Reconstruction of Natural Visual Scenes from Neural Spikes with Deep Neural Networks

Yichen Zhang[1], Shanshan Jia[1], Yajing Zheng[1], Zhaofei Yu[1], Yonghong Tian[1], Siwei Ma[1], Tiejun Huang[1], Jian K. Liu[2]

Abstract

Neural coding is one of the central questions in systems neuroscience for understanding of how the brain processes stimulus from the environment, moreover, it is also a cornerstone for designing algorithms of brain-machine interface, where decoding incoming stimulus is highly demanded for better performance of physical devices. Traditionally researchers have focused on functional magnetic resonance imaging (fMRI) data as the neural signals of interest for decoding visual scenes. However, our visual perception operates in a fast time scale of millisecond in terms of an event termed neural spike. There are few studies of decoding by using spikes. Here we fulfill this aim by developing a novel decoding framework based on deep neural networks, named spike-image decoder (SID), for reconstructing natural visual scenes, including static images and dynamic videos, from experimentally recorded spikes of a population of retinal ganglion cells. The SID is an end-to-end decoder with one end as neural spikes and the other end as images, which can be trained directly such that visual scenes are reconstructed from spikes in a highly accurate fashion. Our SID also outperforms on the reconstruction of visual stimulus compared to existing fMRI decoding models. In addition, with the aid of a spike encoder, we show that SID can be generalized to arbitrary visual scenes by using the image datasets of MNIST, CIFAR10, and CIFAR100. Furthermore, with a pre-trained SID, one can decode any dynamic videos to achieve real-time encoding and decoding of visual scenes by spikes. Altogether, our results shed new lights on neuromorphic computing for artificial visual systems, such as event-based visual cameras and visual neuroprostheses.

Keywords: Vision; Natural scenes; Neural decoding; Neural spikes; Deep learning; Artificial retina

1. Introduction

In everyday life, various types of sensory information are processed by our brain through different sensory modalities, which are then processed to generate a series of behavior reactions. Neuronal networks in the brain, then, play an essential role of efficient and powerful computation where a sequence of input-output mappings is carried out by single and networked neurons. At the level of the system, single neurons receive and respond to input stimuli by changing their membrane potential to generate a sequence of fast events, termed neural spikes. Thus, spikes have been suggested as a fundamental element to represent input-output neural computation of visual scenes [1, 2]. Such observations lead to a central question for systems neuroscience on the sensory system: how neurons represent the input-output relationship between stimuli and their spikes [3, 4, 5]. This question, formulated as neural coding, consists of two essential parts, encoding and decoding. For the visual system, it is to understand how visual scenes are represented by neural spiking activity, and how to decode them to reconstruct the input visual scenes.

The retina serves as a useful system to study these questions. Visual scenes are projected into the eyes, where the retinal neuronal network processes the input by a few types of neurons, but the only output neurons are the retinal

*Corresponding authors: Zhaofei Yu, Jian K. Liu

[1]National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing, and Peng Cheng Laboratory, Shenzhen, China

[2]Centre for Systems Neuroscience, Department of Neuroscience, Psychology and Behaviour, University of Leicester, Leicester, UK

ganglion cells (RGCs). Therefore, all information of visual scenes is encoded by RGCs that produce a sequence of action potentials or spikes, which are transmitted via the optic nerve to the downstream brain regions. Essentially, visual scenes are represented by spikes of a population of RGCs. One expects that encoding and decoding of visual scenes from RGC spikes can be formulated into a closed form where a computational model can be developed to set up a mapping between visual scenes and RGC spikes. Indeed, much effort has been given to study the encoding part of RGCs. Various neuroscience mechanisms have been identified in understanding the retinal computation of visual scenes by its neurons and neural circuitry [6, 7, 8, 9, 10, 11]. In addition, for understanding the encoding principles of the retina, a number of models are developed based on different properties of neurons and neural circuits in the retina [12, 13, 14, 15, 16].

Decoding of RGC spikes to obtain the essential information of visual scenes is processed by the downstream neurons from the lateral geniculate nucleus to the visual cortex, which involves feedbacks from the higher part of the cortex for cognition. However, as the retina does not receive these feedbacks, the RGC spikes can be thought as a minimal computational device to represent visual information as a whole. Therefore, it is suitable to develop a decoding model that can reconstruct visual scenes from the RGC spikes directly [2, 17]. This is in particular important for neuroengineering, such as neuroprosthesis, where efficient algorithms of decoding neuronal signals for controlling physical devices are highly demanded [18]. In general, an ideal neuronal decoder should be able to read out and reconstruct stimulus from neural responses. The existing methods for visual scenes reconstruction are mainly divided into two types. The first one is to decode the stimulus category: choosing a stimulus according to neural responses, then classifying neural signals to find the corresponding stimulus in a candidate set [19]. The second type is to reconstruct the original stimulus directly by using neural responses, i.e., obtaining every pixel of visual scenes from the neural signal. It is obvious that the second category is more challenging [20].

Reconstruction of visual scenes has been studied over many years. The neural signals of interest can be functional magnetic resonance imaging (fMRI) activity [19, 20, 21, 22, 23], neural spikes in the retina [2, 17, 24, 25] and lateral geniculate nucleus [26], neural calcium imaging signals in the visual cortex [27]. However, the decoding performance of current methods is rather low for natural scenes, in particular for dynamical videos, which can be seen from some typical examples of the videos reconstructed from fMRI signals [19, 20]. For the retina, one would expect to decode visual scenes by using the spiking responses of a population of RGCs in a complete way as these spikes are the only output of the eyes. Decoding of visual scenes, at least for static natural images, is possible when sampling the whole image with a number of RGCs [2]. Yet, it remains unclear how to deal with the dynamic natural scenes, where the temporal complexity of stimulus information has a strong coupling with the temporal adaption of neurons [28].

In this study, we propose such a general-purpose decoding framework, termed spike-image decoder (SID), that performs an end-to-end decoding process from neural spikes to visual scenes, based on deep learning neural networks. The proposed SID can achieve state-of-the-art performance, compared to previous studies, for reconstructing natural visual scenes, including both static images and dynamic videos, from spikes of a population of RGCs recorded simultaneously in the isolated animal retina. The workflow of the SID is illustrated in Figure 1. When a large population of RGCs is recorded simultaneously, and their spikes are extracted, a spike-image converter based on a neural network is used to map the spikes of every RGC to intermediate images at the pixel level. After that, an autoencoder-type deep learning neural network is applied to spike-based intermediate images to match original stimulus images for every pixel.

Essentially, the SID has two stages with one as spike-image converter and another one as image-image autoencoder. Most of the previous studies focused on the first stage, such that the spikes are directly mapped to obtain the image pixels, where, traditionally, a decoder can be optimized by some statistical models in a linear or nonlinear fashion [2, 17, 20, 21, 22, 24, 26]. A recent study used a separate autoencoder network as the second stage to enhance the quality of images [25], which essentially employed a similar network model dedicated for image denoising and enhancement [29]. Here, we systematically tested different training approaches of neural networks from spikes to images, and found the SID can achieve similar performance for static images across three types of loss functions, which is consistent with those previous studies, however for dynamic videos, the direct training from spikes to images shows an outstanding performance.

We tested the performance of SID on experimentally recorded RGCs spikes, and found it can obtain state-of-the-art performance for reconstruction of natural visual scenes of static image and dynamic video. Our SID can also be used to decode the fMRI signal, and has a better performance on the reconstruction of visual stimulus compared to the existing state-of-the-art fMRI decoding models. Furthermore, we found that the computational ability of SID can be

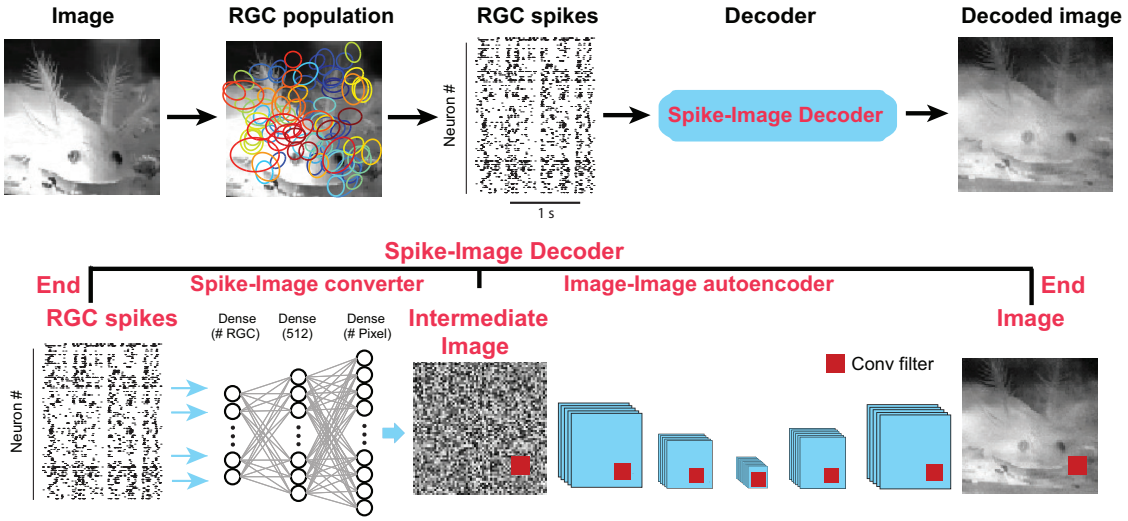


Figure 1: Illustration of the spike-image decoder model. (Top) Workflow of decoding visual scenes. Here a salamander swimming video was presented to a salamander retina to stimulate a firing of spike trains in a population of RGCs. The spikes are used to train the decoder for the given video to reconstruct the visual scene. Receptive fields of ganglion cells are mapped onto the image. Each coloured circle is an outline of receptive field. (Bottom) Spike-image decoder is an end-to-end decoder with two stages: spike-image converter used to map a population of spikes to a pixel-level intermediate image, and image-autoencoder for mapping every pixel to the target pixels in the desired images.

generalized to arbitrary visual scenes with the aid of an encoding model that can simulate spikes for any given images. By using the image datasets of MNIST, CIFAR10, and CIFAR100, the SID can reconstruct natural images with high precision. In addition, with a pre-trained SID model, one can reconstruct any given video scenes to realize real-time encoding and decoding of dynamic visual scenes as an artificial vision system.

2. Methods

2.1. Spike-image decoder

Traditionally, it has been shown that deep neural networks can extract low-level features in earlier layers, and more semantic features in later layers, which is similar to the information processing of the visual cortex in the brain [30, 31]. Here we used a similar deep neural network to reconstruct natural scenes from the retinal spikes. Our decoding model consists of two parts: 1) spike-to-image converter and 2) image-to-image autoencoder. The spike-to-image converter is able to map every spike to every pixel to get intermediate images, which plays a role of nonlinear re-sampling, and the image-to-image autoencoder is used to match all pixels to those of the target images of visual scenes.

2.1.1. Spike-to-image converter

The spike-to-image converter is a three-layer fully-connected neural network, similar to a multilayer perceptron (MLP). The first layer receives the spikes of all RGCs as input such that the number of neurons of the first layer is matched to the number of RGCs used. The second layer is a hidden layer, which consists of 512 neurons. ReLU is used as an activation function of the hidden layer. The output layer has the same number of neurons as the pixel number of the input image, in our case, 4096 neurons were used since all stimulus images are in 64*64 pixels. ReLU as activation function is used for the third layer, and we can get all pixel values of intermediate image from this output layer.

2.1.2. Image-to-image autoencoder

The image-to-image autoencoder is a typical deep autoencoder based on convolutional neural network (CNN), where the whole information processing procedure can be split into two phases. In the first phase, convolution and down-sampling are used to process and decrease the size of the input image. This phase contains four convolutional

layers, the kernel sizes of these four layers are (64,7,7), (128,5,5), (256,3,3), (256,3,3) and the stride sizes are (2,2) for all these four layers. After this phase, the most important components of the input image are kept, and noise and redundant contents are filtered. In the second phase, convolution and up-sampling are used to process the image, which recovers the texture of the down-sampled image while increasing the size of the down-sampled image. The upsampling phase also consists of four convolutional layers, the kernel sizes are (256,3,3), (128,3,3), (64,5,5), (3,7,7) and the stride sizes are (1,1) for all these layers. Up-sampling layers with size (2,2) are added before these four convolutional layers. ReLU is used as an activation function in the whole Image-to-image autoencoder.

2.1.3. SID model architecture

Given an input image $\mathbf{I}$, it will trigger a response $\mathbf{s} = \{s_1, s_2 \dots s_n\}$ on the retinal ganglion cells, here we use rate coding such that s_i represents the spike count of each RGC within a time bin depending on the sampling rate of visual scenes. As in experimental recordings, a typical 30 Hz refreshing rate is used, so the time bin is about 33 ms. Then the triggered responses are first fed into the spike-to-image MLP converter, which outputs an intermediate image $\mathbf{O}_1 = f_1(\mathbf{s})$ from RGC responses. Then the image-to-image CNN autoencoder takes the intermediate image as input, and is able to map it to match the target image, so one can get a clear and refining reconstruction result $\mathbf{O}_2 = f_2(\mathbf{O}_1)$. These two parts are all implemented with deep neural networks, so end-to-end training can be used to train both networks.

Three different ways of defining loss function were explored in this study. In the first one, one can use a multi-task loss function to make sure both stages have the capability to decode natural scenes from spikes, so that we can get a high quality reconstruction result from our decoding model. Here, λ_1 and λ_2 are weights of these two parts mean square error loss, $\mathbf{L}_1 : Loss = \lambda_1||\mathbf{O}_1 - \mathbf{I}|| + \lambda_2||\mathbf{O}_2 - \mathbf{I}||$. In the second one, we optimize each stage of decoding separately such that $\mathbf{L}_2 : (Loss_1 = \lambda_1||\mathbf{O}_1 - \mathbf{I}||, \ \ Loss_2 = \lambda_2||\mathbf{O}_2 - \mathbf{I}||)$. The third one uses one loss for the whole process of two stages as following: $\mathbf{L}_3 : Loss = \lambda||\mathbf{O} - \mathbf{I}||$, where there is only one image output $\mathbf{O}$ as the final reconstruction to compare the target image such that one error signal stays in both stages. In this way, the intermediate image is like noise without visible structures as in Figure 1. Three loss functions were compared in details.

The SID was implemented with Keras deep learning library, and Tensorflow as backend was used. During training for both spike-to-image converter and image-to-image autoencoder, batch normalization using 256 as batch size for both converter MLP and autoencoder CNN was used to achieve better performance [32] and dropout [33] was used to deal with overfitting. Adam method was used to train the model. The code is available online at https://sites.google.com/site/jiankliu.

2.2. Experimental spike data

Our decoder was used to reconstruct natural visual scenes, including both static image and dynamic video, from spikes of a population of RGCs recorded simultaneously in isolated retinas of salamanders. Datasets of spikes recorded and stimulus images and movies used in this study can be found publicly [34]. Experimental details can be found in previous studies [34, 35], Briefly, neural data had been collected in isolated retinas obtained from axolotl salamanders. The retina was placed onto a 60- or 252-channel multielectrode array in a recording chamber. Visual stimuli were displayed on an OLED monitor that was projected onto the photoreceptor layer through a telecentric lens above the retina.

2.2.1. Image stimulus

The static images were taken from a dataset used previously [34, 35], which is part of the McGill Calibrated Colour Image Database [36], including 300 natural images with a size of 64*64 pixels. Briefly, each image covers 1920*1920 μm area on the retina, in which there are 80 RGCs recorded. Each individual image was presented for 200 ms and then followed by an 800 ms empty display in a pseudo-random sequence. For each image to each RGC, the number of spikes during 300 ms in one trial was collected. In addition, we averaged the spike counts among 13 trials for each image. Finally, we got the spike counts of a population of 80 RGCs for one image as the input data for reconstruction. Among all 300 images, 270 images were used to train the decoder model and the remaining 30 images were used as the test set.

2.2.2. Video stimulus

The video data consists of salamander swimming which includes 1800 frames as the same dataset in [34]. Each frame covered a total area of 2700*2700 μm on the retina with a spatial resolution of 360*360 pixels, which was down-sampled into 90*90 pixels, and cropped into 64*64 pixels around the central region of images, in which 90 RGCs responses were collected. A segment of the video was presented at a frame rate of 30 Hz. For the recorded responses of each RGC, one can bin them into 33 ms that is same with the presentation time of each frame. As a result, a spike train was collected with 1800 counts for each RGC in response to the video. Eventually for the population of RGCs, a 90*1800 spike count matrix, averaged over 31 trials of repeated presentation, were used to reconstruct the video. To train the decoder model, we randomly selected 1620 frames from the video as a training set and the remaining frames as a test set. In this way, the strong temporal correlation within the video can be washed out.

2.3. Simulated spike data

To test our SID model for any arbitrary visual scenes, we used a simple encoder based on the typical linear-nonlinear model [1, 35] to simulate neural responses of a population of RGCs. For comparison, the encoding model is based on the same experimental retinal data used for the analysis of the video. The linear filters are based on the receptive fields of a population of 90 RGCs obtained with white noise analysis [1, 34, 35] fitted with a 2D Gaussian for each cell. With the given natural scene stimulus (a static image or a frame of dynamic video), each cell implements a linear computation on pixels within its receptive field, and generates a spike count for this stimulus after a rectification nonlinearity. All these simulated spike counts (modeled responses of 90 RGCs) were then fed into our SID model to reconstruct natural scene stimulus.

With this retinal encoding model, one can simulate RGC responses under any given stimuli. In this study, simulations were done on three popular image datasets, MNIST, CIFAR10, and CIFAR100 [37, 38]. MNIST is a dataset of handwritten digits that consists of a training set of 60,000 examples and a test set of 10,000 examples. CIFAR10 and CIFAR100 are datasets used for image classification, consisting of 60,000 color images in 10 and 100 classes respectively. There are 50,000 training images and 10,000 test images in both of CIFAR10 and CIFAR100. All images were first resized to the size of 64*64 pixels, which is comparable to our retinal experimental data. Three channels of each color image were first encoded into spikes by the encoding model separately. The encoded (90,3) spikes are fed into spike-to-image converter to get a color intermediate image, here the weights are shared for all three channels. Then image-to-image autoencoder takes the intermediate image as input to get final reconstructed color image.

Similarly, for comparison, dynamic videos were also used for testing the SID. With the same encoding model, we used the same movie chips from the fMRI decoding experiments [20], and obtained a set of RGC spiking responses similar to the biological data. Then the SID model pre-trained with CAFAR10 dataset was used to decode these dynamic videos directly.

Finally, we implemented real-time coding of arbitrary visual scenes by embedding our models into the power-efficient computing device, here NVIDIA Jetson TX2, which is built around a 256-core NVIDIA Pascal GPU. In order to better take advantage of the GPU to accelerate the computation, we combined the image-spike encoding process into the decoding neural networks. In the real-time coding system, we can choose the input files such as image or video from the embedded device, or directly read the inputs, i.e. real-time videos, from a camera. However, as our SID was pre-trained with 64×64 pixels as input stimulus and target output, the resolution is not very high such that it misses the fine detailed structure of visual scenes. Therefore, based on the pre-trained SID, we cropped the inputs stimulus into several patches with 64×64 pixels, and then joined the decoding results after sending these data as a batch into the model simultaneously. Considering the GPU processing speed as well as the resolution, we cropped the input frames as 5×5 patches, such that the resolution of input images and decoding results can be increased to 320×320 in a real-time fashion. Higher resolution with real-time speed can be further achieved depending on the hardware used for this artificial vision system. The reconstruction results of the same videos used by the previous study [20] and our real-time video decoding are available online (https://sites.google.com/site/jiankliu).

2.4. Experimental fMRI data

For comparison, we consider experimental fMRI data recorded in human. The stimuli contain a set of 100 gray-scale handwritten digit images (equal number of 6's and 9's) extracted from the MNIST dataset, and the corresponding fMRI recordings were from V1-V3 areas of a single participant as detailed in previous studies [39, 40]. We evaluated

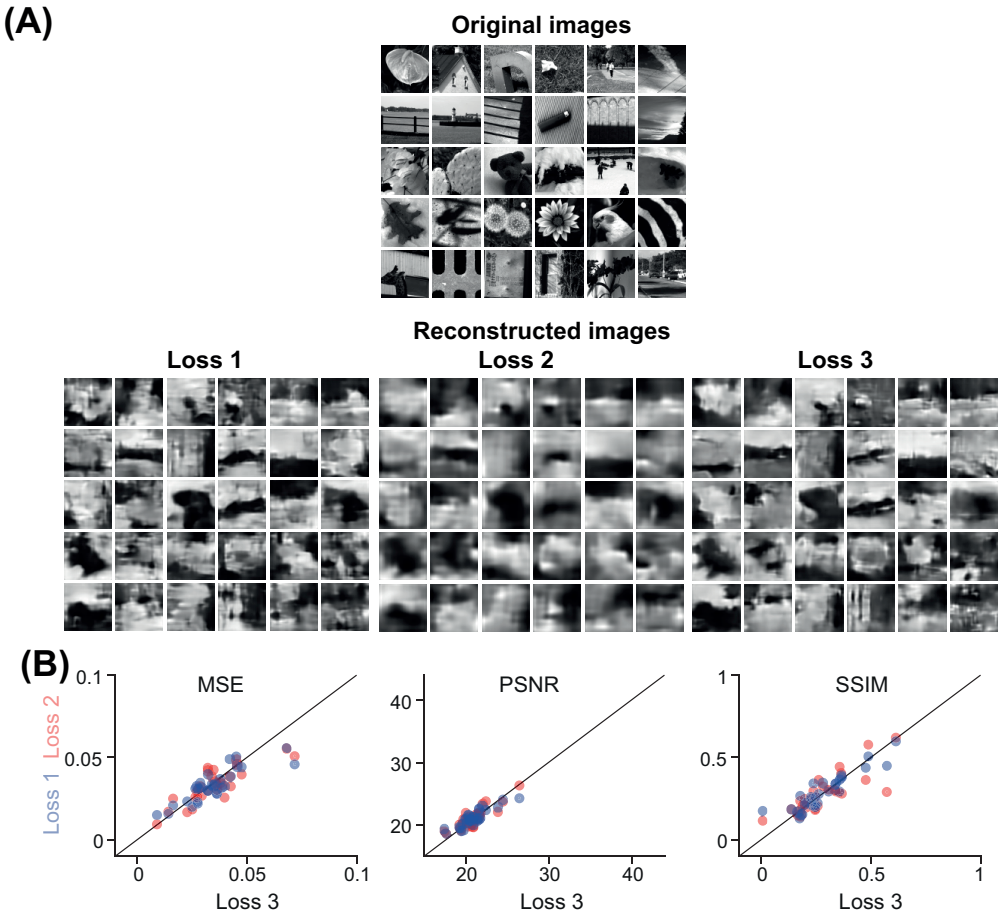


Figure 2: Reconstructed natural images from experimental RGC spikes with different loss functions. (A) Example original stimulus images in the test data (top) and final reconstructed images (bottom) from RGC spikes with three loss functions. (B) Reconstruction errors measured by MSE, PSNR, and SSIM are compared between Loss 3 and Loss 1 (blue) and Loss 2 (red).

this fMRI dataset by our SID directly with the same network structure. First, we convolved the stimulus images, here, 6's and 9's, with the same set of receptive field of RGCs used previously to simulated spikes, which is a population of 90 simulated cells, similar to the simulation of MNIST and CIFAR image datasets.

For fMRI data, each stimulus image triggered a response represented by a 3092-dimension vector as activity patterns. For a detailed comparison, we conducted two-fold cross-validation: decoding two sets of neural signals, fMRI and spikes, with two different models. We choose a recent state-of-the-art method termed deep generative multi-view model (DGMM) [41], which is developed in the context of fMRI decoding. Thus, both models, DGMM and SID, are used for decoding with both fMRI data and our simulated spikes to see if there is an advantage of neural data type, as well as an advancement of decoding method. Within such a validation, one can observe decoding results in four conditions: decoding fMRI by DGMM (DGMM-fMRI), decoding spikes by DGMM (DGMM-Spike), decoding fMRI by SID (SID-fMRI), and decoding spike by SID (SID-Spike). In all cases, we used 90 images for training and 10 images for test.

3. Results

To test the capability of our SID model for reconstruction of visual scenes, we first use our decoding model to reconstruct natural stimulus from biological experimental data that consists of a population of RGCs recorded in

salamander retina triggered by natural image and video. To further test the generalization capability of our model, a set of simulated experiments with the encoding model were conducted, in which we reconstructed stimulus images of MNIST, CIFAR10, and CIFAR100 from simulated spikes. In addition, dynamic videos were also tested with simulated RGC spikes and with the SID model trained from CAFAR10. The results show that our SID can decode natural scenes from both biological and simulated spikes with very good precision.

3.1. Decoding visual scenes from experimental data

Here the experimental data are a population of RGCs recorded from salamander retina (see Methods), where the stimuli include static natural images and a dynamic video, and the spike counts of the recorded RGCs are used as the input of our decoding model to reconstruct natural scenes.

3.1.1. Decoding natural image stimulus

For natural images, there are 300 images as stimulus and 80 RGCs showing spike responses. For the decoder, the spike counts of 80 RGCs in response to each image are collected at one end, the other end is a training set of 270 images of 64*64 pixels each. For each image, spike-image converter up-sample a population of 80 spike counts into an image with 64*64 pixels to get an intermediate image, which then is mapped to the target training image with the image-image autoencoder.

Figure 2 (A) shows the decoding results on some example test images, where the global contents of images are reconstructed well while some details are missed. Depending on the loss functions used for training, intermediate images can be very different (Supplemental Figure 1), however, final reconstructions are similar to each other with some fine differences in the detailed textures. When there are two losses for each stage of the decoder as in $\mathbf{L}_1$ and $\mathbf{L}_2$, the intermediate image (denoted as $\mathbf{O}_1$) is an preliminary image close to the target image. When there is one loss for both stages of the model as in $\mathbf{L}_3$, the intermediate image is like noise without any visual structures.

Figure 2 (B) shows the performance of our model characterized by three typical measures of reconstructed images: the mean square error (MSE) that describes the absolute difference of every pixel, the peak-signal-to-noise-ratio (PSNR) that characterizes the global quality, and the structural similarity index measure (SSIM) that captures the details or image distortion, for evaluating the reconstruction results. Whereas there is a variation between different test images, in general, all three loss functions give similar performances in terms of three measures. However, these measures are not perfect characteristics. For example, the reconstructed images with $\mathbf{L}_2$ look as not good as the other two, however, it gives the similar measure errors. Therefore, these measures are a general guideline rather than the precise characteristic, as it seems that there is no perfect measure [42, 43].

3.1.2. Decoding dynamic video stimulus

So far, most of the decoding studies focus on static natural images, or some simple artificial dynamic scenes [2, 17]. To study dynamical visual scenes, such as videos with highly complex dynamics in both spatial and temporal domains, one has to, traditionally, understand how to deal with strong temporal adaptation observed in single neurons and neural circuits in the retina [28]. Here with the video stimulus containing 1800 frames, and a population of 90 RGCs spike trains, our SID can overcome this difficulty by training of randomly selected frames out of the whole video, and testing on a subset untrained frames (see Methods).

From the reconstructed sample frames of test dataset shown in Figure 3 (A), one can see that our decoding model obtains high precise results, which can not only reconstruct the global content of each frame, but also some details of images that are missed in previous decoding studies [17, 20, 19]. The quantified performance is shown in Figure 3 (B). In contrast to the image results, all three measures, MSE, PSNR, and SSIM, indicate that the best results are obtained with $\mathbf{L}_3$. Such a good performance can also be seen from the sample images where both global contents and fine details are obtained.

3.2. Decoding visual scenes from simulated spikes

To further test the generalization capability of our decoding model, we did numerical experiments on simulated RGC data by creating a retinal encoding model based on the typical linear-nonlinear model and biological experimental data to simulate RGC responses (see Methods). In this way, one can simulate RGC spiking responses for any given stimuli. For comparison with experimental RGC data, we simulated the same number of 90 RGCs to obtain spikes.

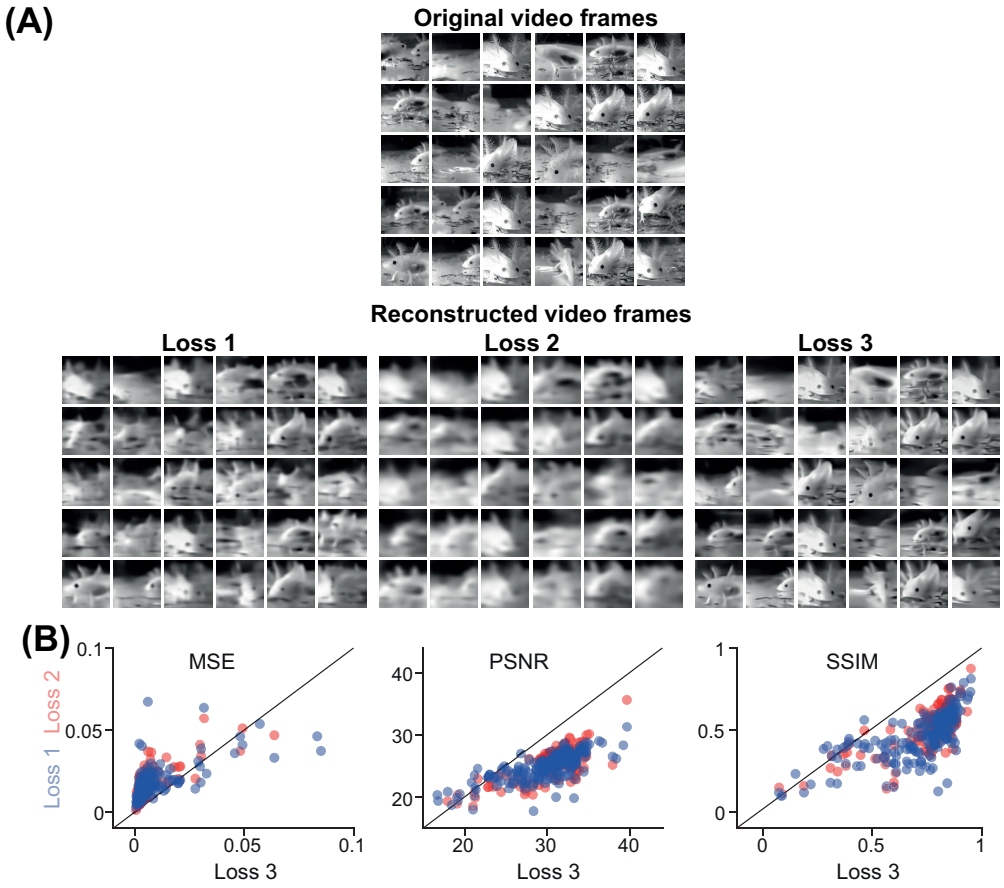


Figure 3: Reconstructed video frames from experimental RGC spikes with different loss functions. Similar to Figure 2, except that here are 30 random test video frames of a continuous video in (A), and a set of 50 random test frame images in (B), where each point is one of the test images.

We first did experiments on three popular image dataset, MNIST, CIFAR10, and CIFAR100. A population of neural spikes was obtained for each image with the encoding model. Then the SID was trained for the training set, and tested on a separate test data. The reconstructed example images are shown in Figure 4 with good quality. In particular, the decoding performance for MNIST is almost perfect.

With the good reconstruction from simulated spikes with a population of only 90 RGCs, one can ask the question that how the decoding performance can be changed by a reduced number of RGCs. Thus, we re-trained the SID with a randomly selected subset of 90 RGCs as the input for the model, as shown in Figure 5. Surprisingly, with only 20 RGC input, the SID can still recover a large part of the information from the original stimulus images. Indeed, the performance measured by MSE, PSNR, and SSIM is systematically decreasing when fewer RGCs are used for input. However, the performance tends to be convergent with 90 RGCs, which is the maximum number of RGCs recorded in our experimental data. It might be possible that the performance can be a little higher when more than 90 RGCs are used. We leave this possibility for the future study when more RGCs can be recorded (See Discussion).

As our decoder reconstructs the image per pixel directly, we expect that the training of SID can be done without any information about the category. Both MNIST and CIFAR datasets are originally used for image classification with labels of category. Thus, we re-trained our SID on these datasets but removing one of the known categories. Then we tested the reconstruction of the missing images from this untrained category. As shown in Figure 6, the 8's digital images are reconstructed from the SID that is trained without 8's of digital images. In both training conditions, the sample size of the training set is the same, so that the only difference between the two is the missing category. We found that the 8's can be decoded with a high accuracy, which means that the training of SID does not use any

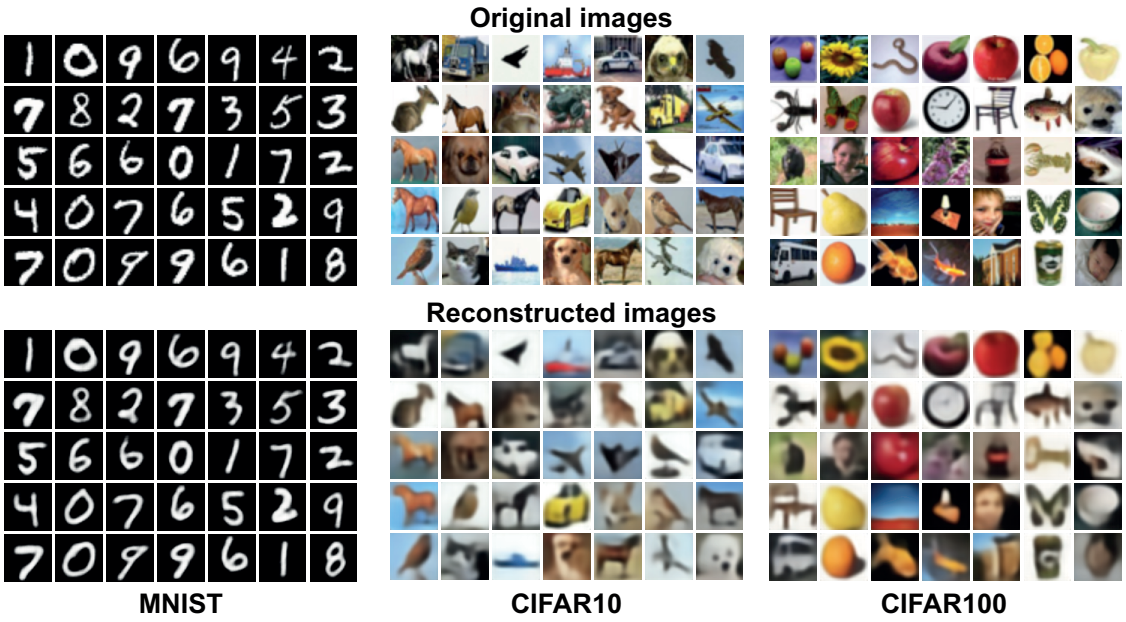


Figure 4: Reconstructed images from simulated RGC spikes in datasets of NMIST, CIFAR10, and CIFAR100. (Top) Example original images in test dataset. (Bottom) Reconstructed images from the spikes of 90 simulated RGCs.

information about the category of the dataset. Similarly, the various ship images from the missing category during the training can also be reconstructed well on the more complex CIFAR dataset as in Figure 6. However, the performance characterized by three measures of MSE, PSNR, and SSIM shows that there is a subtle difference between two versions of the test results: the SID trained with all classes has better performance, compared to the SID training without one category. Taken together, these results suggest that our SID has a good generalization ability, and the rich image content of the dataset can improve the decoding ability of the SID.

3.3. Cross-validation with existing methods

As there is no direct experimental data of neuronal spikes recorded for MNIST so far, we compared our results with the decoding results based on the models developed for fMRI signals. Previous studies evaluated the performance of quite a few decoders based on one particular dataset where a small part of MNIST images, 6's and 9's, were used as stimulus and the fMRI data was collected for human subject [39, 40, 44, 45, 46, 19, 41, 23]. We compared our SID with a recent state-of-the-art method, termed Deep Generative Multi-view Model (DGMM), developed for decoding of fMRI data [41]. For detailed comparison, we used the same fMRI dataset but evaluated the decoding performance based on both fMRI responses and our simulated spikes, as well as two decoding models: DGMM and our SID, such that there are four pairs of model-data test: DGMM-fMRI (*e.g.* fMRI decoding by DGMM), DGMM-Spike, SID-fMRI, and SID-Spike.

The reconstructed example images are shown in Figure 7, which can be characterized by MSE, PSNR and SSIM from a single run and a set of 10 runs with different initial conditions of the decoding models (Supplemental Figure 2). Interestingly, the SID gives a better performance for decoding of fMRI as well. Compared to DGMM, our SID outperforms on decoding with both fMRI signals and simulated spikes. Figure 7 shows the normalized model performance for each individual test images, where all model-data pairs are normalized by the SID-Spike. In this way, SID-Spike gives a ratio as 1, and other model-data pairs have a performance smaller than 1. In particular, the performance increases in ascending order: DGMM-fMRI $<$ DGMM-Spike $<$ SID-fMRI $<$ SID-Spike. These results suggest that our SID, as a general decoding method, shows a good capability for fMRI decoding as well.

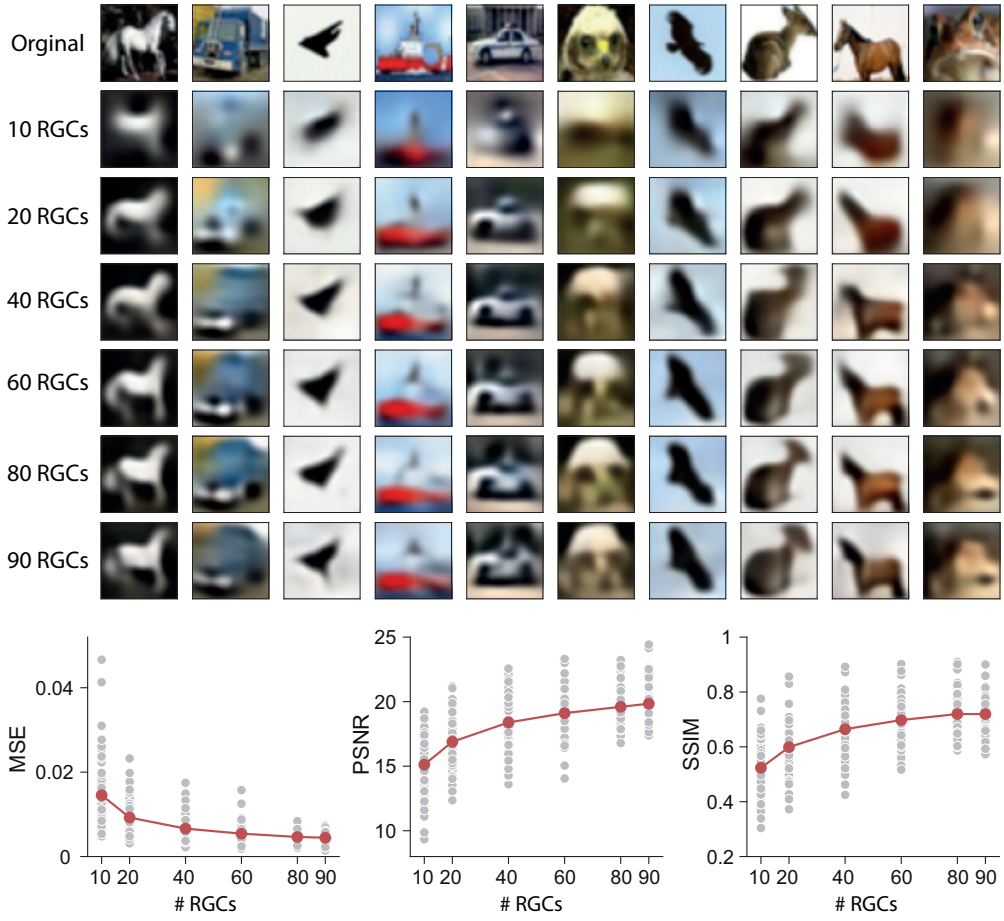


Figure 5: Reconstruction with a reduced number of RGC inputs. (Top) Reconstructed example images from CIFAR10 with different subsets of RGCs, from 10 to 90 RGCs. (Bottom) The performance of SID increased with more RGCs. Each gray point represents one of 50 randomly selected test images. Red points are the averaged measures over all test images.

3.4. Model generalization to real visual scenes

Next, we used the encoding model to simulated RGCs in response to dynamic videos. For comparison, the same set of videos used in [20] were encoded into RGC spikes. Then the SID model trained with CIFAR10 was used to reconstruct the videos from simulated RGC spikes. As shown in Fig 8 (A), the SID can reconstruct the dynamic videos from spikes very well. Compared with the fMRI decoded results [20], our reconstructed videos contain more semantic contents and are much more refined. What is more, the SID model used here is trained with CIFAR10, which means that our SID model has a great generalization capability for arbitrary visual scenes. The complete videos reconstructed are available online (see Methods).

Finally, based on the results above, we implemented a real-time artificial vision system (see Methods), where real-time videos are captured by a standard camera, then encoded into a population of neural spikes, which are fed into the SID to reconstruct the visual scenes. Figure 8 (B) shows a snapshot of a real-time recorded video. Certainly, the speed of reconstruction and resolution of decoded video can be greatly improved with more advanced hardwares. For this real-time system, the SID is also trained with the same CIFAR-10 dataset. These results suggest that our model can sever as a general framework to analyze those data recorded by artificial event-based vision systems [47].

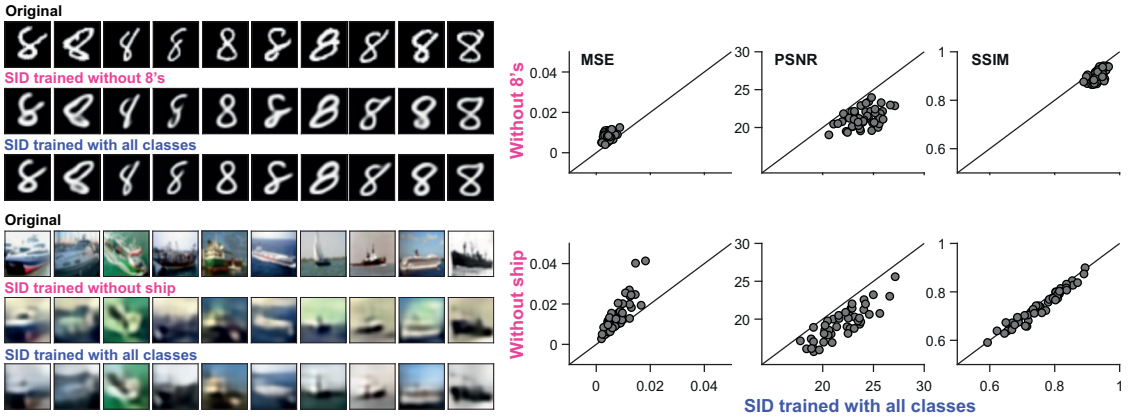


Figure 6: SID generalization to the untrained images of one category of MNIST and CIFAR. (Left, top) Original test 8's images from MNIST data, reconstructed example 8's images by the SID trained on other images but without the 8's category, and reconstructed example 8's images by the SID trained on images with all the categories. (Left, bottom) Similar results from CIAFR data, where the ship category is not used for the SID training. (Right) Comparison of generalization ability between the SID with and without one category of MNIST (top) and CIAFR (bottom). Each gray point represents one of 50 randomly selected test images.

4. Discussion

In this study, we proposed an end-to-end spike-image decoder to reconstruct stimulus images from neural spikes based on the retinal ganglion cells. The capability of SID was tested on experimental RGC data for both static and dynamic natural scenes to obtain state-of-the-art reconstruction performance. Furthermore, by using an additional encoding model, the great performance of reconstruction of arbitrary visual scenes was demonstrated on popular image datasets. In addition, our SID can also be used for decoding of fMRI data and show a very good performance compared to other methods developed in the context of fMRI decoding. Finally, we examined the generalization ability of SID by using a pre-trained SID to decode various types of dynamic videos to achieve real-time encoding and decoding visual scenes by neural spikes.

4.1. Rate code v.s. temporal code

In our current study, the firing rate averaged over a number of trials for experimental data was used for our decoder and reconstruction. Both firing rate code and temporal code have been used in the field of neuronal coding [48, 2, 34]. In practice, for analyzing experimental data, the difference between these two depends on many factors. In our case presented here, experimental data were collected under a condition that a 30 Hz frame rate was used for video stimulation, therefore, to match the sample rate of video frames, we binned the spike trains into 30 Hz as well. For image stimulation, each individual image was presented for 200 ms and then followed by an 800 ms empty display, therefore, for each image, we used the number of spikes during 300 ms after the onset of image presentation. We also conducted the reconstruction by binarizing the firing rate of RGCs into a single spike train with 0's and 1's only, and the results are similar to the current presentation.

From the experimental side, one could conduct new experiments with a high frame rate of video sampling, for instance, 120 Hz can be used, so that one only need to consider a smaller bin, 1000/120≈8.3 ms, in which more precise spike times can be recovered given that the spare firing of the retinal ganglion cells in general can not generate 2 or more spikes within 8.3 ms in salamander, the animal species used here. On the other hand, these neurons in this animal species have a very low firing rate such that within a bin of 33 ms used in our study, there are 93.12% of zero spike, 5.37% of single spike, 1.14% of double spikes, and 0.37% of triple or more spikes. Therefore, although the firing rate was used for our decoding, the low firing rate in neurons is an approximation of temporal coding of spike times.

Given the experimental issues listed above, that is the reason why we use simulated spikes to demonstrate our method. In our simulations of image datasets and real-time videos, the decoding is based on the temporal code with

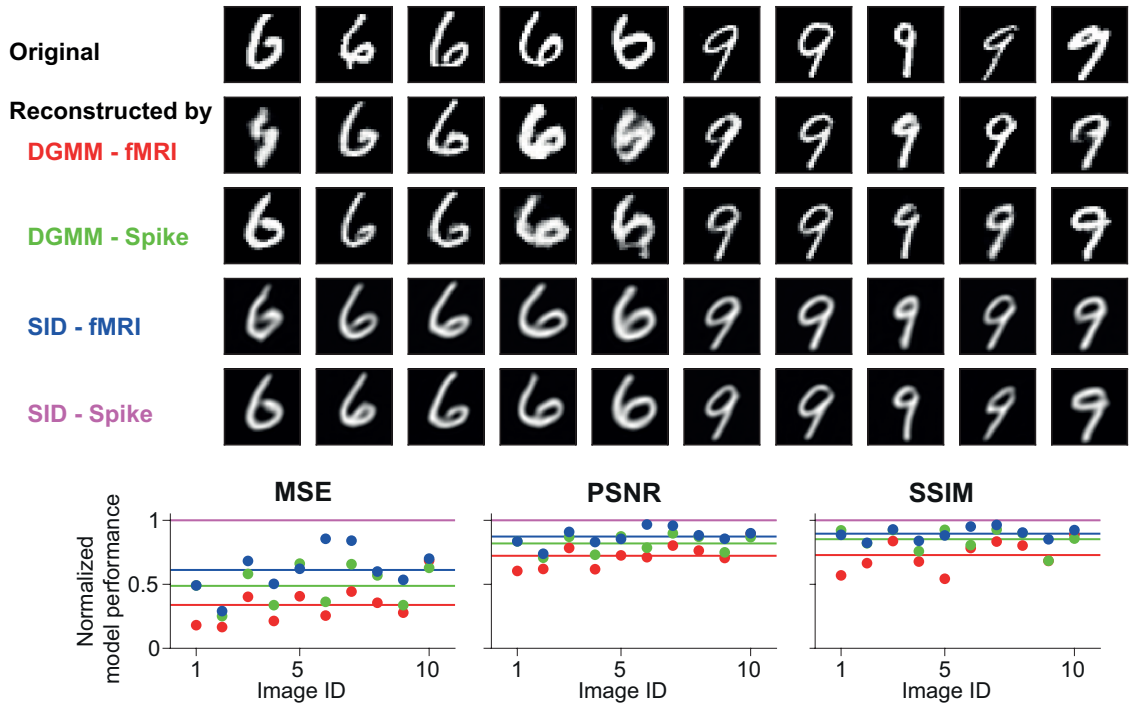


Figure 7: Reconstructed stimulus images from fMRI responses and simulated spikes. (Top) All of the 10 original images in the test dataset used for fMRI recordings, and the reconstructed images from fMRI data and the simulated spikes with the decoder of DGMM and SID cross-validated by a pair of model-data test, *e.g.* DGMM-fMRI stands for fMRI decoding by the DGMM model. For comparison, cross-validation is done by using DGMM for simulated spikes and our SID for fMRI data. (Bottom) Normalized model performance measured by MSE, PSNR, and SSIM. The performance of each pair of model-data is normalized by the SID-Spike. Each colored data point corresponds to each test image in the same order of image ID as in the original stimulus images on the top row. Colored solid lines represent the mean values of each error measure: DGMM-fMRI (red), DGMM-Spike (green), SID-fMRI (blue), and SID-Spike (purple).

the exact spikes in a single trial. Thus, our results presented here are provided as a proof-of-concept for using a population of single-trail spikes to reconstruct visual scenes.

4.2. Implication for other neuronal systems

Our method could serve as a general model of neural coding, similar to other neural coding models, such as the general linear model [13], to study the relationship between neuronal signal and stimulus. In our case, the SID is a non-linear decoder to reconstruct image stimulus from neural spikes, which can be seen as an inverse function of spike-image mapping. Thus our method could be generalized to other parts of the brain as long as the neurons deliver different spike patterns for different stimuli.

The retina has been thought of as a relatively simple neuronal circuit, however, it has been suggested that many complex computations can be realized by their neurons, in particular, the RGCs [11]. It has been studied intensively to develop various types of models for the encoding of visual scenes in the retina [1, 13, 16]. Similar to our current study, other recent work also focuses on neuronal spikes in the retina to develop the novel decoding methods [17, 24, 25]. All the methods used these studies take a similar approach by using the advantage of the recent advanced machine learning technique. Furthermore, the similar CNN decoding approach has been used for other parts of the visual system and different types of neural signals, such as calcium imaging in animal visual cortex [27], and classical fMRI signal in human visual cortex [19]. Therefore, our method could be applied to these types of neural signals from other parts of the visual system.

In the current study, we provide a case for using the SID to the fMRI signal. It is easy to expect that the decoding performance is higher for neural spikes compared to the fMRI signal due to the nature of neural signal, where spikes are much more refined than fMRI. The two-fold cross-validation shown in Figure 7 confirm this expectation. With

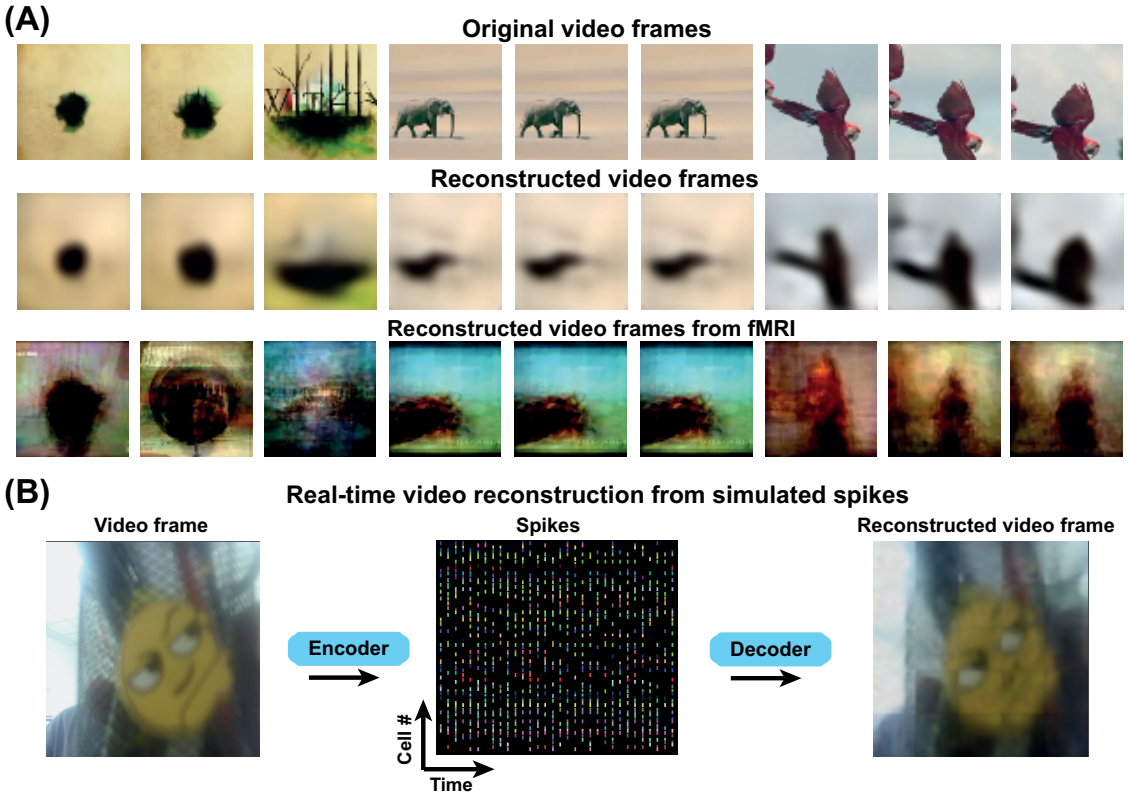


Figure 8: SID can be generalized to reconstruct arbitrary videos with a pre-trained model on CIFAR10 without any fine tunning. (A) Three segments of original videos (top), reconstructed videos from simulated spikes with a SID model pre-trained on CIFAR10 data (middle), and reconstructed videos based on fMRI signals from [20] (bottom . (B) Decoding of real-time videos. (Left) One example frame of a real-time video. (Middle) A population of spike trains obtained by the encoder. Each line represents one RGC's spike train. (Right) The same frame reconstructed in a real-time fashion by the SID. Real-time videos are available to view online.

the same model, either SID or DGMM, the decoding results by using neural spikes are better than those by fMRI. However, interestingly, we also observe a methodological advantage by comparing two decoding methods: SID and DGMM. It should be noted that DGMM is a recent and state-of-the-art method developed in the context of fMRI decoding [41]. Even in this condition, our SID can achieve a better decoding performance than DGMM by using the same fMRI data. Therefore, our method can be seen as a more general model to cope with different types of neural signals beyond neural spikes.

Recent experimental advancements in neuroscience can collect a large population of neurons simultaneously. In particular, in the retina, a population of spike trains from hundreds of retinal ganglion cells can be obtained with well-controlled visual scenes, such as images and movies [17, 34]. New techniques may record several thousands of neurons simultaneously [49, 50, 51]. Together with other recent studies [17, 25], our results in this work show that a better decoding performance can be archived with neural spikes, which is beyond the scope of fMRI signal [19, 20, 41].

4.3. Model limitation and further study

Our decoder can be considered as a general framework, in which the detailed network structures are flexible. Here we only explored the simple and perhaps the most basic structure, e.g., a simple dense connected and layered network for the spike-image converter, similar to a multilayer perceptron, and a typical network for the image-image autoencoder. One of the advantages of the current framework is that one can use any other types of spike-image converter, which is similar to existing models of decoding by transferring spikes into image pixels (see [17]), and any other types of image-image encoder, which is similar to existing models of image analysis (such as generative

adversarial network or GAN [52]) in the deep learning field. Given the rapid development of architecture design for artificial, spiking, and a mixture of both [53], neural networks, it is possible to use other network architectures to further improve the decoding ability.

Our SID does not include the temporal filter, but only the spatial filter, i.e., the spatial receptive field of RGCs. RGCs have both spatial and temporal filters to take into account of spatial structure [35] and temporal dynamics of stimulus [28]. Nonlinear structure in space has been intensively studied in the retina [35, 54, 55]. In contrast, the temporal dynamics are more complex and entangled with the stimulus itself. Even for the simple artificial stimulus, such as white noise, there is a strong neuronal adaption although there is no correlation within white noise stimulus [28]. For video stimulus with natural scenes, neuronal adaption entangled with correlation in the video in the temporal domain makes the investigation of neuronal coding of natural scene extremely difficult [56].

Here we overcome this difficulty by using a training set of randomly selected frames from the video in experimental data. In this way, one can break the temporal correlation between video frames. Essentially, this manipulation resembles presenting a large set of natural images sequentially, termed as naturalistic video, such that the temporal correlation between frames can be reduced to a minimal level [56]. The limited recording time during neuroscience experiments restricts the stimulus to a short segment of video used for RGCs. Thus, we tested the decoding capability of our method by using the stimulation of a large set of images, CIFAR, to mimic the full pixel space encountered by the model. To our surprise, our model trained by CIFAR can be generalized to any dynamic natural scenes without any re-training or fine-tunning.

Dynamic visual scenes are highly complex with the information presented in a spatiotemporal fashion and high-order correlation [4]. Recent advancements of computer vision make some breakthroughs for analyzing these complex natural scenes [57], for instance, the PredNet model can decode the next frame based on the previous frame of video [58]. However, the efficiency, generalization ability, and adaption or transfer learning between different tasks, of well-trained models are still far from human performance [59]. Instead, our modeling framework shows a great generalization ability for the decoding of visual scenes with a pre-trained model. In future work, we expect to develop a time-based SID and use temporal filters for decoding.

4.4. Implication for artificial visual systems

The proposed model framework could also be used for visual neuroprostheses and other bioengineering devices of brain-machine interface. For the retinal neuroprostheses, the benefit of a decoding model is to justify the spiking patterns produced by the targeted downstream neurons, such that electrical stimulation should be able to close to those desired patterns of retinal neural activity in a prosthesis [60]. Such control is important and beyond the traditional way of computing the distance between two or more spike trains in general [61, 62, 63], as recent studies suggest that the performance of neuroprosthesis for restoring vision can be improved by the encoding/decoding algorithms besides of hardware design [60, 64, 65]. Our study here proposes a novel decoding framework, as well as provides a proof-of-principle of encoding and decoding visual scenes with only a small population of neurons. Given the limitation of physical devices, such as only a small patch of brain area can be stimulated at one time, our framework suggests a feasible approach by using a small set of neurons for high precision decoding, which may systematically improve the capability of brain-machine interface.

The model framework proposed in this study could be used for other artificial visual systems. The main feature of our framework is to make use of neural spikes. Advancements of recent artificial intelligence computing align with the development of the next generation of neuromorphic chips and devices, where the new data format is processed as spikes or events [47, 66, 67, 68, 69]. Therefore, our method can be applied to neuromorphic retina cameras with spiking or events, such as the dynamic vision sensor [70]. Taken together with neuromorphic hardware and event/spiking computing algorithm, the next generation of computational vision can develop a better system for artificial vision.

References

[1] E. J. Chichilnisky, A simple white noise analysis of neuronal light responses, Network 12 (2) (2001) 199–213.

[2] T. Gollisch, M. Meister, Rapid neural coding in the retina with relative spike latencies, Science 319 (5866) (2008) 1108–11.

[3] F. Rieke, D. Warland, R. de Ruyter van Steveninck, W. Bialek, Spikes: Exploring the Neural Code (Computational Neuroscience), The MIT Press, 1996.

[4] E. P. Simoncelli, B. A. Olshausen, Natural image statistics and neural representation, Annual Review of Neuroscience 24 (24) (2001) 1193.

[5] M. C.-K. Wu, S. V. David, J. L. Gallant, Complete functional characterization of sensory neurons by system identification, Annual review of neuroscience 29 (2006) 477–505.
[6] P. D. Jadzinsky, S. A. Baccus, Transformation of visual signals by inhibitory interneurons in retinal circuits, Annual Review of Neuroscience 36 (2013) 403–28.
[7] W. N. Grimes, A. Songco-Aguas, F. Rieke, Parallel processing of rod and cone signals: Retinal function and human perception, Annual Review of Vision Science 4 (2018) 123–141.
[8] J. O'Brien, S. A. Bloomfield, Plasticity of retinal gap junctions: Roles in synaptic physiology and disease, Annual Review of Vision Science 4 (2018) 79–100.
[9] M. Rivlin-Etzion, W. N. Grimes, F. Rieke, Flexible neural hardware supports dynamic computations in retina, Trends in Neurosciences 41 (2018) 224–237.
[10] J. B. Demb, J. H. Singer, Functional circuitry of the retina, Annual Review of Vision Science 1 (2015) 263–289.
[11] T. Gollisch, M. Meister, Eye smarter than scientists believed: Neural computations in circuits of the retina, Neuron 65 (2) (2010) 150–164.
[12] A. F. Meyer, R. S. Williamson, J. F. Linden, M. Sahani, Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation, Frontiers in systems neuroscience 10 (2017) 109.
[13] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, E. P. Simoncelli, Spatio-temporal correlations and visual signalling in a complete neuronal population, Nature 454 (7207) (2008) 995.
[14] Q. Yan, Z. Yu, F. Chen, J. K. Liu, Revealing structure components of the retina by deep learning networks, NIPS Symposium on Interpretable Machine Learning. arXiv:1711.02837 (2017).
[15] Q. Yan, Y. Zheng, S. Jia, Y. Zhang, Z. Yu, F. Chen, Y. Tian, T. Huang, J. K. Liu, Revealing fine structures of the retinal receptive field by deep learning networks, arXiv preprint arXiv:1811.0229 (2018).
[16] Z. Yu, J. K. Liu, S. Jia, Y. Zhang, Y. Zheng, Y. Tian, T. Huang, Towards the next generation of retinal neuroprosthesis: Visual computation with spikes, arXiv preprint arXiv:2001.04064 (2020). arXiv:http://arxiv.org/abs/2001.04064.
[17] V. Botella-Soler, S. Deny, G. Martius, O. Marre, G. Tkačik, Nonlinear decoding of a complex movie from the mammalian retina, PLoS Computational Biology 14 (5) (2018) e1006057.
[18] V. Gilja, P. Nuyujukian, C. A. Chestek, J. P. Cunningham, M. Y. Byron, J. M. Fan, M. M. Churchland, M. T. Kaufman, J. C. Kao, S. I. Ryu, et al., A high-performance neural prosthesis enabled by control algorithm design, Nature Neuroscience 15 (12) (2012) 1752.
[19] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, Z. Liu, Neural encoding and decoding with deep learning for dynamic natural vision., Cerebral cortex 28 12 (2018) 4136–4160.
[20] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, J. L. Gallant, Reconstructing visual experiences from brain activity evoked by natural movies, Current Biology 21 (19) (2011) 1641–1646.
[21] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, S. Dehaene, Inverse retinotopy: inferring the visual content of images from brain activation patterns, Neuroimage 33 (4) (2006) 1104–1116.
[22] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, J. L. Gallant, Bayesian reconstruction of natural images from human brain activity, Neuron 63 (6) (2009) 902–915.
[23] K. Qiao, C. Zhang, L. Wang, J. Chen, L. Zeng, L. Tong, B. Yan, Accurate reconstruction of image stimuli from human functional magnetic resonance imaging based on the decoding model with capsule network architecture, Front Neuroinform 12 (2018) 62.
[24] O. Marre, V. Botella-Soler, K. D. Simmons, T. Mora, G. Tkačik, M. J. Berry II, High accuracy decoding of dynamical motion from a large retinal population, PLoS Computational Biology 11 (7) (2015) e1004304.
[25] N. Parthasarathy, E. Batty, W. Falcon, T. Rutten, M. Rajpal, E. Chichilnisky, L. Paninski, Neural networks for efficient bayesian decoding of natural images from retinal neurons, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 6437–6448.
[26] G. B. Stanley, F. F. Li, Y. Dan, Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus, Journal of Neurophysiology 19 (18) (1999) 8036–8042.
[27] S. Garasto, A. A. Bharath, S. R. Schultz, Visual reconstruction from 2-photon calcium imaging suggests linear readout properties of neurons in mouse primary visual cortex, bioRxiv (2018) 300392.
[28] J. K. Liu, T. Gollisch, Spike-triggered covariance analysis reveals phenomenological diversity of contrast adaptation in the retina, PLoS Computational Biology 11 (7) (2015) e1004425.
[29] Y. Yue, L. He, G. He, J. Liu, K. Du, Y. Tian, T. Huang, et al., A simple blind-denoising filter inspired by electrically coupled photoreceptors in the retina, arXiv preprint arXiv:1806.05882 (2018).
[30] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, 2014, pp. 818–833.
[31] D. L. K. Yamins, J. J. Dicarlo, Using goal-driven deep learning models to understand sensory cortex, Nature Neuroscience 19 (3) (2016) 356.
[32] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift., arXiv preprint arXiv:1502.03167 (2015).
[33] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors., arXiv preprint arXiv:1207.0580 (2012).
[34] A. Onken, J. K. Liu, P. C. R. Karunasekara, I. Delis, T. Gollisch, S. Panzeri, Using matrix and tensor factorizations for the single-trial analysis of population spike trains, PLoS Computational Biology 12 (11) (2016) e1005189.
[35] J. K. Liu, H. M. Schreyer, A. Onken, F. Rozenblit, M. H. Khani, V. Krishnamoorthy, S. Panzeri, T. Gollisch, Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization, Nature Communications 8 (1) (2017) 149.
[36] A. Olmos, F. A. A. Kingdom, A biologically inspired algorithm for the recovery of shading and reflectance images, Perception 33 (12) (2004) 1463–1473.
[37] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324. doi:10.1109/5.726791.
[38] A. Krizhevsky, Learning multiple layers of features from tiny images, University of Toronto (05 2012).

[39] M. A. van Gerven, F. P. de Lange, T. Heskes, Neural decoding with hierarchical generative models, Neural computation 22 (12) (2010) 3127–3142.
[40] M. A. Van Gerven, B. Cseke, F. P. De Lange, T. Heskes, Efficient bayesian multivariate fmri analysis using a sparsifying spatio-temporal prior, NeuroImage 50 (1) (2010) 150–161.
[41] C. Du, C. Du, L. Huang, H. He, Reconstructing perceived images from human brain activities with bayesian deep multiview learning, IEEE Transactions on Neural Networks and Learning Systems 30 (8) (2019) 2310–2323. doi:10.1109/tnnls.2018.2882456.
[42] A. Hore, D. Ziou, Image quality metrics: PSNR vs. SSIM, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010. doi:10.1109/icpr.2010.579.
[43] Z. Wang, E. Simoncelli, A. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, IEEE, 2003. doi:10.1109/acssc.2003.1292216.
[44] Y. Fujiwara, Y. Miyawaki, Y. Kamitani, Modular encoding and decoding models derived from bayesian canonical correlation analysis, Neural computation 25 (4) (2013) 979–1005.
[45] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, International Conference on Machine Learning (2015) 1083–1092.
[46] Y. Miyawaki, H. Uchida, O. Yamashita, M. aki Sato, Y. Morito, H. C.Tanabe, N. Sadato, Y. Kamitani, Visual image reconstruction from human brain activity using a combination of multiscale local image decoders, Neuron 60 (2008) 915–929.
[47] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, D. Scaramuzza, Event-based vision: A survey, arXiv preprint arXiv:1904.08405 (2019).
[48] A. L. Jacobs, G. Fridman, R. M. Douglas, N. M. Alam, P. E. Latham, G. T. Prusky, S. Nirenberg, Ruling out and ruling in neural codes, Proceedings of the National Academy of Sciences of the United States of America (2009).
[49] G. Portelli, J. M. Barrett, G. Hilgen, T. Masquelier, A. Maccione, S. Di Marco, L. Berdondini, P. Kornprobst, E. Sernagor, Rank order coding: a retinal information decoding strategy revealed by large-scale multielectrode array retinal recordings, Eneuro (2016) ENEURO–0134.
[50] A. Maccione, M. H. Hennig, M. Gandolfo, O. Muthmann, J. Coppenhagen, S. J. Eglen, L. Berdondini, E. Sernagor, Following the ontogeny of retinal waves: pan-retinal recordings of population dynamics in the neonatal mouse, The Journal of physiology 592 (7) (2014) 1545–1563.
[51] G. Hilgen, S. Pirmoradian, D. Pamplona, P. Kornprobst, B. Cessac, M. H. Hennig, E. Sernagor, Pan-retinal characterisation of light responses from ganglion cells in the developing mouse retina, Scientific reports 7 (2017) 42330.
[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Advances in Neural Information Processing Systems 3 (06 2014).
[53] Q. Xu, Y. Qi, H. Yu, J. Shen, H. Tang, G. Pan, Csnn: An augmented spiking based framework with perceptron-inception., in: IJCAI, 2018, pp. 1646–1652.
[54] G. Schwartz, F. Rieke, Perspectives on: information and coding in mammalian sensory physiology: nonlinear spatial encoding by retinal ganglion cells: when 1 + 1 not equal 2, The Journal of General Physiology 138 (3) (2011) 283–90.
[55] G. W. Schwartz, H. Okawa, F. A. Dunn, J. L. Morgan, D. Kerschensteiner, R. O. Wong, F. Rieke, The spatial structure of a nonlinear receptive field, Nature Neuroscience (2012).
[56] A. Heitman, N. Brackbill, M. Greschner, A. Sher, A. M. Litke, E. Chichilnisky, Testing pseudo-linear models of responses to natural scenes in primate retina, bioRxiv (2016).
[57] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[58] W. Lotter, G. Kreiman, D. Cox, Deep predictive coding networks for video prediction and unsupervised learning, arXiv:1605.08104v5 (2017).
[59] A. H. Marblestone, G. Wayne, K. P. Kording, Toward an integration of deep learning and neuroscience, Frontiers in Computational Neuroscience 10 (2016) 94.
[60] S. Nirenberg, C. Pandarinath, Retinal prosthetic strategy with the capacity to restore normal vision, Proceedings of the National Academy of Sciences 109 (37) (2012) 15012–15017.
[61] J. D. Victor, Spike train metrics, Current opinion in neurobiology 15 (5) (2005) 585–592.
[62] M. v. Rossum, A novel spike distance, Neural computation 13 (4) (2001) 751–763.
[63] N. P. Shah, S. Madugula, E. Chichilnisky, J. Shlens, Y. Singer, Learning a neural response metric for retinal prosthesis, bioRxiv (2017) 226530.
[64] L. Yue, J. D. Weiland, B. Roska, M. S. Humayun, Retinal stimulation strategies to restore vision: Fundamentals and systems, Progress in retinal and eye research 53 (2016) 21–47.
[65] B. Yan, S. Nirenberg, An embedded real-time processing platform for optogenetic neuroprosthetic applications, IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society 26 (1) (2018) 233–243.
[66] D. Ma, J. Shen, Z. Gu, M. Zhang, X. Zhu, X. Xu, Q. Xu, Y. Shen, G. Pan, Darwin: A neuromorphic hardware co-processor based on spiking neural networks, Journal of Systems Architecture 77 (2017) 43–51. doi:10.1016/j.sysarc.2017.01.003.
[67] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y. Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie, L. Shi, Towards artificial general intelligence with hybrid tianjic chip architecture, Nature 572 (7767) (2019) 106–111. doi:10.1038/s41586-019-1424-8.
[68] S. Esser, P. Merolla, J. Arthur, A. Cassidy, R. Appuswamy, A. Andreopoulos, D. Berg, J. McKinstry, T. Melano, D. Barch, et al., Convolutional networks for fast, energy-efficient neuromorphic computing., Proceedings of the National Academy of Sciences 113 (41) (2016) 11441–11446.
[69] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, et al., Loihi: A neuromorphic manycore processor with on-chip learning, IEEE Micro 38 (1) (2018) 82–99.
[70] P. Lichtsteiner, C. Posch, T. Delbruck, A 128*128 120 db 15 us latency asynchronous temporal contrast vision sensor, IEEE journal of solid-state circuits 43 (2) (2008) 566–576.

Supplemental Materials: Reconstruction of Natural Visual Scenes from Neural Spikes with Deep Neural Networks

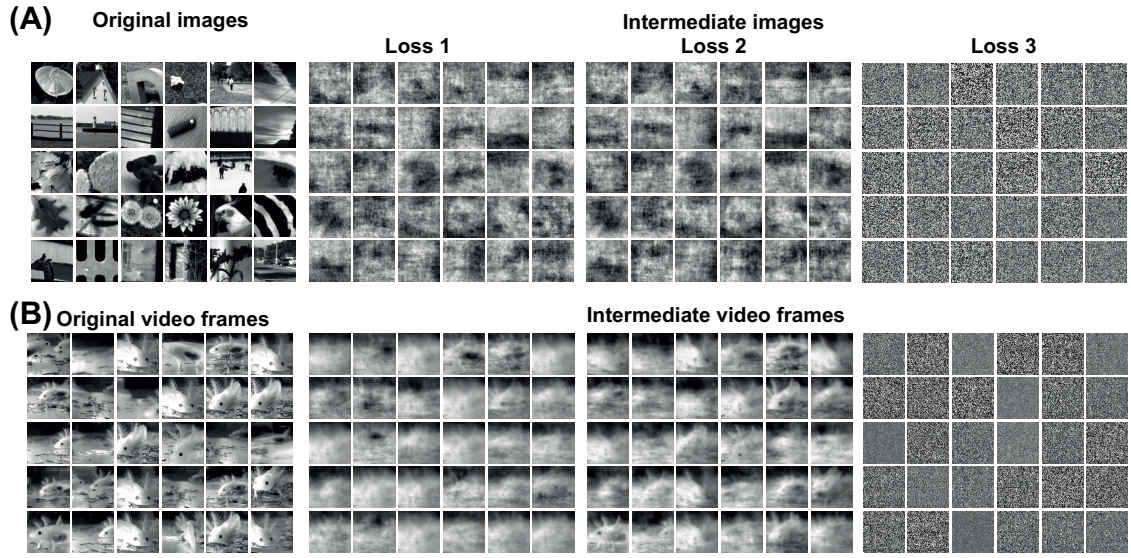


Figure S1: Reconstruction from experimental RGC spikes with different loss functions. Original stimulus images in the test data. Intermediate images from RGC spikes with three loss functions. Different losses can lead to quite different intermediate images, in particular, random images for Loss 3.

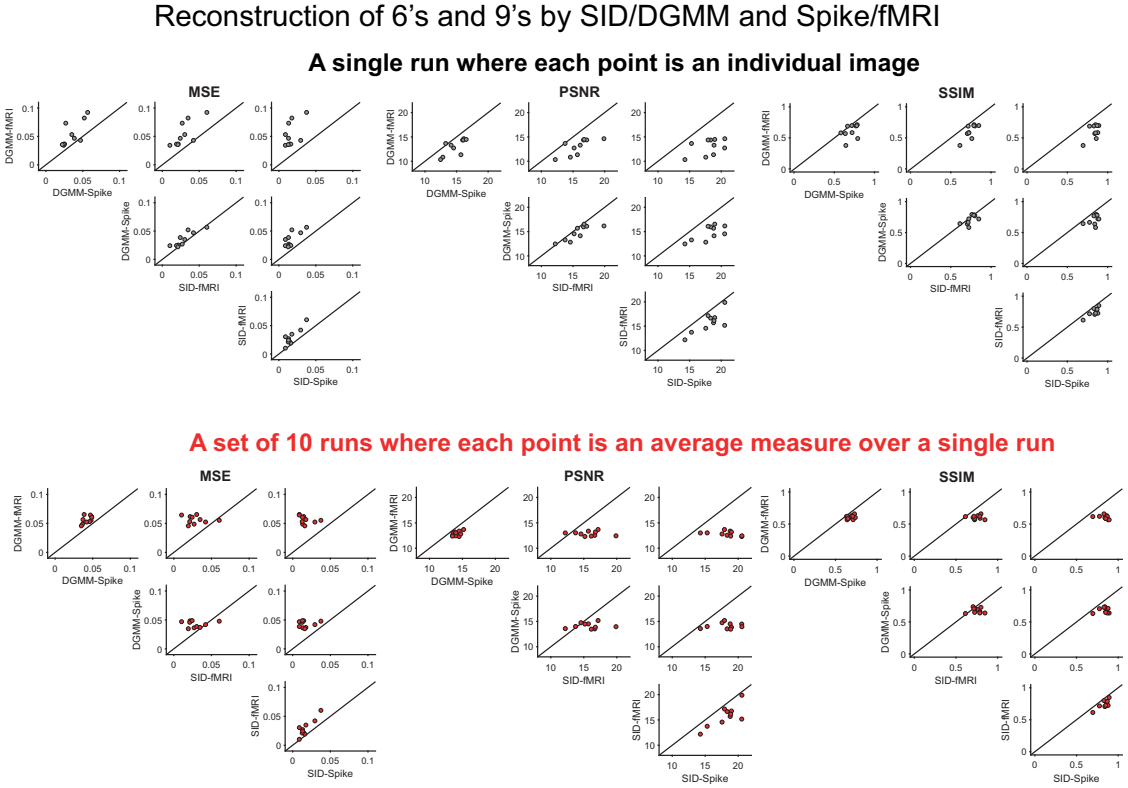


Figure S2: Reconstruction from fMRI data and simulated spikes with the decoder of DGMM and SID cross-validated by a pair of model-data test, *e.g.* DGMM-fMRI stands for fMRI decoding by the DGMM model. For comparison, cross-validation is done by using DGMM for simulated spikes and our SID for fMRI data. Model performance measured by MSE, PSNR, and SSIM in a single run of the SID model, and a set of 10 SID runs with different initial conditions. Each plot is a model-data pairwise comparison of MSE, PNSR, and SSIM.