists of salamander swimming which includes 1800 frames, and the tiger movie includes 1600 frames. Both movies are at a frame rate of 30 Hz. The RGC response data were recorded from the retinal cells of salamanders in previous experimental works [44]. Briefly, The spike train of each RGC was collected from isolated retinas put in a recording chamber with 60- or 252-channel multielectrode array. Visual stimuli were presented to the photoreceptor layer through a telecentric lens above the retina. Each frame covered 2700*2700 μm area on the retina with a spatial resolution of 360x360 pixels. Each frame from the video was displayed with a rate of 30Hz, while 1218 RGC responses for the salamander movie and 1407 RGC responses for the tiger movie were recorded, and binned into the same presentation time period as each frame display. As a result, 1800 (salamander) and 1600 (tiger) spike counts in each RGC spike train were collected for every recording trial. In this work, 18 trials for both datasets were used for a series of analyses. The distribution of cells with maximal firing spikes and the firing rasters of 9 trials are illustrated in Fig S1.

Datasets were divided into 9:1, 90% for the training set, and 10% for the testing set. We used the input spike signal as an array composing M vectors, and each length is N, M is the number of all stimulus frames, and N is the number of all recorded RGCs where N is 1218 in the salamander dataset and 1407 in tiger dataset. In traditional all-cell neural decoding, N remains unchanged.

In the main decoding experiments of this work, specific methods applied to RGC spike data by means of a neural decoding model are introduced as follows:

- Sparse encoding of RGCs. We tested the reconstructions with different scales of RGCs, i.e., changing the N of the input array for the decoding model. Different numbers of RGCs were randomly chosen from all cells. Then, these responses of different scales were fed into our decoding model to observe the resulting reconstructions.
- Characteristics of cross-trial RGC spike trains. Here, we divided 18 spike trials into two parts: 9 trials for training and 9 trials for testing. The single-trial model was trained with one trial from training trials, while the multi-trial model was trained with 9 trials. The 9 trials in the testing part were reserved for testing the performance of different decoding models.
- Information uniformity of RGC subsets. We randomly chose 100 cells from all cells. Then these 100 cells were used to train the decoding model. We have done this process 10 times, and these models were used for later analysis.

C. Image reconstruction metrics

For stimulus reconstruction, it is intractable to perceive the differences between reconstructed images and original stimuli. For this reason, we evaluated the quality of reconstruction using different image assessment metrics. Referring to the development of the image quality assessment field, six full-reference metrics were applied to compare reconstructed frames with original stimuli. This set of metrics would evaluate the similarity from different aspects. Individual characteristics of these metrics are briefly introduced as follows.

1) Mean Square Error (MSE): MSE equals the final expectation of the squared error between desired and original values. Given an original image I and its reconstruction K, MSE is defined as:

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{1}$$

2) Peak Signal-to-Noise Ratio (PSNR): The PSNR (unit is decibel) is defined as:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{2}$$

MAX_I is the maximum possible pixel value of the original image, that is 255. Besides, the larger PSNR, the better the image quality, and the range of PSNR is not limited.

3) Structural Similarity Index Metric (SSIM) [45]: SSIM is inspired by the assumption that the human visual processing system is competent to extract structural information in scenes highly adaptively. SSIM index is calculated on various windows of an image. Luminance (l), contrast (c), and structure (s) are included when measuring two windows x and y. SSIM value is in the range $[0, 1]$. The more similar the reconstructed image is to the original images, the higher the SSIM value.

$$\begin{cases} l(x,y) = \dfrac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\ c(x,y) = \dfrac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \\ s(x,y) = \dfrac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \end{cases} \tag{3}$$

μ_x and σ are the mean and the variance for the corresponding window. $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are the constant, $c_3 = \frac{c_2}{2}$. L is located in the range of pixel value range, i.e., $[0\,255]$. $k_1 = 0.01$ and $k_2 = 0.03$. We can get the SSIM equation as

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^{\alpha} \cdot [c(\mathbf{x}, \mathbf{y})]^{\beta} \cdot [s(\mathbf{x}, \mathbf{y})]^{\gamma} \tag{4}$$

where α, β and γ equal to 1.

4) Most Apparent Distortion (MAD) [46]: MAD is intended to rate image quality from two aspects: detection-based perceived distortion in high-quality images, and appearance-based perceived distortion in low-quality images. The combination of these two measures is regarded as effective in predicting subjective ratings of image quality.MAD must be a positive number. The image quality declines with increasing MAD values. The equation is given by:

$$\mathrm{MAD} = (d_{\text{detect}})^{\alpha} (d_{\text{appear}})^{1-\alpha} \tag{5}$$

d_{detect} and d_{appear} are measured respectively by specific processes for distortion in high-quality and low-quality image levels. The entire degree of distortion determines the weight $\alpha \in [0, 1]$.

5) Feature Similarity Index (FSIM) [47]: FSIM is designed based on the fact that the human visual system (HVS) understands an image in terms of its low-level features. Specifically, the phase congruency (PC), a dimensionless measure of the significance of a local structure [48], is taken as the primary feature in FSIM. Given the contrast invariance property of PC, the image gradient magnitude (GM) is used as the second feature in FSIM. The range of FSIM value is the same as SSIM. FSIM is defined as:

$$\mathrm{FSIM} = \frac{\sum_{\mathrm{x} \in \Omega} S_L(\mathrm{x}) \cdot PC_m(\mathrm{x})}{\sum_{\mathrm{x} \in \Omega} PC_m(\mathrm{x})} \tag{6}$$

where $S_L(\mathbf{x}) = [S_{PC}(\mathbf{x})]^{\alpha} \cdot [S_G(\mathbf{x})]^{\beta}$, usually $S_L(\mathbf{x}) = \mathrm{S}_{PC}(\mathbf{x}) \cdot \mathrm{S}_G(\mathbf{x})$ for the sake of clarity. $S_L(\mathbf{x})$ represents the similarity at each location $\mathbf{x}$ combining PC similarity $S_{PC}(\mathbf{x})$ and GM similarity S_G, whose computation process is removed here for briefness. $PC_m(\mathbf{x}) = \max(PC_1(\mathbf{x}), PC_2(\mathbf{x}))$ is used as weight for the significance of $S_L(\mathbf{x})$ in total similarity between two images.

6) Gradient Similarity (GSM) [49]: Gradients are well known for conveying significant visual information and are essential for scene understanding. Besides, gradients can also be used to capture structural and contrast changes. In fact, luminance changes have a significant impact on image quality. GSM, whose value is in the range $[0, 1]$, integrates changes in luminance and contrast structure through an adaptive method to obtain a holistic image quality score. The quality of the reconstructed image improves with increasing GSM value. The proposed gradient similarity is defined as:

$$g(x, y) = \frac{2 g_x g_y + C}{g_x^2 + g_y^2 + C} \tag{7}$$

where g_x and g_y are the gradient values of the central pixel of image x and y. C is the small constant to avoid the zero-value denominator.

D. Shuffle noise & Gaussian noise in spikes

We used two types of noise, Gaussian noise, and shuffle noise. to disturb the RGC spike signal. Concretely, Adding Gaussian noise in our work was putting zero-mean Gaussian distribution values to each RGC normalized spike signal. And, adding shuffle noise here was shuffling the responses of specified-percentage cells of all RGCs. The illustrations of these two types of noise are shown in Fig S3 and Fig S4. Then, we tested the performance of the noise-disturbed RGC responses on our decoding models, i.e., the single-trial model & multi-trial model.

E. Low-pass filters

Two types of low-pass filters were used in this work for decoding analysis and comparison, i.e., frequency low-pass filter & Gaussian low-pass filter. These two types of low-pass filters are both two-dimensional image filters, with a frequency filter in the frequency domain and a Gaussian filter in the time domain.

Here we briefly introduce their filter principles. Whereas convolution is used to evaluate filters in the spatial domain, filtering is implemented in the frequency domain by multiplication. The image in the frequency domain is multiplied by the filter's frequency response in the frequency domain. Thus, the filter transfer function is simply a matrix of the same size as the image. The complex values of the image in the frequency domain are simply multiplied element by element with the filter transfer function to amplify or attenuate specific frequencies of the image. The inverse Fourier transform is then used to generate the output image. The frequency filter transfer function equation is defined as:

$$H_{FL}(u, v) = \begin{cases} 1 & \text{if } D(u, v) \leq D_0 \\ 0 & \text{if } D(u, v) > D_0 \end{cases} \tag{8}$$

Where $D(u, v)$ is the distance of each element of the transfer function to the origin (0,0), and D_0 is the stop frequency. The Gaussian filter transfer function equation is defined as:

$$H_{GL}(u, v) = e^{-\left(D(u,v)^2\right)/2\sigma^2} \tag{9}$$

where $sigma$ determines the size of the Gaussian filter.

III. Results

A. Reconstruction movie frames from RGC spikes

We first developed a deep neural network model to reconstruct stimulus frames drawn from two presented movies (the main contents are salamander swimming and tiger strolling) from the recorded neural spike signal in a salamander's retina. As illustrated in Fig 1, our model is extended from our recent model [17] while using a new feature, the U-Net autoencoder with skip connections (Concatenate 1, 2, and 3 as part of feature fusion in Fig 1). We then used the publicly released experimental neural data collected in a previous study [44], including two datasets: salamander and tiger datasets. The visual stimuli in the salamander dataset include 1800 frames drawn from a movie whose main context is a salamander swimming scene, while the stimuli in the tiger dataset include

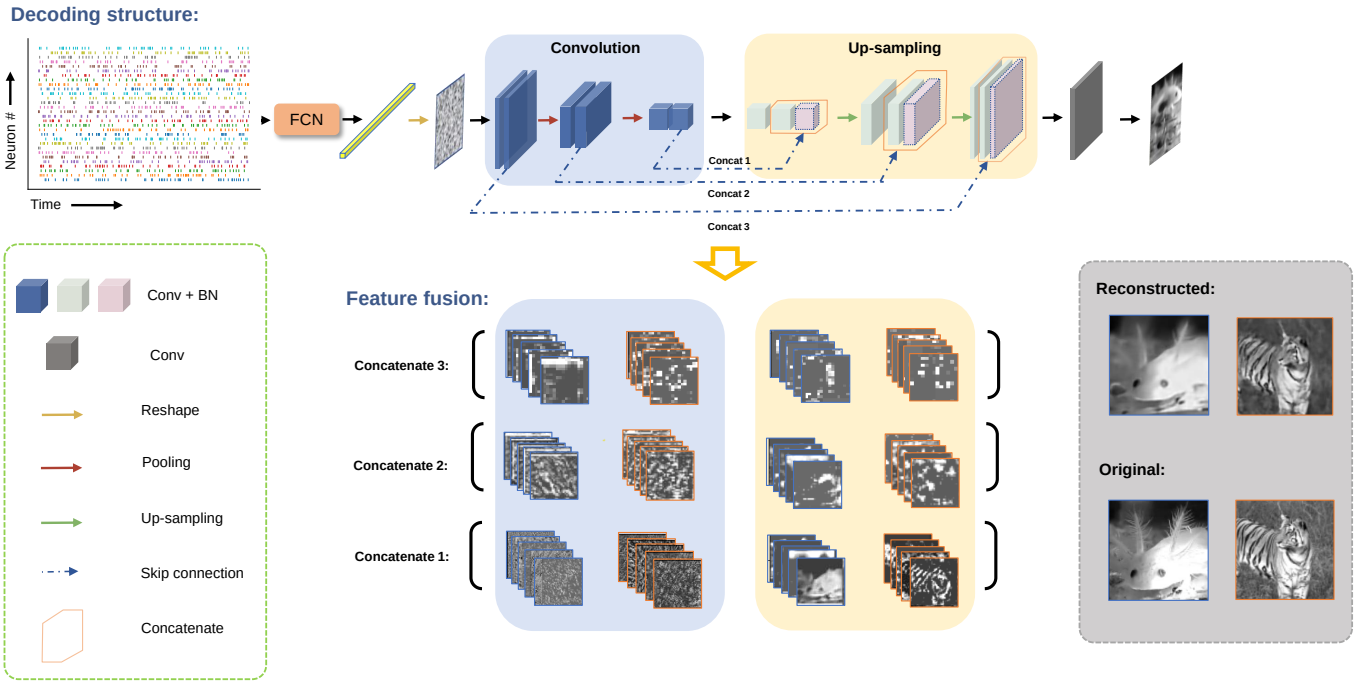

Fig. 1: Illustration of the neural decoding model. The input RGC spike train is converted to one vector whose length equals the product of the length and width of stimulus frames through a multi-layer fully connected neural network (FCN). Then, the output is transferred to a U-Net-like structure autoencoder. The whole autoencoder part can be regarded as the two processes of ”convolution (down-sampling)” and ”up-sampling” (marked with light blue and light orange). Cubes of different colors indicate feature maps with different sizes processed through different layers in the network, i.e., convolution (Conv), batch normalization (BN), spatial dropout, and concatenate layers. Remarkably, three skip layer connections are added during the autoencoder process. The respective features are merged by means of skip connections (Concatenate 1, 2, and 3 in the figure). Here, we chose two frames from two movie datasets for reconstruction illustration. It is observed that the low-level features (shallow layer) and high-level features (deep layer) during reconstruction complement each other for a better reconstruction.

1600 frames drawn from a movie whose main context is a tiger strolling scene. Additionally, 1218 RGC responses for the salamander dataset and 1407 RGC responses for the tiger dataset were recorded with multiple trials. In this work, 18 random trials of two datasets were used in the overall analysis experiments. Sample cell responses and overall responses in all cells were similar in both movies (Fig S1).

As is well known, the number of input samples has a great impact on the performance of decoding models, especially for neural decoders based on deep learning, which are highly demanding on the amount of input data. From the perspective of decoding with neuronal responses, the number of cells usually contributes to a better decoding performance [17], [21]. Here, we use different scales of randomly chosen RGCs to train the decoding model. To quantify the performance of reconstructions without bias, we used six different metrics (MSE, PSNR, SSIM, GSM, FSIM, and MAD, see Methods) for image quality assessment as a reconstruction index. Besides common measures of MSE, PSNR, and SSIM, we selected three more novel measures (GSM, FSIM, and MAD), since it is still debated which metric gives a more reasonable and accurate description of image quality [50], [51], [52], [53].

The resulting reconstructed movie frames and assessment metrics are shown in Fig 2. From the reconstructions and metric changes, using more cells ameliorates the decoding model performance in general. However, the changes in metrics are specific. In particular, MAD failed to tell the difference between the two movies. From the reconstructed frames, reconstructions are blurred with 100 cells and start to be distinguishable from about 400 cells. The reconstructions of the tiger movie are not as good as those of the salamander movie. The contents of reconstructed frames corresponding to one original frame at different RGC scales in the tiger movie are more varied than those in the salamander movie.

Different scenes in the tiger movie are not as uniform as those in the salamander movie, which has an impact on the decoding model performance, presumably due to the scene complexity of the tiger movie being higher than the salamander movie [54].

We also studied the uniformity of decoded information distribution in RGC population response. With randomly chosen 100 cells from all cells for for 10 times, we trained 10 decoding models using these 10 subset cells in both datasets. The metrics are illustrated in Fig S2, where the stability of 10 subset models with the tiger dataset is more varying than the salamander dataset.

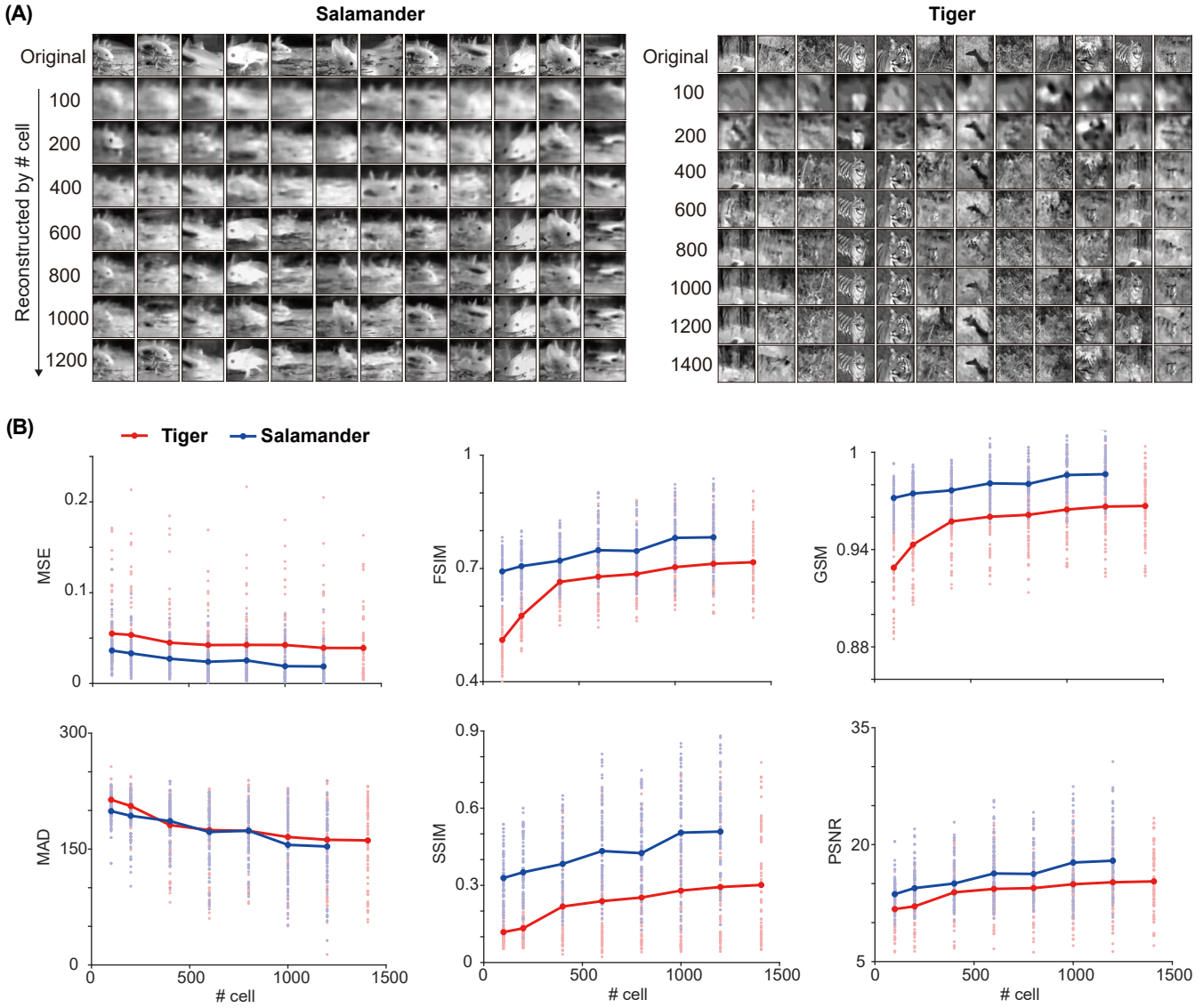

Fig. 2: Reconstructions with different numbers of cells. (A) Original and reconstructed image frames from salamander (left) and tiger (right) movies using different numbers of cells. (B) Change of assessment metrics with different cells randomly chosen in datasets. Solid lines indicate the mean values of all testing frames. Data points are metric values of half of the testing frames (90 for the salamander movie and 80 for the tiger movie).

B. Compare with other decoding methods

We compare our method with other existing decoding models. Here we chose the linear decoding model (LD) and the spike-image decoder (SID) [17] for comparison. We selected one trial in both datasets and split them by 9:1 into train and test sets. For the linear decoding model, we converted the spike trains to the reconstructed stimulus by linear transformation [12]. Considering the sparsity of spike trains, we added the L1-regularized in linear regression and used Lasso regression with a regularization parameter of 1 to solve the linear decoder. For the SID method, we also designed a SID of the same depth as our model, and used the same training parameters for training and testing. The final evaluation metrics are shown in Tab I. The bold metrics values in this table are the best of the three. We can find that the image quality generated by our decoding model outperforms other models in all five metrics and both tiger and salamander datasets. This demonstrates the generality and superiority of our model.

C. Cross-trial decoding

Benefiting from the multi-trial recordings in physiological experiments, we are able to incorporate these trial-varying spike trains under the same visual stimulus conditions to explore the overall characteristics of neural responses to the same stimuli and the inner relationship between the multi-trial neural signals. Here, we first trained our decoding model with single-trial (spike patterns) and multi-trial RGC spike data, respectively. After the single-trial model (decoding model trained with one single-trial spike train) and multi-trial model

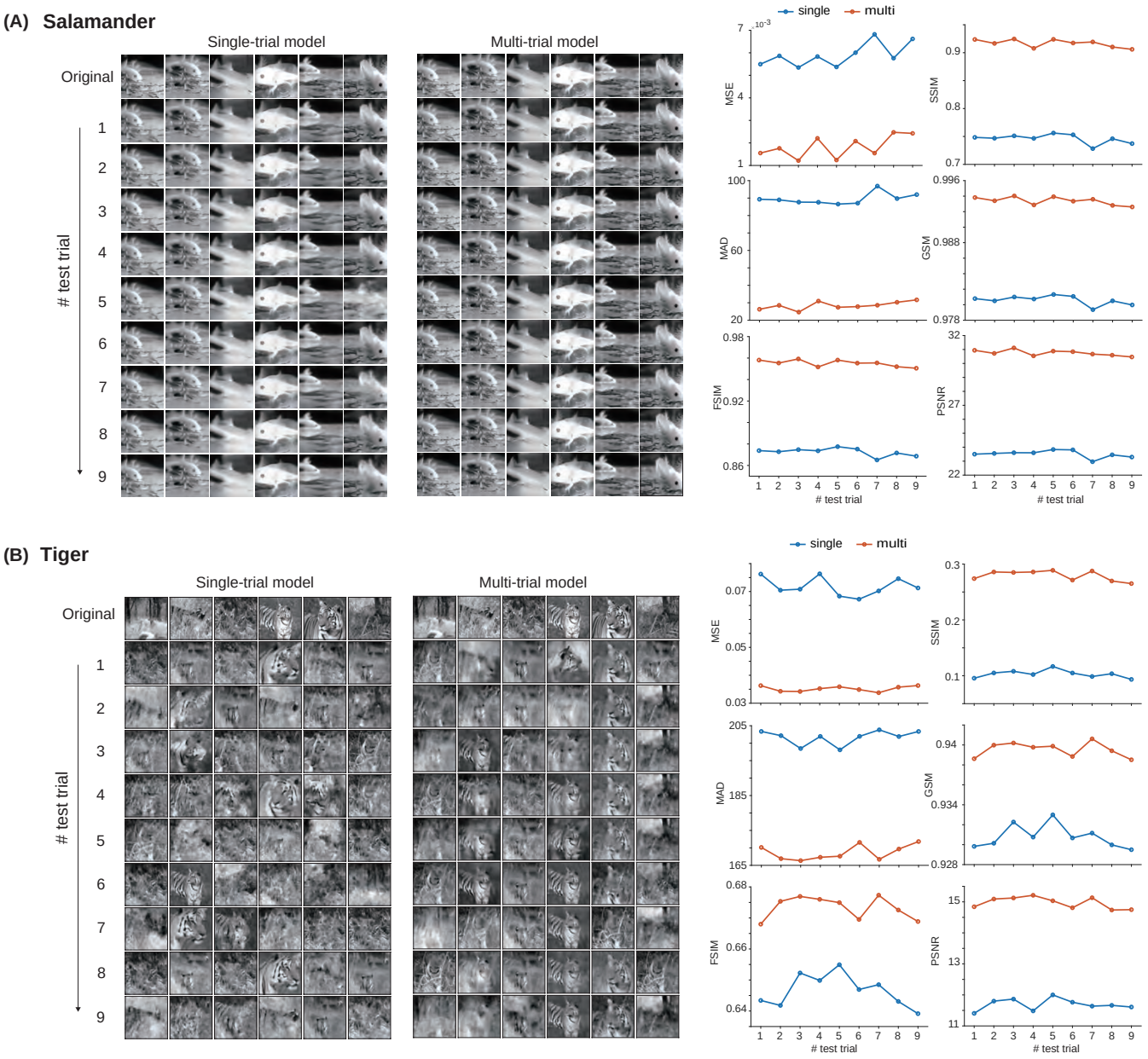

Fig. 3: Cross-trail test metrics of single-trial and multi-trial model. (A) (Left) Original and reconstructed frames of single/multi-trial model over 9 different test trials. (Right) Change of metrics over different test trials. (B) Similar to (A) but for the tiger movie. Markedly, reconstructed frames of single-trial model in the tiger dataset are far worse than those of multi-trial model.

(decoding model trained with multi-trial firing rates, specifically, 9 trials) were finished, we then tested the applicability of cross-trial neural spike trains over these two decoding models. Specifically, we put another 9 trials into a single-trial model and multi-trial model, respectively, to see their decoding performance over multi-trial spike signals. These experiments were carried out over both datasets, and the results are illustrated in Fig 3.

As shown in the reconstructed frames, the overall reconstructions of single/multi-trial models in the salamander dataset are better than those in the tiger dataset. Besides the image quality, the cross-trial reconstruction stability in the salamander is better than that in the tiger. In the tiger movie, the contents of reconstructed frames corresponding to the same original frames but from different spike trials are varied, especially for the multi-trial model. On the contrary, the contents of cross-trial reconstructed frames of both single & multi-trial models in the salamander dataset remain consistent. We then named it as a ”wrong match”, i.e., the image quality of the reconstructed object is clear enough, but the content of that is not in accord with the original object, which is an indication of unstable decoding. From the aspect of the single/multi-trial model, the metric change trends show that the multi-trial model has a strong generalization capability compared with the single-trial model.

TABLE I: Metrics value for reconstructed images by different methods

Method	MSE	PSNR	SSIM	GSM	FSIM	MAD
Salamander						
LD	0.0331	12.57	0.329	0.946	0.677	206.6
SID	0.0400	9.69	0.218	0.931	0.619	186.8
Ours	**0.0192**	**16.76**	**0.489**	**0.962**	**0.755**	**166.4**
Tiger						
LD	0.0553	10.52	0.084	0.928	0.634	215.5
SID	0.0404	8.11	0.117	0.894	0.521	202.5
Ours	**0.0315**	**16.78**	**0.378**	**0.953**	**0.745**	**140.4**

D. Noise immunity of RGC decoding model

In the basic reconstruction experiment described above, the neuron order of the input RGC response array was fixed, i.e., the position of an individual cell was the same during the training and test process of the decoding overall. We added the shuffling noise to the test spike data in the tiger dataset to test the noise immunity of our decoding model. The shuffle noise percentage means the portion of cells whose positions were disordered in the test set (Fig S3). We also added Gaussian noise to the RGC responses of the tiger dataset. The Gaussian noise was set to zero mean value with varying sigma values to see the influence of the fluctuation of Gaussian noise (Fig S4).

From the reconstructed frames in Fig 4, the effect of shuffle noise on reconstruction in a salamander movie increases gradually as the percentage grows. Two phenomena have appeared under shuffle noise, i.e., fuzziness and wrong match. Notably, there were more wrong matches for the multi-trial model compared with the single-trial model. The overall reconstructions of single/multi-trial models in the tiger movie are not as satisfied as those in the salamander movie. Still, both fuzziness and wrong matches happened, while more fuzziness in the single-trial model and more wrong matches in the multi-trial model. Interestingly, most reconstructed frames of single/multi-trial models in both the salamander and tiger movies turned out to be several frames with similar contents under 100 percent shuffle noise.

From the metrics aspect, the shuffle noise effect on the single/multi-trial model in the salamander movie is non-linear. Specifically, the trends of metrics are similar with *sigmoid* function: before about 40 percent, the effect is little, while the metrics change rapidly after that. For both single and multi-trial cases, the variances of metrics are not large (tiger slightly larger than salamander). On the other hand, the shuffle noise effect on metrics of the multi-trial model in the tiger movie is linear, while only a slight influence is reflected on the single-trial model.

On the whole, shuffle noise results in the wrong match, with slight fuzziness in image quality. When the neuronal responses are totally shuffled, most of all reconstructions will tend to evolve into several frames with similar contents.

As for Gaussian noise, seen from the reconstructed frames in Fig 5, compared with the shuffle noise situation, most deteriorating of reconstructed frames is in image details, though the overall reconstructions get worse as the Gaussian noise increases. The wrong match phenomenon is obvious in tiger movie reconstructions. As more fuzziness becomes worse when Gaussian noise increases, the reconstructed frames turn out to be similar in the tiger movie.

From the metrics aspect, different from shuffle noise, Gaussian noise results in larger variances of metrics as the sigma increases. The effect of Gaussian noise on metrics is non-linear in the salamander movie. Specifically, little degradation in the front part of the sigma growth process and a larger degradation in the latter part. Contrary to the coordinated change of single/multi-trial models in the salamander movies, variances of single/multi-trial models in the tiger movies are varied at the same sigma. The Gaussian noise effect on the multi-trial model is non-linear in the tiger movie, while the variances are not as large as those in the salamander movie.

It seems strange that the metrics of the single-trial model in the tiger movie get improved as the sigma increases. We have performed the noise experiments repeatedly and ruled out the model problem or experiment setting. We thought the poor generalization capability of the single-trial model in the tiger movie should be responsible for that. In the context of impractical reconstructions, the slight change of metrics makes no sense.

E. Decoding low-pass filtered images

Recent studies suggested stimulus images decoded from neural responses of the retina resemble low-pass filtered original images [41]. Here we examined this relationship between our decoded frames and low-pass filtered ones. We processed the original frames in our datasets through low-pass filters at different frequencies, in which the low-frequency filter would blur the details of images while keeping the global feature. We quantified the similarity and computed the image quality metrics between them and low-pass filtered images (1-30 Hz in Fig 6 and 1-150 Hz in Fig S5). There is a preferred low frequency around 3-10 Hz where the reconstruction measured by metric values is peaked depending on specific assessment metrics (except MSE). Such a preferred frequency is presumably due to the filtered computation from upstream cells in the retina.

To further investigate this, we compared the effect of frequency low-pass filters with Gaussian blurred filters on frames in our datasets. Notably, the smooth filter size in Gaussian low-pass filtering is usually taken as the approximated size of receptive fields of upstream cells and RGCs in the retina. We computed the image quality metrics of frames low-filtered at different frequencies and Gaussian sigma values, and put them in the same coordinate system so as to compare their low-pass effects on original visual stimuli, which are shown in Fig 7. While the Gaussian low-pass filter performs the low-pass operation in the spatial domain, the filtered frames are similar to frequency low-passed ones as the sigma increases. At the same coordinate system, the effect of two low-pass filters on the original frames is apparent. Both frequency and sigma have a non-linear influence on original frames. The difference is that the metrics change faster as the frequency decreases in the frequency low-pass filter, while the metrics

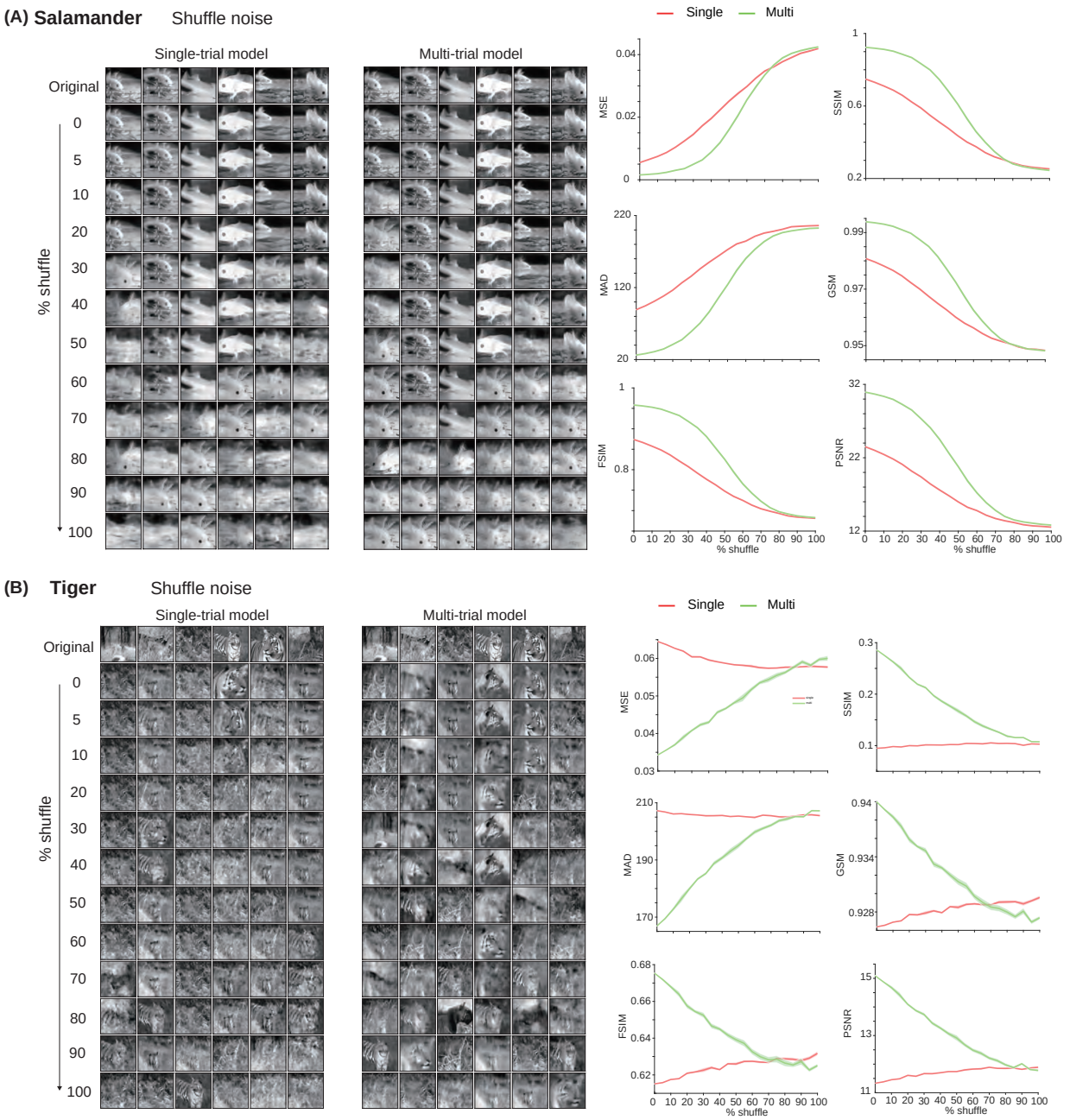

Fig. 4: Shuffle noise effect on decoding movies. (A) (Left) Original and reconstructed frames of a salamander movie with different levels of shuffle noise in a single-trial/ multi-trial model. (Right) Change of reconstruction metrics. Decoding was run 10 times, with the solid lines as the average values and the spread areas as standard errors. (B) Similar to (A) but for the tiger movie.

change slower as the sigma value grows. It can be seen that under the same level of reconstruction, the interaction between low-passed filters and Gaussian filters is arranged between 3-10 Hz as well. The values of Gaussian sigma are comparable to the receptive fields of upstream bipolar cells.

IV. Discussion

Here using a neural network decoder and spike datasets under the stimulation of two dynamic movies, we systematically explored the robust decoding with different scenarios of single and multiple trials, the effect of noise, and low-frequency filters. The RGC spike data in this work was collected over multi trials, with visual stimuli of two movies, including thousands of consecutive frames. Specifically, neuronal responses of a large scale of more than one thousand RGCs were recorded. We studied the characteristics of cross-trial spike trains whereby observing the decoding performance of models trained with single-trial or multi-trial spike data. Researchers usually study the raw spike train data directly or extract stimulus-evoked responses from spike train signal with various methods [55], [56], [57], [58]. Hereby we showed that

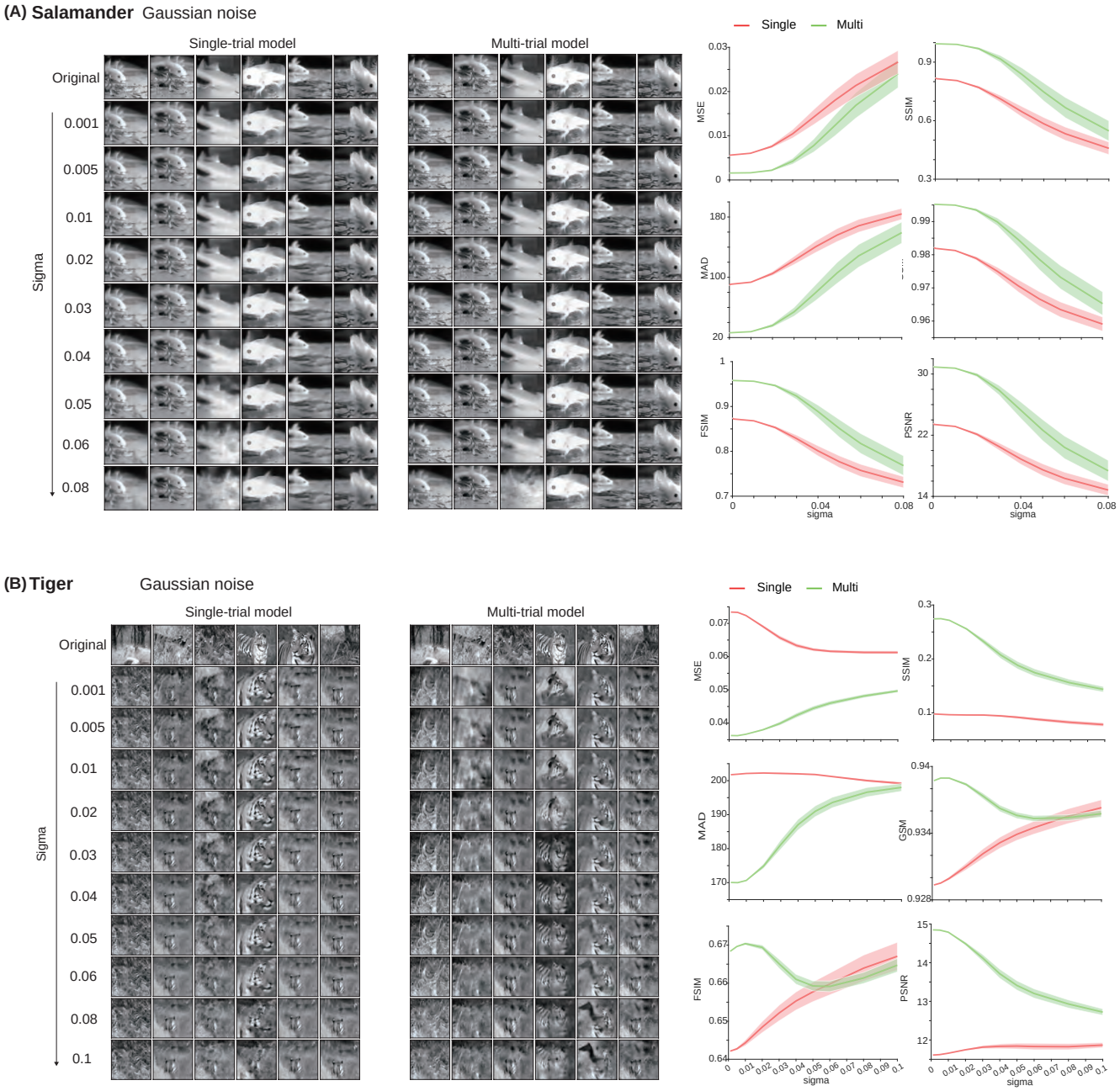

Fig. 5: Gaussian noise effect on decoding. (A) (Left) Original and reconstructed frames of a salamander movie with different levels of shuffle noise in a single-trial/multi-trial model. (Right) Change of reconstruction metrics. Decoding was run 10 times, with the solid lines as the average values and the spread areas as standard errors. (B) Similar to (A) but for the tiger movie.

decoding with single trial data can also give robust results. An interesting future direction is to combine encoding and decoding to study the effectiveness of the model as a whole. Additionally, the same decoding model can be used to evaluate the capabilities of different encoding models. Furthermore, it is worth exploring the direction of using spiking neural networks [59], [60] as a replacement for artificial neural networks to achieve encoding and decoding.

As known, the retinal ganglion cells transmit information about visual stimuli to the cortex via the thalamus. Specifically, the final informativeness conveyed from the retina is reliant on the optic nerve responses and the transmission efficacy of the next part LGN, i.e., how much information is decoded to the cortex. In the whole process, the corruption derived from endogenous noise and the external environment would make a difference in information fidelity. Despite this negative effect, biological visual systems could usually remain robust and even make use of it. Some have focused on the robust information propagation problem from the perspective of neural encoding. With added noise in the first layer of the encoding model, it was found that the covariance structures optimized information propagation through noisy circuits [61]. It is possible for

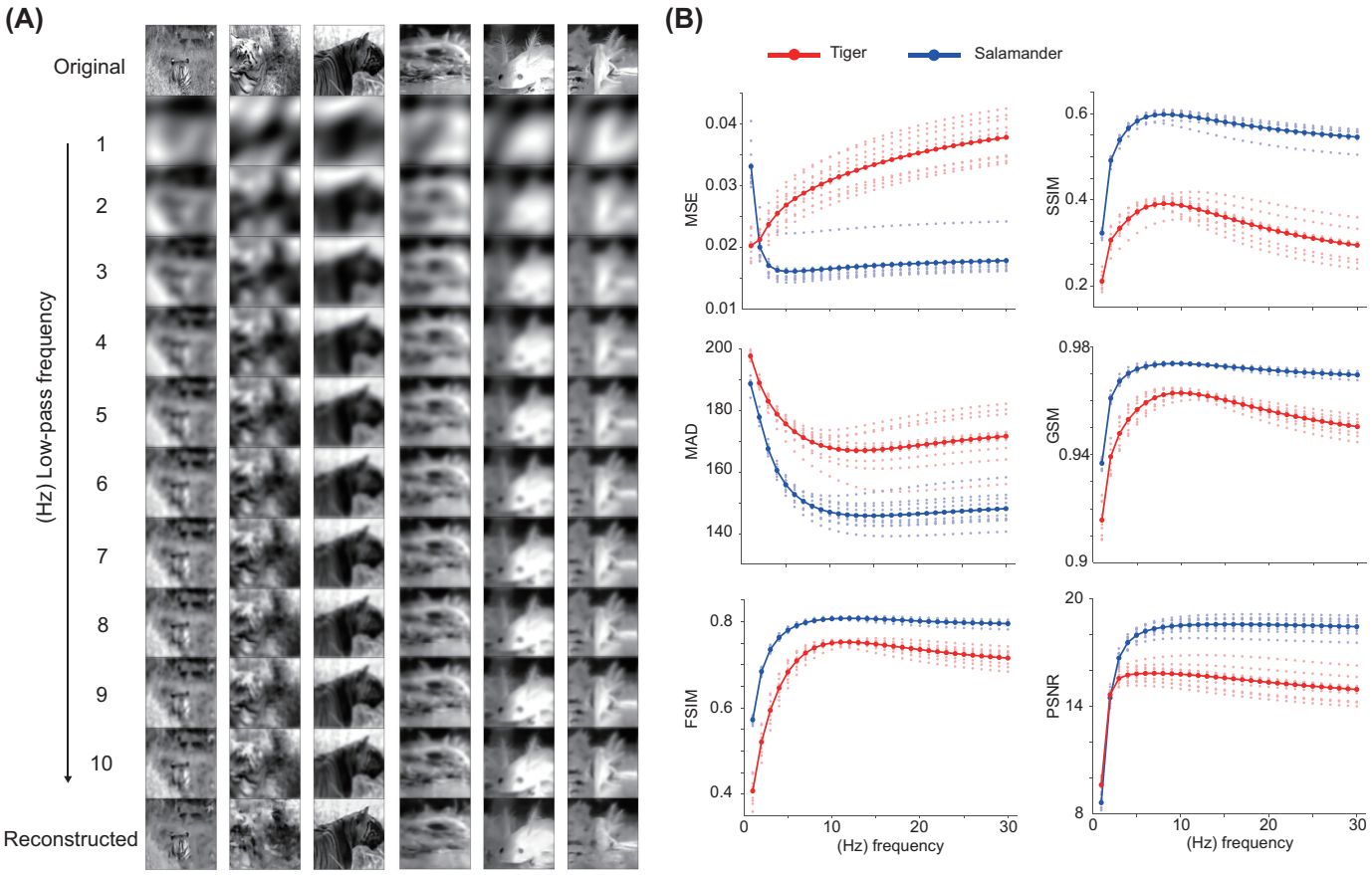

Fig. 6: Image reconstruction metrics change through low-pass frequency reference frames in both datasets. (A) Original frames, reconstructed frames, and low-pass frequency (1-10Hz) reference frames in the tiger dataset. (B) Change of reconstruction metrics over low-pass frequency. 10 light-color points at each x-axis represent metric values of 10 decoding models with different random initialization. The dark-color line indicates the mean metric values of the 10 different models described above.

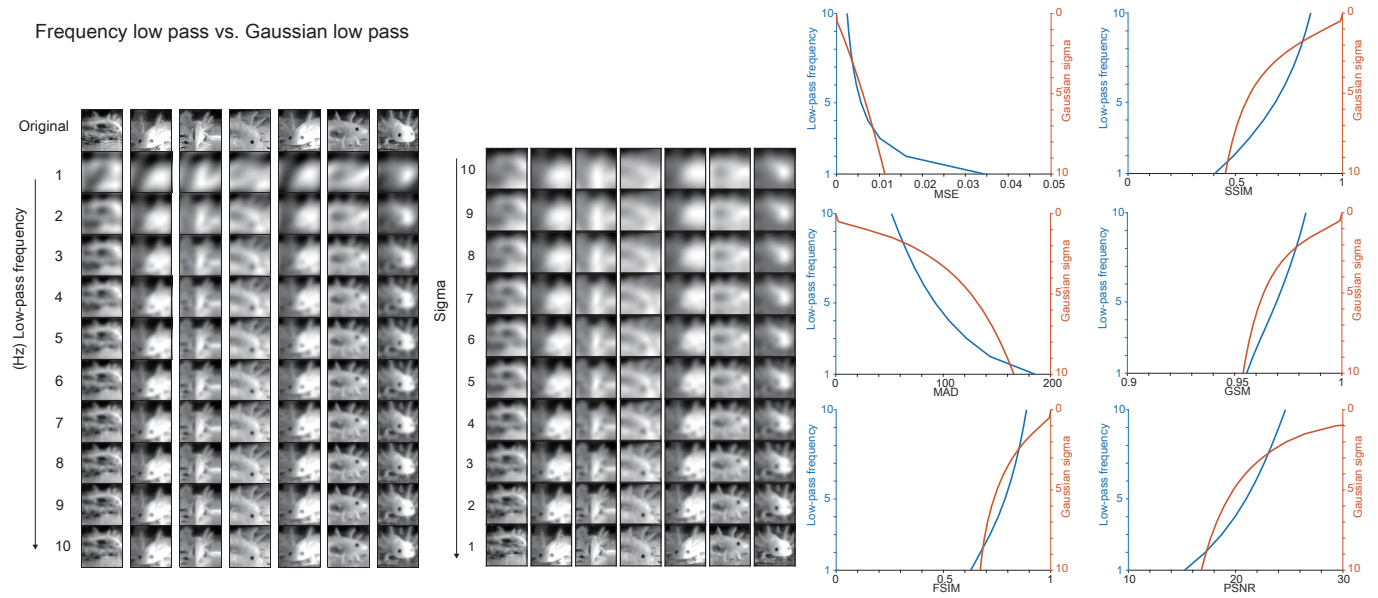

Fig. 7: Reconstruction of low-pass filtered and Gaussian smoothed images. (A) (Left) Original frames, reconstructed frames, and low-pass frequency (1-10 Hz) and (right) Gaussian low-pass sigma (1-10) frames in the salamander dataset. (B) Change of reconstruction metrics for low-pass filter and Gaussian filter.

decoding by adjusting and introducing iteration structure to a neural encoder model and making an analog hardware implementation that is energy efficient but also robust [62]. An adaptive spike threshold was proposed [63] to ensure robust information transmission across cortical states. On the other hand, in the field of artificial neural networks, a similar problem is called adversarial attacks [64], [65], [66], [67], [68], [69]. Concretely, a surprising failure to recognize objects would emerge in images corrupted with different noise patterns that humans have no trouble with if the artificial neural networks were not particularly designed [70], [71], [72]. Recent work creatively combined a biologically constrained Gabor filter bank with an artificial neural network back-end [73]. This so-called VOneBlock is substantially more robust than the base convolutional neural network while maintaining high performance on the classification tasks. Although researchers have tried to overcome noise corruption from the neural encoding field, no prior work has analytically paid attention to noise corruption in neural decoding. Our work filled this gap with a detailed examination of the effect of different types of noise.

Taken together, our results provide a detailed guideline for using a neural decoder to read out external stimuli from retinal spikes, which can serve to develop advanced methods for neuroprosthesis and other types of brain-machine interface devices. Together with recent advances in retinal neuroprosthesis hardware [74], [1], [75], our results suggest that one can employ a decoder to quantify the quality of reconstructed dynamical visual scenes from retinal spikes generated by neuroprosthesis, which could enhance the further design of neuroprosthesis hardware.

REFERENCES

[1] N. P. Shah and E. J. Chichilnisky, "Computational challenges and opportunities for a bi-directional artificial retina," *Journal of Neural Engineering*, vol. 17, no. 5, p. 055002, 2020.

[2] T. Gollisch and M. Meister, "Eye smarter than scientists believed: Neural computations in circuits of the retina," *Neuron*, vol. 65, no. 2, pp. 150–164, 2010.

[3] Z. Yu, J. K. Liu, S. Jia, Y. Zhang, Y. Zheng, Y. Tian, and T. Huang, "Toward the next generation of retinal neuroprosthesis: Visual computation with spikes," *Engineering*, vol. 6, no. 4, pp. 449–461, 2020.

[4] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, "Bayesian reconstruction of natural images from human brain activity," *Neuron*, vol. 63, no. 6, pp. 902–915, 2009.

[5] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology*, vol. 21, no. 19, pp. 1641–1646, 2011.

[6] K. Qiao, C. Zhang, L. Wang, J. Chen, L. Zeng, L. Tong, and B. Yan, "Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture," *Frontiers in Neuroinformatics*, vol. 12, p. 62, 2018.

[7] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, and S. Dehaene, "Inverse retinotopy: Inferring the visual content of images from brain activation patterns," *Neuroimage*, vol. 33, no. 4, pp. 1104–1116, 2006.

[8] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, and Z. Liu, "Neural encoding and decoding with deep learning for dynamic natural vision," *Cerebral Cortex*, vol. 28, no. 12, pp. 4136–4160, 2018.

[9] C. Du, C. Du, and H. He, "Sharing deep generative representation for perceived image reconstruction from human brain activity," in *International Joint Conference on Neural Network*. IEEE, 2017, pp. 1049–1056.

[10] C. Du, C. Du, L. Huang, and H. He, "Reconstructing perceived images from human brain activities with bayesian deep multiview learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 8, pp. 2310–2323, 2018.

[11] K. Fu, C. Du, S. Wang, and H. He, "Multi-view multi-label fine-grained emotion decoding from human brain activity," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[12] V. Botella-Soler, S. Deny, G. Martius, O. Marre, and G. Tkačik, "Nonlinear decoding of a complex movie from the mammalian retina," *PLoS Computational Biology*, vol. 14, no. 5, p. e1006057, 2018.

[13] T. Gollisch and M. Meister, "Rapid neural coding in the retina with relative spike latencies," *Science*, vol. 319, no. 5866, pp. 1108–1111, 2008.

[14] N. Parthasarathy, E. Batty, W. Falcon, T. Rutten, M. Rajpal, E. J. Chichilnisky, and L. Paninski, "Neural networks for efficient bayesian decoding of natural images from retinal neurons," in *Advances in Neural Information Processing Systems*, 2017, pp. 6434–6445.

[15] O. Marre, V. Botella-Soler, K. D. Simmons, T. Mora, G. Tkačik, and M. J. Berry II, "High accuracy decoding of dynamical motion from a large retinal population," *PLoS Computational Biology*, vol. 11, no. 7, p. e1004304, 2015.

[16] Z. Yu, S. Guo, F. Deng, Q. Yan, K. Huang, J. K. Liu, and F. Chen, "Emergent inference of hidden markov models in spiking neural networks through winner-take-all," *IEEE Transactions on Cybernetics*, vol. 50, no. 3, pp. 1347–1354, 2018.

[17] Y. Zhang, S. Jia, Y. Zheng, Z. Yu, Y. Tian, S. Ma, T. Huang, and J. K. Liu, "Reconstruction of natural visual scenes from neural spikes with deep neural networks," *Neural Networks*, vol. 125, pp. 19–30, 2020.

[18] N. Brackbill, C. Rhoades, A. Kling, N. P. Shah, A. Sher, A. M. Litke, and E. Chichilnisky, "Reconstruction of natural images from responses of primate retinal ganglion cells," *Elife*, vol. 9, p. e58516, 2020.

[19] Q. Zhou, C. Du, D. Li, H. Wang, J. K. Liu, and H. He, "Simultaneous neural spike encoding and decoding based on cross-modal dual deep generative model," in *International Joint Conference on Neural Networks*. IEEE, 2020, pp. 1–8.

[20] Q. Xu, J. Shen, X. Ran, H. Tang, G. Pan, and J. K. Liu, "Robust transcoding sensory information with neural spikes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 1935–1946, 2022.

[21] T. Yoshida and K. Ohki, "Natural images are reliably represented by sparse and variable populations of neurons in visual cortex," *Nature Communications*, vol. 11, no. 1, p. 872, 2020.

[22] S. Garasto, W. Nicola, A. A. Bharath, and S. R. Schultz, "Neural sampling strategies for visual stimulus reconstruction from Two-photon imaging of mouse primary visual cortex," in *International IEEE/EMBS Conference on Neural Engineering*, 2019, pp. 566–570.

[23] S. Garasto, A. A. Bharath, and S. R. Schultz, "Visual reconstruction from 2-photon calcium imaging suggests linear readout properties of neurons in mouse primary visual cortex," *bioRxiv*, p. 300392, 2018.

[24] J. B. Demb and J. H. Singer, "Functional circuitry of the retina," *Annual Review of Vision Science*, vol. 1, pp. 263–289, 2015.

[25] W. N. Grimes, A. Songco-Aguas, and F. Rieke, "Parallel processing of rod and cone signals: Retinal function and human perception," *Annual Review of Vision Science*, vol. 4, pp. 123–141, 2018.

[26] P. D. Jadzinsky and S. A. Baccus, "Transformation of visual signals by inhibitory interneurons in retinal circuits," *Annual Review of Neuroscience*, vol. 36, no. 1, pp. 403–428, 2013.

[27] J. O'Brien and S. A. Bloomfield, "Plasticity of retinal gap junctions: Roles in synaptic physiology and disease," *Annual Review of Vision Science*, vol. 4, pp. 79–100, 2018.

[28] M. Rivlin-Etzion, W. N. Grimes, and F. Rieke, "Flexible neural hardware supports dynamic computations in retina," *Trends in Neurosciences*, vol. 41, no. 4, pp. 224–237, 2018.

[29] J. K. Liu and T. Gollisch, "Spike-triggered covariance analysis reveals phenomenological diversity of contrast adaptation in the retina," *PLoS Computational Biology*, vol. 11, no. 7, p. e1004425, 2015.

[30] 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*, vol. 10, 2017.

[31] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, no. 7207, pp. 995–999, 2008.

[32] J. K. Liu, H. M. Schreyer, A. Onken, F. Rozenblit, M. H. Khani, V. Krishnamoorthy, S. Panzeri, and T. Gollisch, "Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization," *Nature Communications*, vol. 8, no. 1, p. 149, 2017.

[33] B. Yan and S. Nirenberg, "An embedded real-time processing platform for optogenetic neuroprosthetic applications," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 1, pp. 233–243, 2018.

[34] Q. Yan, Y. Zheng, S. Jia, Y. Zhang, Z. Yu, F. Chen, Y. Tian, T. Huang, and J. K. Liu, "Revealing fine structures of the retinal receptive field by deep-learning networks," *IEEE Transactions on Cybernetics*, vol. 39, pp. 39–50, 2022.

[35] J. K. Liu and T. Gollisch, "Simple model for encoding natural images by retinal ganglion cells with nonlinear spatial integration," *PLoS Computational Biology*, vol. 18, no. 3, p. e1009925, 2022.

[36] D. K. Warland, P. Reinagel, and M. Meister, "Decoding visual information from a population of retinal ganglion cells," *Journal of Neurophysiology*, vol. 78, no. 5, pp. 2336–2350, 1997.

[37] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

[39] B. C. McCann, M. M. Hayhoe, and W. S. Geisler, "Decoding natural signals from the peripheral retina," *Journal of Vision*, vol. 11, no. 10, pp. 19–19, 2011.

[40] S. B. Ryu, J. H. Ye, Y. S. Goo, C. H. Kim, and K. H. Kim, "Decoding of temporal visual information from electrically evoked retinal ganglion cell activities in photoreceptor-degenerated retinas," *Investigative Opthalmology & Visual Science*, vol. 52, no. 9, p. 6271, 2011.

[41] Y. J. Kim, N. Brackbill, E. Batty, J. Lee, C. Mitelut, W. Tong, E. J. Chichilnisky, and L. Paninski, "Nonlinear decoding of natural images from large-scale primate retinal ganglion recordings," *Neural Computation*, vol. 33, no. 7, pp. 1719–1750, 2021.

[42] Y. Zhang, T. Bu, J. Zhang, S. Tang, Z. Yu, J. K. Liu, and T. Huang, "Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex," *Neural Computation*, vol. 34, no. 6, pp. 1369–1397, 2022.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980 [cs]*, 2017.

[44] A. Onken, J. K. Liu, P. P. C. R. Karunasekara, I. Delis, T. Gollisch, and S. Panzeri, "Using matrix and tensor factorizations for the single-trial analysis of population spike trains," *PLoS Computational Biology*, vol. 12, no. 11, p. e1005189, 2016.

[45] 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*, vol. 13, no. 4, pp. 600–612, 2004.

[46] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.

[47] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[48] M. C. Morrone, J. Ross, D. C. Burr, and R. Owens, "Mach bands are phase dependent," *Nature*, vol. 324, no. 6094, pp. 250–253, 1986.

[49] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2012.

[50] R. Dosselmann and X. D. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image and Video Processing*, vol. 5, no. 1, pp. 81–91, 2011.

[51] M. Pedersen and J. Y. Hardeberg, "Full-reference image quality metrics: Classification and evaluation," *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, no. 1, pp. 1–80, 2012.

[52] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *International Conference on Pattern Recognition*. IEEE, 2010, pp. 2366–2369.

[53] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2002, pp. IV–3313.

[54] Y. Zheng, S. Jia, Z. Yu, J. K. Liu, and T. Huang, "Unraveling neural coding of dynamic natural visual scenes via convolutional recurrent neural networks," *Patterns*, vol. 2, no. 10, p. 100350, 2021.

[55] V. Lopes-dos-Santos, S. Panzeri, C. Kayser, M. E. Diamond, and R. Quian Quiroga, "Extracting information in spike time patterns with wavelets and information theory," *Journal of Neurophysiology*, vol. 113, no. 3, pp. 1015–1033, 2015.

[56] V. Lopes-dos-Santos, H. G. Rey, J. Navajas, and R. Quian Quiroga, "Extracting information from the shape and spatial distribution of evoked potentials," *Journal of Neuroscience Methods*, vol. 296, pp. 12–22, 2018.

[57] M. Okun, P. Yger, S. L. Marguet, F. Gerard-Mercier, A. Benucci, S. Katzner, L. Busse, M. Carandini, and K. D. Harris, "Population rate dynamics and multineuron firing patterns in sensory cortex," *Journal of Neuroscience*, vol. 32, no. 48, pp. 17 108–17 119, 2012.

[58] M. Okun, N. A. Steinmetz, L. Cossell, M. F. Iacaruso, H. Ko, P. Barthó, T. Moore, S. B. Hofer, T. D. Mrsic-Flogel, M. Carandini, and K. D. Harris, "Diverse coupling of neurons to populations in sensory cortex," *Nature*, vol. 521, no. 7553, pp. 511–515, 2015.

[59] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[60] T. Bu, W. Fang, J. Ding, P. DAI, Z. Yu, and T. Huang, "Optimal ann-snn conversion for high-accuracy and ultra-low-latency spiking neural networks," in *International Conference on Learning Representations*, 2022.

[61] J. Zylberberg, A. Pouget, P. E. Latham, and E. Shea-Brown, "Robust information propagation through noisy neural circuits," *PLoS Computational Biology*, vol. 13, no. 4, p. e1005497, 2017.

[62] C. Zhao, J. Li, and Y. Yi, "Making neural encoding robust and energy efficient: An advanced analog temporal encoder for brain-inspired computing systems," in *IEEE/ACM International Conference on Computer-Aided Design*, 2016, pp. 1–6.

[63] C. Huang, A. Resnik, T. Celikel, and B. Englitz, "Adaptive Spike Threshold Enables Robust and Temporally Precise Neuronal Encoding," *PLoS Computational Biology*, vol. 12, no. 6, p. e1004984, 2016.

[64] W. Brendel, J. Rauber, M. Kümmerer, I. Ustyuzhaninov, and M. Bethge, "Accurate, reliable and fast robustness evaluation," *arXiv:1907.01003 [cs, stat]*, 2019.

[65] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *arXiv:1608.04644 [cs]*, 2017.

[66] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "EAD: Elastic-net attacks to deep neural networks via adversarial examples," *arXiv:1709.04114 [cs, stat]*, 2018.

[67] J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," *arXiv:1811.09600 [cs]*, 2019.

[68] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv:1312.6199 [cs]*, 2014.

[69] T. Bu, J. Ding, Z. Hao, and Z. Yu, "Rate gradient approximation attack threats deep spiking neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 7896–7906.

[70] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," *arXiv:1705.02498 [cs]*, 2017.

[71] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *arXiv:1808.08750 [cs, q-bio, stat]*, 2020.

[72] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv:1903.12261 [cs, stat]*, 2019.

[73] J. Dapello, T. Marques, M. Schrimpf, F. Geiger, D. D. Cox, and J. J. DiCarlo, "Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations," *bioRxiv*, 2020.

[74] L. Yue, J. D. Weiland, B. Roska, and M. S. Humayun, "Retinal stimulation strategies to restore vision: Fundamentals and systems," *Progress in Retinal and Eye Research*, vol. 53, pp. 21–47, 2016.

[75] J. Tang, N. Qin, Y. Chong, Y. Diao, Z. Wang, T. Xue, M. Jiang, J. Zhang, G. Zheng *et al.*, "Nanowire arrays restore vision in blind mice," *Nature Communications*, vol. 9, no. 1, p. 786, 2018.

Supplemental Materials:

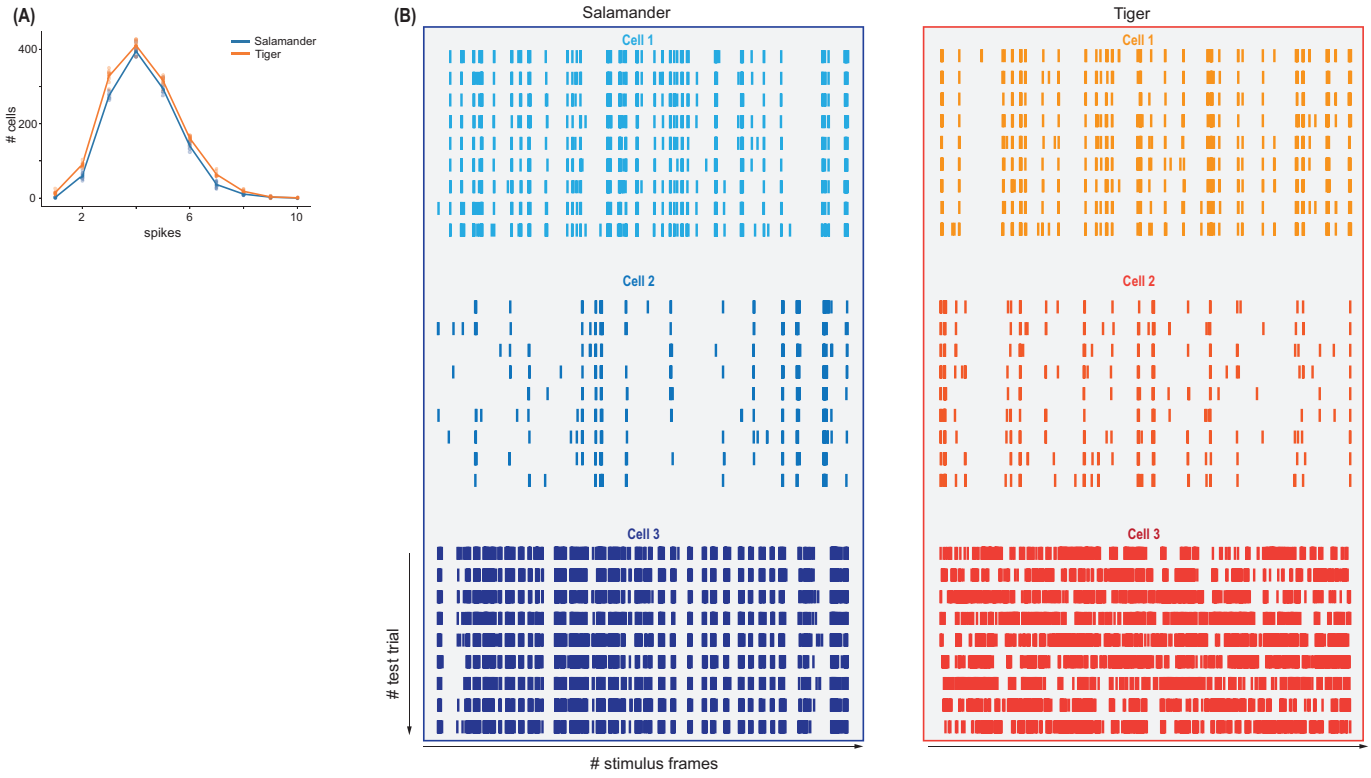

Fig. S1: Spike trains in salamander and tiger datasets. (A) Distribution of cells with maximal firing spikes to different stimulus frames in salamander and tiger datasets. The x-axis represents the number of firing spikes. And the corresponding colored points represent the cell numbers in 9 trials (blue for salamander and yellow for tiger). The lines are joined with the average values of the cell numbers in 9 trials. (B) We illustrated each 9 spike rasters of 3 respective cells in salamander and tiger datasets. We take the Cell 3 on the left panel as an example. The y-axis represents different stimulus frames (which are 1800 in salamander movie and 1600 in tiger movie), and the nine rows indicate firing rasters of 9 trials. Each bar with dark blue means this cell has fired at the very stimulus frame.

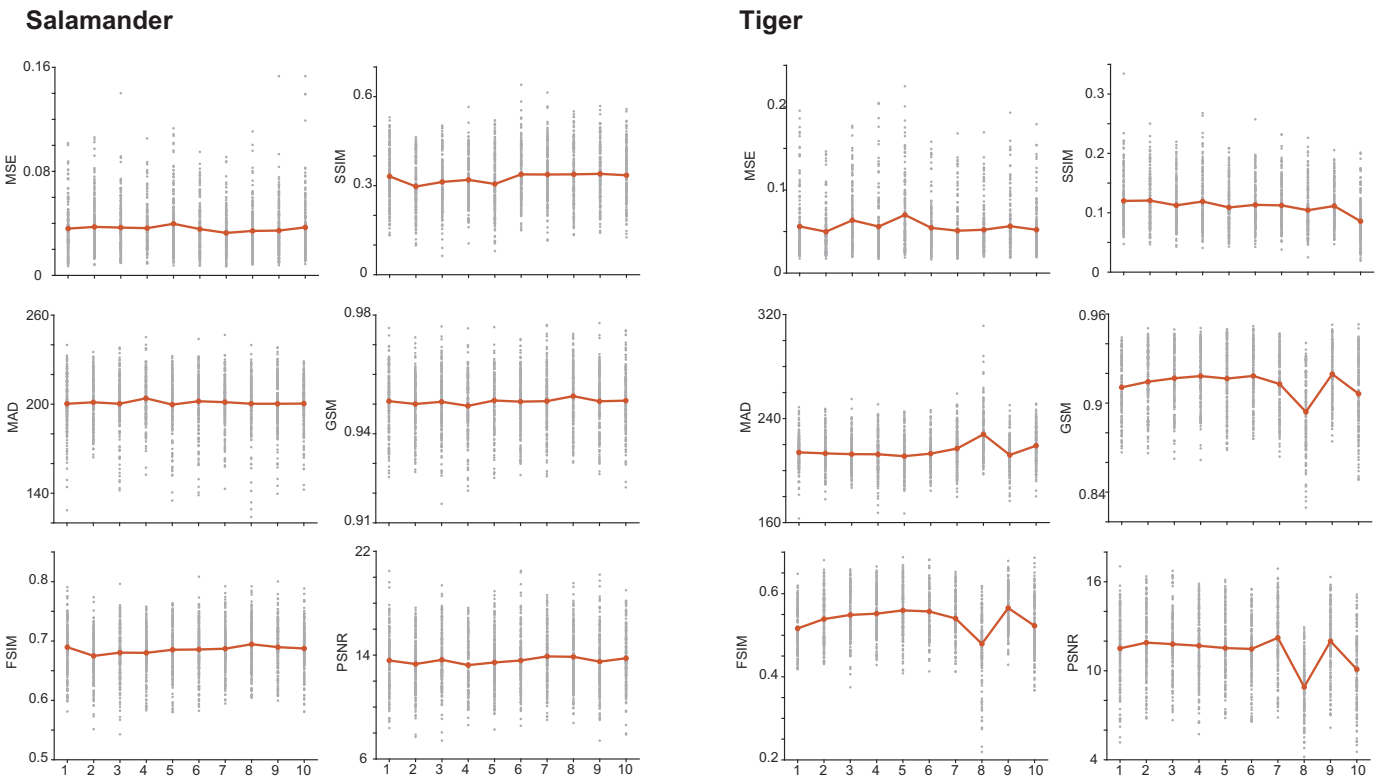

Fig. S2: Metrics of 10 decoding models of 100 RGCs. The x-axis represents 10 different decoding models trained with 100 randomly chosen RGCs from salamander or tiger dataset. The gray points are the metric value of each frame from half of the test set size, specifically, 90 for the salamander and 80 for the tiger. The red line indicates the average value of all frames from the test set. red

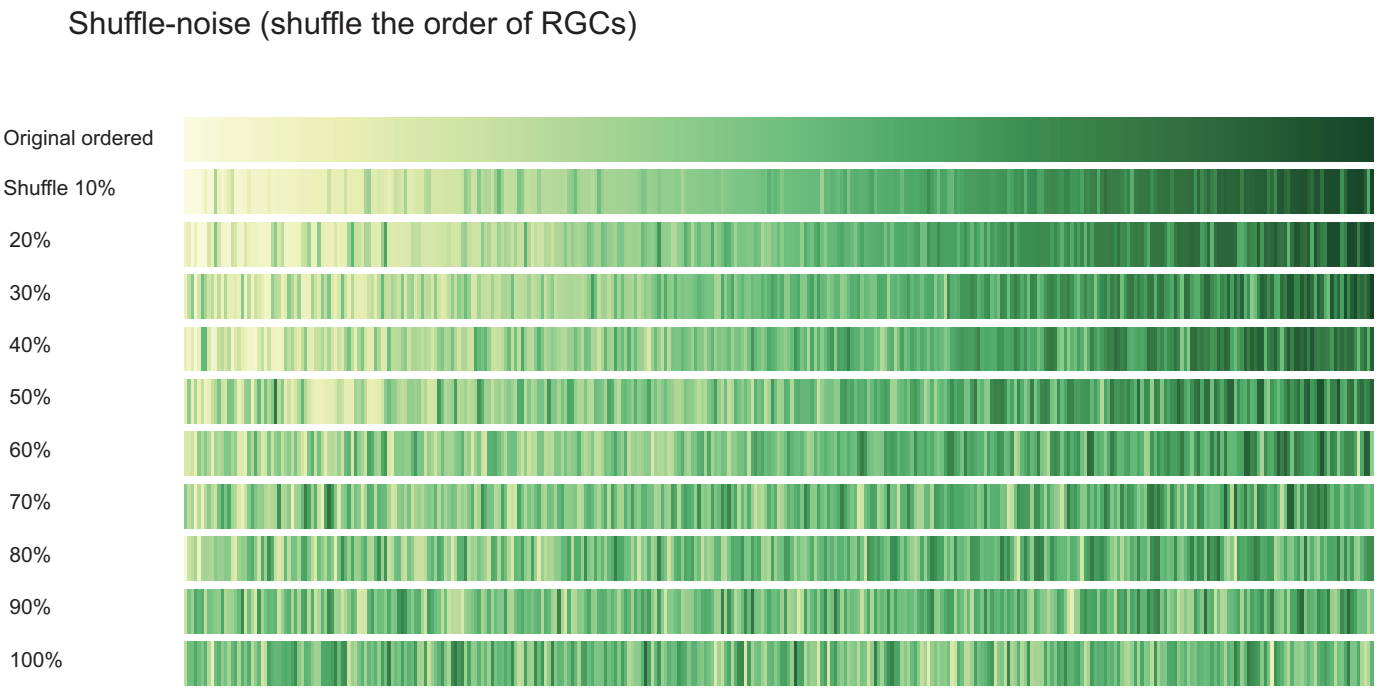

Fig. S3: Illustration of shuffle noise. We explain the principle of shuffle noise in this work. First, we assume the first row is the original order of RGCs, which are fed into the decoding model in the context of no-noise added. The whole color bar at the first row is the original color bar named 'YlGn', which can be divided into 1218 parts average at the x-axis, indicating 1218 cells. The original color bar indicates the color evolution from the left to the right uniformly, while more shuffle noise, more chaotic for the color evolution. Adding shuffle noise is to shuffle the input order of a certain percent of all RGCs, which are illustrated as following rows. red

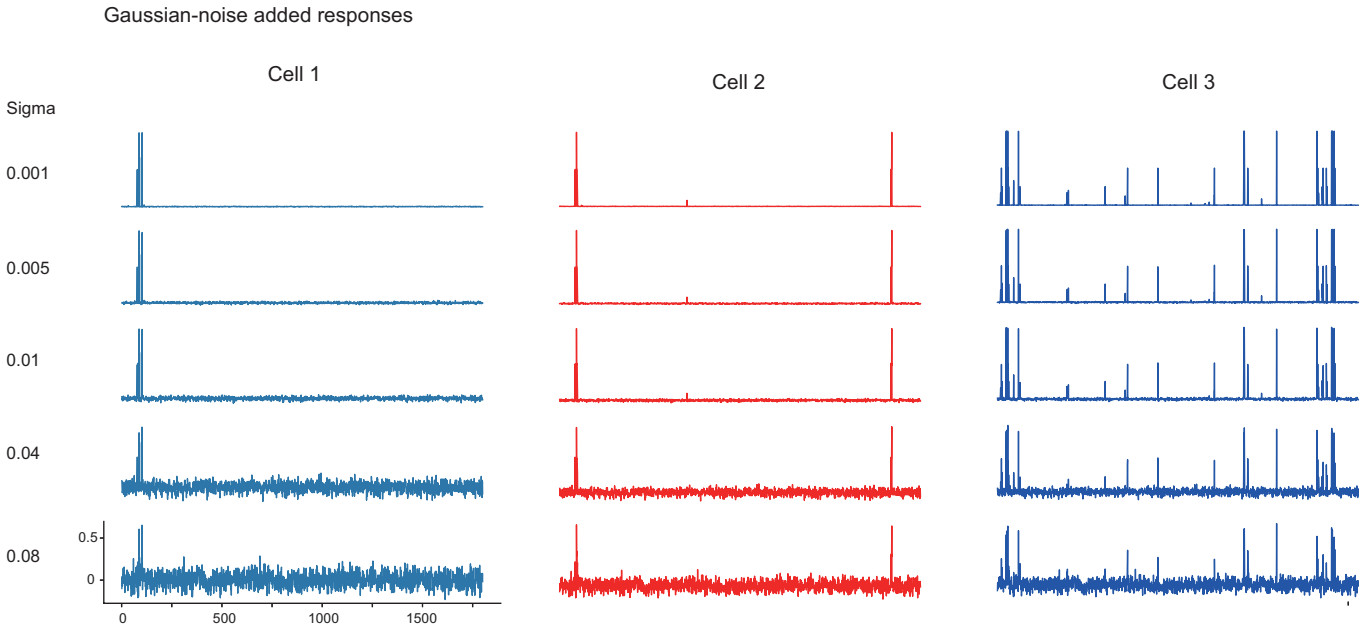

Fig. S4: Illustration of Gaussian noise. We add zero-mean Gaussian noise with different sigmas to the normalized spike trains of RGC response to 1800 stimuli. Here, we take 3 cells for instance, which represent different response cases of concentration firing, sparse firing to different stimuli, and strong firing to multi stimuli. Different cells are indicated by different colors. From up to down, each row represents the noise-added normalized spike train (added larger-sigma Gaussian noise). red

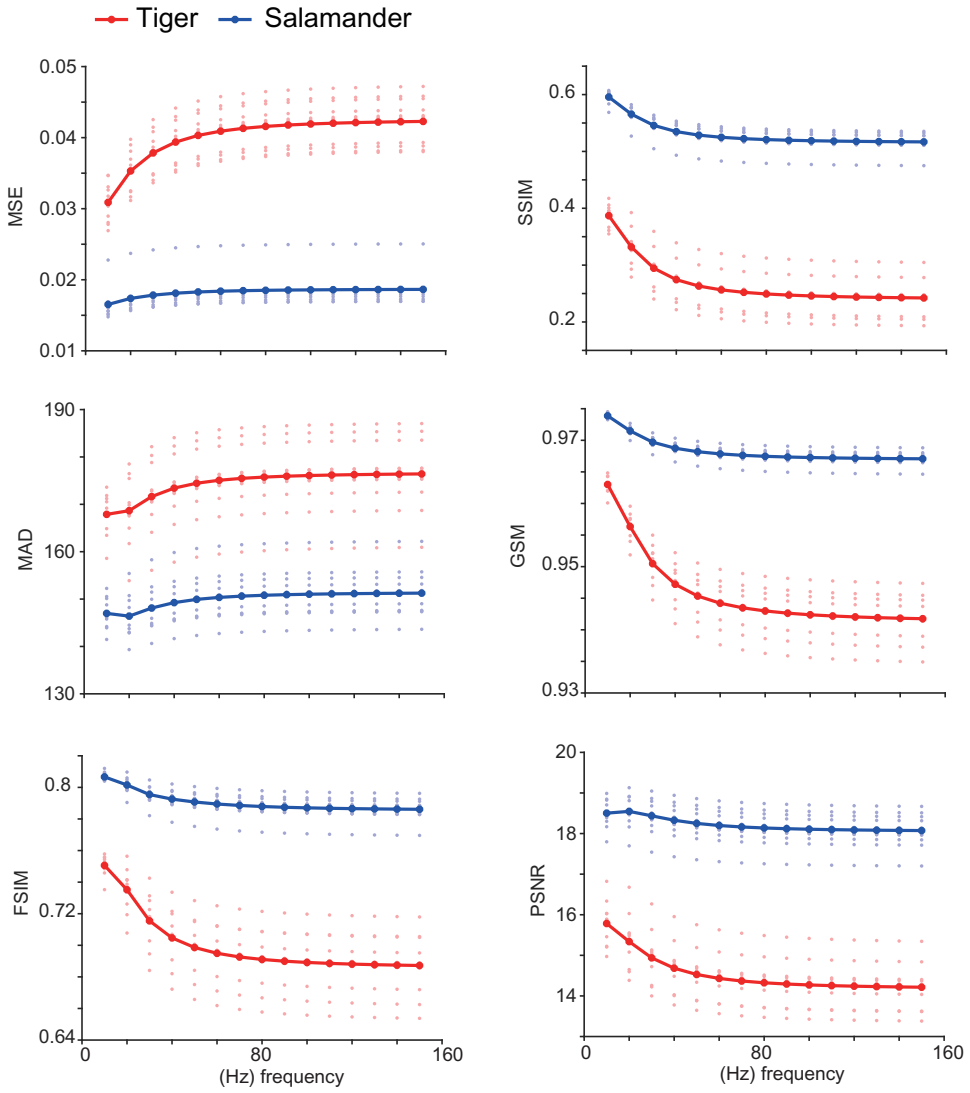

Fig. S5: **Image reconstruction metrics change through low-pass frequency (1-150Hz) reference frames in both datasets.** 10 light-color points at each x-axis represent metric values of 10 decoding models with different random initialization. The dark-color line indicates the mean metric values of 10 different models described above for each subplot. red green