



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/179515/>

Version: Accepted Version

Article:

Xu, Q, Shen, J, Ran, X et al. (2022) Robust Transcoding Sensory Information With Neural Spikes. *IEEE Transactions on Neural Networks and Learning Systems*, 33 (5). pp. 1935-1946. ISSN: 2162-237X

<https://doi.org/10.1109/tnnls.2021.3107449>

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Robust transcoding sensory information with neural spikes

Qi Xu, Jiangrong Shen, Xuming Ran, Huajin Tang, Gang Pan and Jian K. Liu

Abstract—Neural coding, including encoding and decoding, is one of the key problems in neuroscience for understanding how the brain uses neural signals to relate sensory perception and motor behaviors with neural systems. However, most of the existed studies only aim at dealing with the analogy signal of neural systems, while lacking a unique feature of biological neurons, termed spike, which is the fundamental information unit for neural computation as well as a building block for brain-machine interface. Aiming at these limitations, we propose a transcoding framework to encode multi-modal sensory information into neural spikes, then reconstruct stimuli from spikes. Sensory information can be compressed into 10% in terms of neural spikes, yet re-extract 100% of information by reconstruction. Our framework can not only feasibly and accurately reconstruct dynamical visual and auditory scenes, but also rebuild the stimulus patterns from functional magnetic resonance imaging brain activities. Importantly, it has a superb ability of noise-immunity for various types of artificial noises and background signals. The proposed framework provides efficient ways to perform multimodal feature representation and reconstruction in a high-throughput fashion, with potential usage for efficient neuromorphic computing in a noisy environment.

Index Terms—Neural Spikes, Cross-Multimodal, Reconstruction, Decoding, Spatio-temporal Representations, Denoising.

I. INTRODUCTION

SENSORY information is an essential and integrative part of the brain for processing the environment we are in [1]. The most basic stage of sensory perception is to recall the information perceived for higher cognition. Thus, intelligence machines are demanding an ability of representation and reconstruction of sensory information captured by various sensors, to achieve remarkably good computational intelligence tasks. Although various engineering effort has been made in this area, the biological information processing system still outperforms the best artificial systems in many fields such as processing cross-modalities and noise-immunity.

Qi Xu is with School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, 116024, China. (xuqi@dlut.edu.cn)

Qi Xu, Jiangrong Shen, Huajin Tang and Gang Pan are with the College of Computer Science and Technology, Zhejiang University, Zhejiang, 310027, China. (xuqi123@zju.edu.cn, jrshen@zju.edu.cn, htang@zju.edu.cn, gpan@zju.edu.cn)

Qi Xu, Jiangrong Shen and Jian K. Liu are with the Centre for Systems Neuroscience, Department of Neuroscience, Psychology and Behaviour, University of Leicester, Leicester, LE1 7RH, UK.

Jiangrong Shen is also with Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, 310027, China and Gang Pan is also with the State Key Lab of CAD & CG, Zhejiang University, Hangzhou 310027, China.

Xuming Ran is with the Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, 518055, China.

Jian K. Liu is also with the School of Computing, University of Leeds, Leeds, LS2 9JT, UK. (j.liu9@leeds.ac.uk)

Currently, our brain brings various types of sensor information with different sensory modalities from our surrounding environment. For which, neural coding is very essential for comprehending how neural systems respond to outside stimuli [2]. From the functional part of view, an efficient and effective coding system consists of two elementary parts, neural encoding and decoding [3] [4]. Encoding methods try to transfer outside stimuli into specific responses for further processing by downstream neural systems, then decoding aims to analyse and predict external stimuli from those specific format of data encoded by the encoding system. In biological coding system, neurons transmit the information when they receive the external stimuli by changing their membrane potential to fire a series of fast event termed spikes, forming spatio-temporal representations [5]. Thus spikes have been suggested as a more biological format to represent the input-output relations in neural systems than any other artificial one [6] [7], such as choosing real value based data as transmission media in artificial neural networks [8].

For encoding and decoding in biological information processing systems, there still remain big challenges to understanding the mapping between those external stimuli and fundamental spiking activities. For decoding, although some traditional methods have made significant progresses [9] [10], most of them tried to build artificial models with simple linear models and the questions are limited to either brain activity pattern classification or visual stimuli recognition measured by functional magnetic resonance imaging (fMRI) [11] [12]. On the other hand, deep learning models have enjoyed a great success in many areas of computer vision [8], it is very common for modern artificial deep neural networks (DNNs) to have tens of millions of parameters which lead to higher dimensional complexity and hierarchical structures. Inspired by biologically visual systems, hierarchical DNNs, using convolutional and pooling units to code external stimuli, have already shown in resembling some complex visual representations in human visual system [13]. For visual scenes, convolutional neural networks (CNNs) have been adopted to model the encoding of visual neurons, such as from retina, visual cortex to inferotemporal cortex [14] [15]. Thus, it is promising to build a more reasonable coding system between external stimuli and neural information processing with the aid of spiking activities and the structures of DNNs [7] [16]. Recent studies show that it is promising to use DNNs working with neural spikes for both encoding and decoding [17], [18], [2].

Inspired by the aforementioned studies, this paper proposes an efficient and effective coding system with neural spikes

for sensory information based on deep learning network models, named as deep spike pattern decoder (DSPD), that universally transcodes sensory information across multiple sensory modalities using neural spikes. Based on our recent work on decoding with neural spikes [18], the DSPD is an uniform coding framework consists of two parts: encoding and decoding. The encoding part maps outside sensory stimuli into image pixels, than transcodes pixels into neural representations efficiently in two ways. First in the spatial domain, compared to the high dimension of thousands of pixels, it only use a few hundreds of neurons to represented 100% of image pixels into 10% of neural spikes. Secondly, in the time domain, it can sample high-frequency images in videos into a spare temporal patterns, e.g., 30-60 Hz frame rates down to a few Hzs neural spikes firing sparsely over time. The transcoded spatialtemporal patterns in terms of neural spikes can be outputted and transferred in a high-throughput fashion to any downstream hardware for future processing.

Based on transcoded spiking representations, one can conduct any types of neural computation for practical tasks, ranging from classification, semantic recognition, to full-frame reconstruction. Here we show the capacity of our proposed framework in the context of coding of cross-multimodal sensory information, and its good capability of transfer learning, few-shot learning, and stimulus denoising. We evaluated our model on three different types of modal inputs: images, fMRI brain activities, and sound signals. In order to show the generalization ability, we applied the model to the clean and noise-free MNIST dataset and its four variations with strong noises and unrelated background signals. We also take the subsets from these datasets to show the capability of our model on few-shot learning. Experimental results demonstrate that our model is not only capable of perceiving and reconstructing corss-multimodal inputs (images, fMRI and sounds), but also having a good ability of generalization and noise-immunity. The qualitative and quantitative measurements show that our model can construct multimodal stimuli with a performance comparable to some other cognitive models. All together, our model provides an uniform and consistent coding system for efficiently and effectively transcodng sensory information via neural spikes. Inspired by biological underpinnings of how cross-multimodal patterns are perceived and represented by neural processing systems, our work suggest an approach of neuromorphic computing with neural spikes for handling multiple sources of sensor information.

II. METHODS

The proposed DSPD is a framework with a mixture of a biological encoding part and a deep neural nwtwork (DNN) based decoding part as illustrated in Figure 1. The encoding part is similar to an neural pathway of the sensory systems, which receive sensory information in the format of images, sound waves, or other types of artificial sensor data represented spatial, temporal, or spatiotemporal patterns. The output of the encoder is a sequence of spikes similar to biological neurons in response to stimuli. After encoding, the encoded information will be delivered to the decoding part. Depending on practical

tasks, the different decoders can be built for signal reconstruction, object recognition, semantic classification, etc. One can decode the spikes directly with spiking neural networks as decoder. Or one can also convert spikes into different format of data, for example, image pixels, to take advantage of the state-of-the-art computer vision techniques. The benefit of transcoding sensory information with neural spikes is to utilize the core concept of neuromorphic computing, e.g., energy and data efficient computing without loss of any information. Thus, our proposed framework is a unified spike transcoding system functioning as data compression, feature extraction, temporal encoding and decoding.

In this study, we put our proposed framework into the context of signal reconstruction in terms of image pixels. However, it is noted that our framework is fixable to account for other purposes, so that the exact architectures of the encoder and decoder are fixable to adapt to be other types of neural networks, or simple traditional statistical methods.

A. Transcoding with spikes

A spiking based encoding method differs from which in conventional DNNs. For a pattern recognition such as image classification task, DNNs usually take the raw pixel based value as input directly. In contrast, the spiking based encoding method would map those pixels into binary spike events that happen over time. Depending on data format, one can preprocess the raw sensory information by converting them into image pixels, for example, transferring sound waveforms into spectrograms of image pixels. Here the input images were unified as a size of 64×64 . Then an encoder is applied to images to convert them into spikes.

Unlike the previous study [18] where the encoder consists of a small number of retinal neurons. Here we used a set of 300 neurons to cover the whole image space. It is noted that with larger sizes of input images, one can use more number of neurons for encoding. All the encoding neurons were sampled over the entire image space, such that each neuron is located at a specific position in image space. The nonlinear filters are based on the receptive fields of 80 RGCs measured in experiment with white noise analysis [19] fitted with a 2D Gaussian for each cell. We then resampled the receptive fields of all 300 cells by rotating and shifting those experimental 80 cells to cover the pixel space of images, in this way one can overcome the underrepresented location bias due to the limitation of experimental recordings [20]. In addition, we used three subunits for each encoding neuron to utilize the idea of nonlinear subunits of sensory neurons. Each subunit has a Gaussian filter as the receptive field to capture a local image patch. Then the filtered image generates a value of mean over all pixels, which is transferred to obtain a spike count. Binary spikes are sampled from this processing to obtain a spatiotemporal spike pattern. We also tested other filters to generate spikes from inputs. Parameters of encoding neurons are not sensitive to the model outputs, as the spike pattern from the encoding neurons is playing a role of low-dimension representing of inputs, which is not participated into the training of the decoding part.

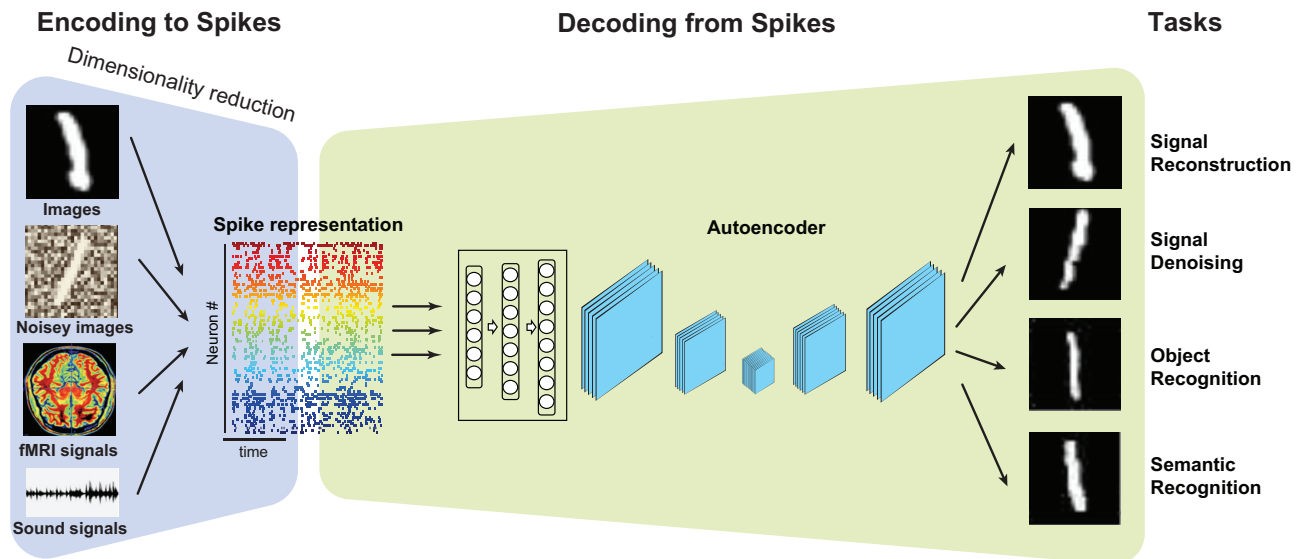


Fig. 1: The schematic diagram of DSPD framework.

197 B. Pattern decoding with spikes

198 After encoding, sensory information is represented by a
 199 sequence of spatiotemporal spiking pattern. To fulfill our aim
 200 of signal reconstruction, we used a similar decoder as in our
 201 recent work [18]. We first upsampled the spatial dimension
 202 into the original input image size. Then we used a three-
 203 layer fully-connected neural network, which is similar to
 204 a multilayer perceptron. The first layer receives the spikes
 205 coming from the neural encoding layer and the number of the
 206 neurons in the first layer is the same as the neurons of neural
 207 encoding layer. here 300, e.g., the same dimension as the
 208 number of neurons used for spiking representation. With the
 209 512 neurons in second layer (hidden layer) and 4096 neurons
 210 in third layer (output layer), we used the ReLU as activation
 211 functions to filter the non-negative value into image pixels.
 212 As input images are 64×64 , 4096 neurons were used in third
 213 layer as the output for signal reconstruction.

214 The propose of this upsampled image from spikes is to
 215 reconstruct the original signals, such that both have the same
 216 dimension. In case of implementing other tasks, upsampled
 217 images are not necessary. For the signal reconstruction, we
 218 adopt a typical autoencoder based on convolutional neural
 219 networks. This autoencoder consists of two parts as shown
 220 in Figure 1. In the first phase, the convolutional parts down
 221 sample the spike-based images. Notably, the most important
 222 part of the spike-based images are kept for recovering the
 223 texture and increasing the size. Meanwhile, through the de-
 224 creasing size of convolutional units, the noise and redundant
 225 components are filtered. Then the filtered images will recover
 226 through the increasing size of convolutional units in the up-
 227 sampling phase.

228 The size of the autoencoder here we used is 64C7-128C5-
 229 256C3-256C3-US2-256C3-US2-128C3-US2-64C5-US2 (C
 230 means convolutional layer, US means upsampling). The
 231 activation function is ReLU and the dropout rate is 0.25, we
 232 also use strides (2, 2) for padding and batch normalization
 233 for accelerating the training to achieve the convergence state

respectively.

234 Given an input pattern X , it will trigger a response $\mathbf{s} =$
 235 $\{s_1, s_2, s_3 \dots s_n\}$ within the encode method we just described
 236 on the 300 RGCs, here we adopt spike firing rate such as s_i
 237 in \mathbf{s} to represent the spike count of each RGC cell within a
 238 bin based on the sampling rating of pattern. Then the triggered
 239 responses are first fed into spike-image dense layer based con-
 240 verter which output an intermediate image $Y_1 = f_1(X)$, then
 241 the image-image autoencoder takes the Y_1 as input to map it
 242 to match the target pattern. So we can get a refining recon-
 243 struction pattern $Y_2 = f_2(Y_1)$, and the end-end training could
 244 be implemented by the two joint parts. f_1 and f_2 are their
 245 corresponding activation function, in this paper we adopted
 246 ReLU. Based on this information flow, we could get the
 247 training loss function, $loss = \lambda_1 \|Y_1 - X\| + \lambda_2 \|Y_2 - X\|$.
 248 With this loss function, the proposed model could be trained
 249 successfully.
 250

251 C. Datasets and codes

252 As shown in Figure 1, we evaluate our model on three
 253 different types of signals (visual images, fMRI brain activ-
 254 ity patterns, and sound signals [21] [22]). Specifically, we
 255 employed various different datasets: original MNIST with 10
 256 digital letters [23], MNIST with random white noise [24],
 257 MNIST with background images [24], MNIST with different
 258 level of artificial noise. fMRI brain activity datasets [25]
 259 Fig. 5) and sound signals of 10 spoken letter datasets [26].
 260

261 We used a dataset of fMRI brain activity using handwritten
 262 letter images as stimuli [25], which is fMRI imaging of hu-
 263 mans containing 360 gray-scale handwritten character images.
 264 It has equal number of character B, R, A, I, N, S. The original
 265 image resolution is 56×56 and the corresponding fMRI
 266 signals contain voxels (each fMRI character pattern has 2420
 267 voxels) from V1 and V2 areas of all three subjects S1, S2 and
 268 S3.

269 We also test our model on sound signals. We choose 0-9
 digits of T1-46 speech corpus [27] with the audio samples

270 read by 16 speakers for the 10 digits as in MNIST with 4136
 271 audio samples totally. This sound-image dataset is divided into
 272 4000 for training and 136 for testing. During the training
 273 process, the pairs of audio-image are used as the training
 274 samples simultaneously which are the same digital samples
 275 in noise image-image datasets and fMRI-image datasets. We
 276 used Auditory toolbox [28] for pre-processing the data, such
 277 that all of the audio samples are converted as the spectrograms
 278 with 1500 time steps and 39 frequencies, then we can get the
 279 a $58,500 \times 1$ vector (1500×39) for each sample.

280 Although these signals have different dimensionality, we
 281 adjusted their sizes and the number of encoding neurons
 282 according to the computational ability of the machine. In
 283 our cases, the experiments were conducted on a workstation
 284 equipped with two-processor Intel(R) Xeon(R) Core CPU and
 285 one NVidia GeForce GTX 2080Ti GPU. The operating system
 286 is Ubuntu 16.04. Tensorflow [29] and Keras [30] were used
 287 for implementing our model.

288 D. Performance evaluation

289 We choose three different evaluating metrics to evaluate
 290 the performance on the proposed DSPD and other compared
 291 methods.

292 1) Mean Square Error (MSE): MSE represents the final
 293 expectation of the squared error between the desired and
 294 original values. A detailed description of the MSE about the
 295 pair of patterns $\langle \mathbf{X}_1, \mathbf{X}_2 \rangle$, with the resolution of $H \times W$ is as
 296 follow:

$$MSE = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W ((\mathbf{X}_1(i, j) - \mathbf{X}_2(i, j))^2), \quad (1)$$

297 Generally, lower MSE value means better pattern quality.

298 2) Structural Similarity Index Metric (SSIM): SSIM is
 299 used for evaluating the structure comparison between two
 300 patterns. [31] thought this kind of metric with the assumption
 301 that human visual processing system can perceive the pattern
 302 including its variations and distortion through extracting the
 303 structural information changes.

304 Based on the luminance (l), contrast (c) and structure (s) of
 305 two patterns x and y .

$$SSIM(x, y) = [l(x, y)^\alpha \cdot c(x, y)^\beta \cdot s(x, y)^\gamma] \quad (2)$$

306 When the α, β and γ equal to 1, we can get the SSIM function
 307 which I used in this paper as shown in equation (3).

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (3)$$

308 SSIM could be used for describing the the positive relation
 309 with the pattern quality between the original and reconstructed
 310 patterns. In order to show more detailed performance, we also
 311 introduce another pattern quality metric named Peak Signal to
 312 Noise Ratio (PSNR).

313 3) Peak Signal-to-Noise Ratio (PSNR): Given a clean pat-
 314 tern \mathbf{I}_1 and the reconstructed pattern \mathbf{I}_2 with size $M \times N$ we

can get the MSE as the same in equation (1), we can get the
 $PSNR$ as shown in equation 4:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (4)$$

MAX_I^2 is the max value in whole pixel range. For instance,
 if we used uint8 to represent an image, MAX_I^2 should be 255
 ($2^8 - 1$).

III. RESULTS

A. One framework for multiple tasks

Our proposed model is a framework with a mixture of a
 biological encoding part and a DNN based decoding part as
 illustrated in Figure 1. The encoding part is similar to an
 neural pathway of the sensory systems, which receive sensory
 information in the format of images, sound waves, or other
 types of artificial sensor data represented spatial, temporal,
 or spatiotemporal patterns. The output of the encoder is a
 sequence of spikes similar to biological neurons in response
 to stimuli. After encoding, the encoded information will be
 delivered to the decoding part. Depending on practical tasks,
 the different decoders can be built for signal reconstruction,
 object recognition, semantic classification, etc. One can decode
 the spikes directly with spiking neural networks as decoder.
 Or one can also convert spikes into different format of data,
 for example, image pixels, to take advantage of the state-of-
 the-art computer vision techniques. The benefit of transcoding
 sensory information with neural spikes is to utilize the core
 concept of neuromorphic computing, e.g., energy and data
 efficient computing without loss of any information. Thus,
 our proposed framework is a unified spike transcoding system
 functioning as data compression, feature extraction, temporal
 encoding and decoding.

In this study, we put our proposed framework into the
 context of signal reconstruction in terms of image pixels.
 However, it is noted that our framework is fixable to account
 for other purposes, so that the exact architectures of the
 encoder and decoder are fixable to adapt to be other types of
 neural networks, or simple traditional statistical methods. To
 reconstruct signals, we need to upsample the encoded spikes
 into the remapping image space with the same size of signals,
 4096 in our cases. According to the central limit theorem,
 these remapping images are following a Gaussian distribution.
 The intuition is that if one adds up all of different types of
 images through each detailed pixel, we would get a white-
 noise picture. In this sense, these remapping images are the
 reservoir of input information and crucial for reconstructing
 the final output signals to match the input signals.

As shown in Figure 1, we evaluate our model on various
 different datasets for different tasks.

- MNIST data [23], where there are 10 digital images,
 is used to demonstrate the feasibility of our model for
 transcoding with neural spikes.
- MNIST with random noise [24], where each digital image
 is embedded with a certain level of noise. Furthermore,
 we also used data with different levels of noise to test the
 model behavior, e.g. varied Gaussian noise with different
 noise intensities.

- MNIST with background images [24], where each digital image is embedded with a background natural image. A random patch from a white and black was used as the background. Those patches were extracted randomly from a set of pictures downloaded online.
- CIFAR10[32] is a RGB based dataset which consists of 50,000 training images and 10,000 test images in 10 classes, the image size is 32×32 . It has natural images with complex patterns and objects which was used by the proposed DSPD to show its reconstruction ability. The same as Gaussian MNIST, we also used data with different levels of Gaussian noise to test the model denoise behavior.
- fMRI brain activity under viewing handwritten images [25], where the dataset consists of fMRI signals viewing the letters of B, R, A, I, N, S.
- Sound signals of 10 spoken letter datasets [26], where different people read 10 digits of MNIST. The dataset includes audio-image pairs which were used to build the relationship between audio waves and images.

B. Signal Reconstruction and Denoising

In order to show the capability of the proposed DSPD for signal reconstruction, we use visual images regarding to mimic the static image reconstruction as one of the most important functions in biological visual processing system. We applied DSPD on five static image datasets which are divided into two categories: pure dataset MNIST and noisy datasets random-MNIST (with random noise), background-MNIST (with background images), rotation-MNIST (rotated digital) and rotation-background-MNIST (rotated digital with background images) as show in Fig. 3. The dataset is divided into two parts: training set (50,000 training samples) and test set (10,000 test samples) for MNIST and its variation. Different from other reconstruction models [18] [33] which only focus on image without any other noise, DSPD have strong generation ability in noisy environment caused by random (rand), background (bg), rotation (rot) and background-rotation (bg-rot).

In order to further explore the model's generalization ability in noisy environment, we divide the sizes of the training set and test set to verify that the DSPD can achieve better performance on small-size datasets than any other models. For examples, when the training samples are 90 and test samples are 10 means, we choose 90 training samples from the whole 50,000 training samples randomly and they are uniformly distributed in 0-9 ten classes.

As shown in Fig. 3, we choose standard MNIST and its four variations to show the noise immunity of DSPD, these four noisy MNIST datasets have random, background, rotation and rotation-background noise respectively. The first two rows in Fig.2 represent the qualitative evaluations showing that the DSPD have strong denoising ability when it deals with the random-MNIST and background-MNIST, the reconstructed images from random and background MNIST appear clear without noise. However, when the datasets have rotated objects, DSPD cannot reconstruct meaningful images. Presumably, because rotation is symmetrical in all directions,

that break the unity of directionality in digital images, for instances, if a handwritten image 6 is rotated more than 90 degree or even 180 degree, then it becomes some wrong types such as 9, which can not be discriminated by the model.

In order to further demonstrating that the strong rotation is more symmetrical, we used t-SNE [34] to visualize the structure of sample population represented by images after upsampling spikes (Fig. 3). From Fig. 3, one can see that when t-SNE is applied on clean MNIST images, the 0-9 ten classes could be splitted better when rot (rotation) MNIST. As shown in Fig. 3, the encoded patterns from rotation MNIST are mixed together so that them can not be separated well. Although the patterns all look like white-noise, they are significantly different. From the encoding point of view, this could also explain the meaning about the patterns after encoding and give the reason why the reconstructed images from rotation and rotation-background MNIST look like zeros in the last two rows in Fig. 3.

Not only limited by the quality evaluations on visualization, we also make some more detailed quantitative evaluations. Table I. To show the advantage of spike transcoding,, we implement and compare our DSPD with another recent state-of-the-art method termed deep generative multi-view model (DGMM) [35]. DGMM is designed in the context of fMRI decoding, here we test it for signal reconstruction. As DGMM is designed for reconstructing small size datasets, in order to compare the reconstruction performance with DSPD, we extract a small subset from whole dataset as using 90 images for training and 10 images for rebuilding. And the MNIST and its four variations are not uniformly distributed in 50,000 training samples and 10,000 test samples, in order to avoid to the imbalanced training problem, we choose 40,000 and 8000 equally distributed training samples and 8000 test samples as the maximum experimental condition. From table I, we can see that DSPD perform better than DGMM when in small size 90 training samples and 10 test samples on MSE, SSIM and PSNR. DSPD reaches a PSNR peak at 13.11 when reconstructing from random MNIST. If the training and test samples from small size dataset (90/10) move to large size dataset (40,000/8000), these performance evaluation metrics of DSPD on random and background MNIST are better than these evaluated on 90 training and 10 test. On the whole, there is no huge performance gap on random (MSE: 0.032 SSIM: 0.52 PSNR: 14.72), background (MSE: 0.048 SSIM: 0.421 PSNR: 13.77). This is thought to be due to the increasing training samples from random and background MNIST could help train the framework and improve the decoding performance.

We then further test the model with different levels of noise. Based on the clean MNIST images, we added Gaussian noise with increasing levels of noise by varying the parameter of σ . As shown in Fig. 2 left, we varied the degree of σ from 0 (clean) to 0.1 (strong noise). With the increasing of noise level, the images look like more fuzzy. With those noise MNIST images as input, the proposed DSPD could reconstruct the pictures as shown in Fig. 2 right. One can observe that the proposed framework could rebuild the pattern successfully and the reconstructed samples could denoise very well with different level of noise, except the strong noise ($\sigma = 0.1$),

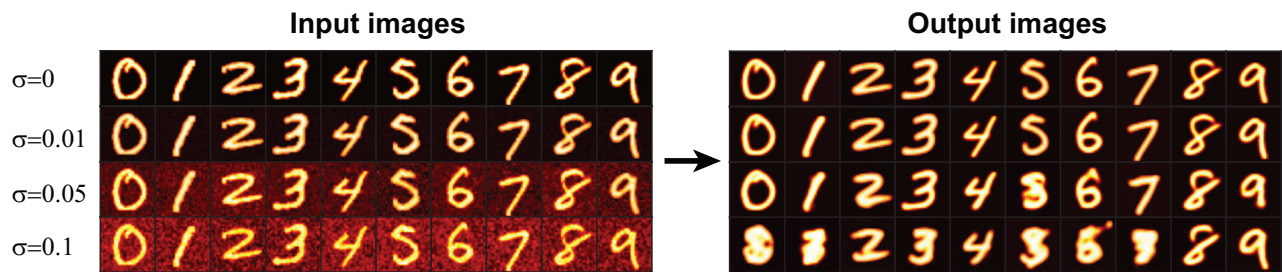


Fig. 2: Reconstructed images from noisy MNIST.

484 which is similar in top right corner of Fig. 3. Although the
 485 reconstructed samples with strong noise is not visually perfect
 486 as those from light noise, we can also recognize the digit shape
 487 easily.

488 The proposed DSPD could not only reconstruct high quality
 489 from noisy handwritten digits, but also get good reconstruction
 490 performance from noisy natural image-complexity dataset,
 491 here we adopted CIFAR10 as experimental dataset.

492 As shown in figure 4, with different levels of Gaussian
 493 noise (from $\sigma = 0$ to $\sigma = 0.1$), the proposed DSPD
 494 could reconstruct images from noisy CIFAR10 dataset. The
 495 proposed DSPD was trained on 50,000 images and rebuilt
 496 from 10,000 test samples. Different from MNIST digits, the
 497 proposed model could reconstruct similar quality figures with
 498 both clean noise or strong noise visually. This also means
 499 more natural images with higher complexity have strong anti-
 500 noise ability. One possible reason is that natural images with
 501 complex patterns contain more information including color,
 502 texture and shape, while digits are much more simple. So from
 503 Figure 4, the proposed DSPD show its strong anti-noise ability
 504 in real-life natural environments.

505 C. Reconstruction of fMRI Signals

506 The presented DSPD framework could not only reconstruct
 507 high-quality images and show strong noise immunity, but
 508 also perform well on object recognition from fMRI signals.
 509 We used a fMRI dataset with the stimuli as handwritten
 510 letter images for testing the model. In order to show the
 511 reconstruction ability of DSPD, we also compared our DSPD
 512 with the DGMM [35]. Visually we observe that proposed
 513 DSPD can rebuild better quality patterns compared the results
 514 from DGMM.

515 Fig. 5 represented the reconstructed samples produced by
 516 DSPD and DGMM. Fig. 5 left are reconstructed patterns of
 517 DSPD and DGMM with 90 training samples and 10 recon-
 518 structing samples. We can observe that the proposed DSPD
 519 show more clear reconstructed samples compared to the results
 520 from DGMM. And there is a similar conclusion no matter on
 521 subjects S_1 , S_2 and S_3 , or brain areas V_1 and V_2 , when the
 522 training samples increased to 300 and reconstructing samples
 523 are 60 as shown in Fig. 5 right. Compared to the results from
 524 DSPD, DGMM generates more blurry reconstructed images.

525 Table II shows more detailed performance quantitative eval-
 526 uation on fMRI Handwritten characters dataset of DSPD and
 527 DGMM. As mentioned before, this fMRI based character
 528 dataset has three subjects S_1 , S_2 and S_3 from V_1 and V_2

529 of human retinal systems. Here we used 300 image-fMRI
 530 pairs for training and 60 for reconstructing. As shown in
 531 table II, in subject 1 (S_1), the proposed DSPD could perform
 532 better than the DGMM on MSE, SSIM and PSNR. As for S_2 ,
 533 DGMM could get better reconstruction performance on MSE
 534 (0.059) and PSNR (13.02) in character patterns from V_2 areas,
 535 DSPD achieve the best performance on SSIM (0.45). When
 536 we observe the performance evaluation metrics located on S_3 ,
 537 except DGMM has the best PSNR (12.508) in V_1 areas, the
 538 proposed DSPD nearly behave better than DGMM on MSE
 539 and SSIM no matter in V_1 and V_2 areas. In short, the proposed
 540 DSPD behave better in most cases, but that is not a big
 541 difference. So, from the quality and quantitative evaluation of
 542 DSPD and DGMM, we can conclude that the proposed DSPD
 543 achieve better reconstruction performance on fMRI character
 544 datasets.

545 D. Decoding Sound Signal

546 In order to further explore the potential of our model frame-
 547 work, we apply it on a sound dataset with audio waveform
 548 by different human subjects reading 10 digits of MNIST. As
 549 shown in Fig. 6, the same as used in [26], we choose 0-9
 550 digits as the audio samples and standard MNIST for images
 551 (see Methods). For a single digit, the samples are collected
 552 from different human subjects reading it for audio data and
 553 writing it for MNIST image data. There are different mappings
 554 between audio digits and image digits. To induce noise and
 555 show the generalization of audio data, we designed two types
 556 of audio-image pairing dataset as shown in Fig. 6. Fig. 6 A
 557 is the dataset A, in which we choose different image samples
 558 for different audio samples in the the sample class as one
 559 image-per audio. Whileas, in dataset B, we use the same image
 560 samples to represent the same class of audio samples, which
 561 means the images in one class are the same for differnt audio
 562 samples.

563 For sound-image dataset A (one image-per audio) and
 564 dataset B (one image-per class), we choose a subset about
 565 90 training samples and 10 test samples to show the recon-
 566 struction performance as shown in Fig. 7A and B. And for
 567 a further comparison, we divide the full size (4136 samples)
 568 as 4000 training samples and 136 test samples respectively,
 569 the selected reconstructed samples are presented in Fig. 7C
 570 and D. We can observe that compared to the generated from
 571 dataset B, dataset A generates more blurry images which
 572 indicate the reconstructed samples from dataset A could learn
 573 the underlying shape, structure and texture of the presented

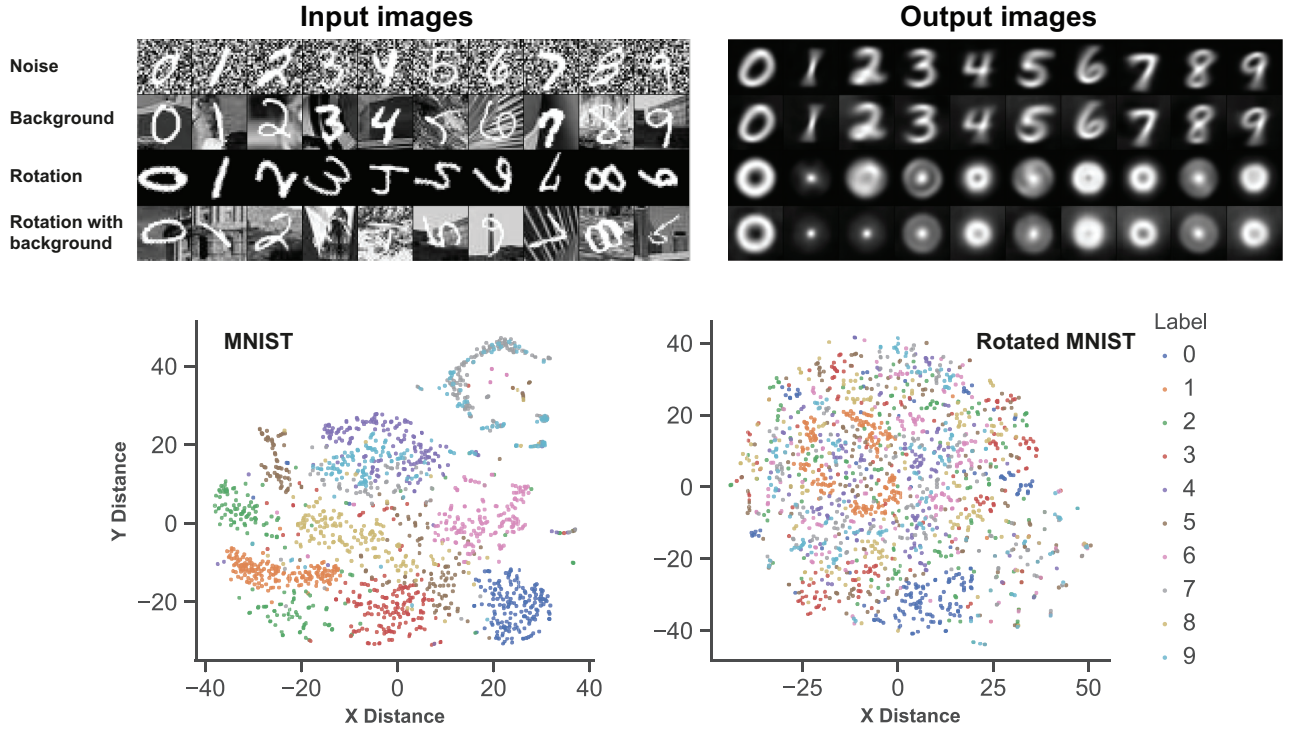


Fig. 3: Reconstructed images from different versions of MNIST. Different t-SNE visualization images between clean and rotated MNIST based spatio-temporal patterns after encoding.

TABLE I: Comparison of noise immunity between DSPD and DGMM on MNIST and its variations.

Model	Random			Background			Rotation			Bg-rotation		
	MSE	SSIM	PSNR	MSE	SSIM	PSNR	MSE	SSIM	PSNR	MSE	SSIM	PSNR
DSPD (90/10)	0.049	0.15	13.11	0.056	0.381	12.90	0.072	0.417	11.67	0.087	0.290	10.99
DGMM (90/10)	0.062	0.36	12.02	0.080	0.358	11.33	0.124	0.243	9.39	0.090	0.288	10.59
DSPD (40K/8K)	0.032	0.52	14.72	0.048	0.421	13.77	0.068	0.489	11.77	0.092	0.276	10.58

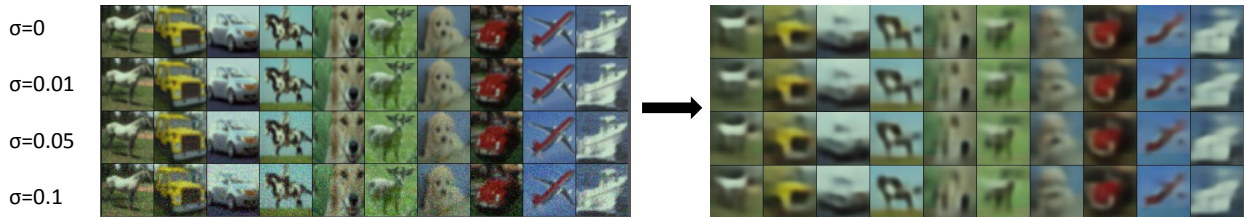


Fig. 4: Reconstructed images from noisy CIFAR10.

574 images, but they could not learn finer details. Although the
 575 images in dataset A are various, the proposed DSPD may learn
 576 some more different basic information such as shape, texture
 577 and structure and extract the common information among them
 578 all, the proposed model could be trained over multiple same
 579 samples of the same class, which is more easier and helpful
 580 for a network model.

581 IV. DISCUSSION

582 In this paper, we proposed a robust cross-multimodal pattern
 583 reconstruction model named deep spike-to-pattern decoder
 584 (DSPD). This cognitive model combines neural encoding and
 585 DNN based decoding parts in a same framework, with the
 586 help of neural encoding method, this biological plausible

587 reconstruction model can encode the outside stimuli to spatio-temporal patterns. Based on these kinds of advantages, the
 588 proposed DSPD has strong generalization ability and become
 589 robust in noisy environment. Furthermore, it is the first attempt
 590 to encode various kinds of stimuli: image, fMRI and sound in a
 591 uniform framework. We show the reconstruction performance
 592 of the presented DSPD applied on MNIST, variational MNIST,
 593 fMRI-digits datasets, fMRI-characters datasets, sound-image
 594 dataset A and dataset B is comparable to some other state-of-
 595 the-art reconstruction models. We argue the encoding method
 596 and decoding structure adopted by DSPD could help to extract
 597 more important features and lead to train a more robust and
 598 efficient cognitive reconstruction model. In the future, we will
 599 adopt more types of external stimuli such as ECoG, EEG and
 600

TABLE II: Evaluation of neural decoding performance of DGMM and proposed DSPD on fMRI character dataset with three subjects S_1 , S_2 and S_3 from v_1 and v_2 areas.

Models	Character fMRI-S1			Character fMRI-S2			Character fMRI-S3		
	MSE	SSIM	PSNR	MSE	SSIM	PSNR	MSE	SSIM	PSNR
DGMM-V1	0.068	0.212	11.87	0.060	0.266	12.79	0.069	0.27	12.508
DSPD-V1	0.063	0.427	12.46	0.067	0.43	12.38	0.064	0.46	12.35
DGMM-V2	0.071	0.210	11.83	0.059	0.27	13.02	0.079	0.29	11.95
DSPD-V2	0.061	0.442	12.44	0.063	0.45	12.79	0.063	0.47	12.506

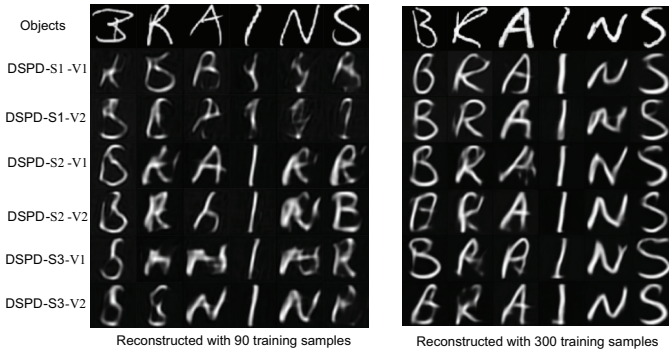


Fig. 5: Presented fMRI characters and Reconstructed Results of DSPD three subjects S_1 , S_2 and S_3 from the V_1 and V_2 areas (the left images are with 90 training samples and the right images are with 300 training samples).

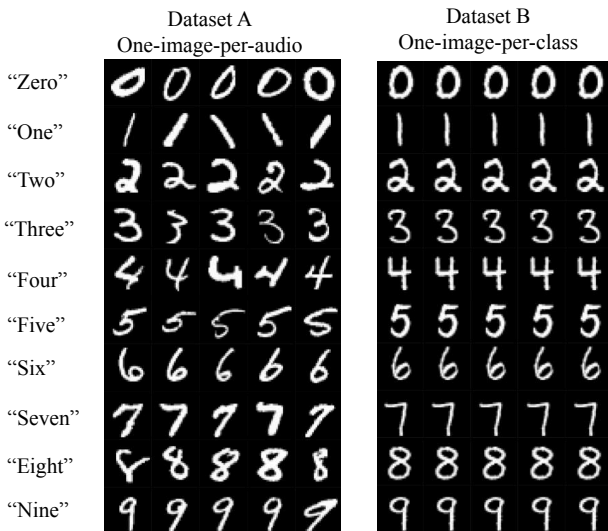


Fig. 6: Two Types of Sound Datasets. Dataset A means one image corresponds one paired audio sample, Dataset B means one image corresponds one audio class.

601 etc.

602 Because of the event driven nature of the spiking activities,
 603 it would be beneficial for implementations of neuromorphic
 604 hardware chips with aid of its structure. Furthermore, this work
 605 proposes a more biological realistic reconstruction framework
 606 which can achieve nearly real-time encoding and decoding
 607 various patterns by neural spikes. The potential showed by
 608 DSPD is promising with the hope that this cognitive model
 609 could help us how mammalian neocortex and neural circuits
 610 are performing computations in high-level visual tasks.

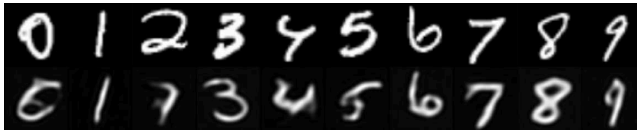
A. Neural Encoding and Decoding

611 How information is represented in the brain still remains
 612 unclear, but this leads to one of the core problems in neural
 613 processing system. However, there is strong evidence [36],
 614 [20] to believe that spike trains are an optimal way for
 615 transmission and information representation. Unlike neurons
 616 in traditional convolutional neural networks (CNNs), which
 617 communicate via real values, neurons in computational sys-
 618 tems such as spiking neural network (SNN) communicate
 619 via spikes. Spiking based systems have been shown to be
 620 more computationally powerful than traditional artificial neural
 621 networks (ANNs), including CNNs. Moreover, these systems
 622 are event-driven, computation in synapses and neurons are
 623 triggered by incoming spikes. Driven by sparse spike trains,
 624 most synapses and neurons in neural circuits are idle for
 625 most of the time, which allows those spiking based models
 626 to run inference with low computational cost and low power.
 627 They are advantageous to deal with spatio-temporal patterns,
 628 through spike-based learning and memory mechanisms [37].
 629

630 However, compared with deep CNNs, typical artificial spik-
 631 ing systems are surely at a great disadvantage about feature
 632 extraction because of shallow structures with few biologically
 633 based neurons. The difficulty for building a deep biological
 634 coding system lies on the complex neural dynamics, shallow
 635 layer cannot detect and capture some deeper and hidden
 636 information. [38] and [39] explored the visual system using
 637 the hierarchical simple cell and complex cell feedforward
 638 model, and showed that there is a high resemblance of the
 639 feature extraction process between the model and biological
 640 brain. Nevertheless, the previous work [38] does not model the
 641 coding flow in a biological realism way, i.e., relying on a non-
 642 biological classifier such as support vector machine. Aiming at
 643 this issue, CSNN [16] proposes a brain-inspired spiking based
 644 coding framework, which consists of a partial CNN and a
 645 SNN. CSNN is able to exploit the powerful feature extraction
 646 ability of the CNN to increase the coding performance of the
 647 computational neural system.

648 There still exist big challenges about constructing robust
 649 coding system which is believed to originate from the invari-
 650 ant representation of cross-multimodal features. In biological
 651 coding processing, the information which is received from the
 652 outside and communicate between the neurons is discrete.
 653 Before run-time, every real value of the outside image is
 654 encoded into spike trains by the feat of encoding methods,
 655 then the spikes are communicated between the corresponding
 656 neurons of the networks. The existed encoding rules can be
 657 classified into rate based coding, temporal based coding and
 658 others.

659 The rate based coding [40] is used to encode images into



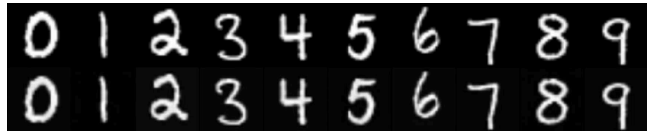
A. Image synthesized from audio-image dataset A with 90 training samples and 10 test samples.



B. Image synthesized from audio-image dataset B with 90 training samples and 10 test samples.



C. Image synthesized from audio-image dataset A with 4000 training samples and 136 test samples.



D. Image synthesized from audio-image dataset B with 4000 training samples and 136 test samples.

Fig. 7: Image synthesized from Dataset A (one image-per audio) and Dataset B (one image-per class) with small size training samples (90) and full size training samples (4000). Images in first line are the presented samples and figures in second line are reconstructed results.

660 dense spikes, a higher firing rate is defined as high sensory
 661 variable which can be represented as the average number of
 662 spikes counting within a temporal encoding window. The rate
 663 based coding always uses dense spikes (Poisson spike trains)
 664 to represent the neurons firing rate. To encode a real value,
 665 rate coding tends to generate many spikes, especially if the
 666 real value is large, which imposes high computational load on
 667 downstream spiking neurons. [41] proposes a novel algorithm
 668 which adopted filtered spike train as transition from original
 669 images. The sparse coding [42] clusters a relatively small
 670 subset of neurons which have nearly the same firing rate.

671 Although these rate based coding mechanisms are to some
 672 extent successful, the power consumption of the whole system
 673 is large. The precision of the encoded value increases with the
 674 time span of the spike train, which is roughly proportional to
 675 the number of spikes in the spike train. In addition, given
 676 the time span of the spike train, the number of spikes in the
 677 spike train is roughly proportional to the encoded value [43].
 678 Therefore, with rate coding, many spikes have to be generated
 679 to encode a large value with high precision, which imposes a
 680 high computational load on downstream neurons. On the other
 681 hand, to generate a spike train, spikes have to be generated
 682 with different spike times. With rate coding, spike times of
 683 individual spikes are not used to convey information at all.

684 Furthermore, studies [44], [45] have proved that neurons
 685 in human retina firing more likely as temporal coding mech-
 686 anism compared to rate based coding ways [20]. Patterns
 687 encoded from temporal coding can carry more information
 688 in spatiotemporal spikes and consume fewer computational
 689 resources than rate based coding. So based on the advantages
 690 lying in temporal encoding, this paper adopts a biological
 691 temporal encoding methods as the primary encoding layer.

692 Compared with the spiking neuron models such as IF, LIF,
 693 Adex, Izhikevich in SNN or Aurel Lazar's Time Encoding
 694 Machines[46], our model is not a spike-in spike-out model.
 695 We only consider the question of reconstructing visual stimuli
 696 from neuron responses, i.e. decoding is an essential part in this
 697 study. Here we propose a decoding model that reconstructs
 698 natural scenes directly from neural signals. Different from
 699 HTM[47] (hierarchical temporal memory) which focuses on
 700 time-coherent information in analysis of brain's model, we

expect that our decoder will help to solve some problems on
 neural decoding (e.g. what characters of spikes are important
 for neural coding), and provide some clues on the questions
 of brain-machine interface, such as neural neuroprosthesis.

Some recent work[48], [49], [50] have encoded dynamic
 video scenes, speech and biomedical signals with DVS (Dy-
 namic Vision Sensors) or other Neuromorphic hardware chips
 successful. Our proposed model is so far implemented on
 Ubuntu software system, in the future, we will take DVS
 sensors as one of the beginning of sensory information
 acquisition equipment and implement the DSPD model on
 our designed Darwin[51] Neuromorphic hardware system to
 achieve a software-hardware integrated spiking recognition
 framework for artificial machine vision.

B. Multimodal Pattern Reconstruction

There has already been various studies for how to con-
 struct the visual pattern reconstruction systems. Typical vi-
 sual reconstruction aim at reconstructing the original stimuli
 by using the neural response, for instances, rebuilding the
 visual scenes which the animals saw before through ob-
 taining each pixel of those scenes from the neural signals
 produced by visual system, including neural spikes and fMRI
 activity [18] [52] [53]. [54] proposed a Bayesian canonical
 correlation analysis model to build a bridge between visual
 scenes and the corresponding brain activities, however due to
 the limitation of simple linear shallow framework, it cannot get
 some complex features. [18] [55] constructed the rebuilding
 systems with the aid of deep neural networks, compared to
 traditional simple mapping methods, these models could obtain
 more meaningful and complex features, thus leading to better
 performance. [56] combined the probabilistic inference with
 the generative adversarial networks and applied it into a face
 image - evoked brain activities, which usually cannot converge
 to the global optimum with the constrain of a n equilibrium
 between the generator and discriminator [57].

Although the aforementioned work greatly promote the
 research in the area of pattern reconstruction, accurately recon-
 structing the cross-multimodal still remains challenging from
 two main aspects: 1. Those models are short of more biological
 coding activities such as spikes encoding and decoding from

with neural coding method, since the spikes generated with neural coding are the unique output neurons of retinas. They only focused on one or two modals pattern reconstruction tasks such as fMRI and images, cross-multimodal pattern rebuilding is necessary and pivotal for understanding how neural representation in biological neural system. In order to address these limitations, this paper proposed a cross multi-modal pattern reconstruction with hierarchical structures from spiking activities, named deep spike-to-pattern decoder (DSPD). Recent advances in experimental techniques enables us to record neural signals from multiple brain areas simultaneously [58]. Thus, our proposed decoding approach make it possible to decoding of multimodal information from neural signals of multiple brain areas with one single decoding framework. We expect that the method presented here will advance the methodology of analyzing neural spikes, as well as the applicability of neuromorphic computing.

REFERENCES

- [1] M. Jazayeri and J. A. Movshon, "Optimal representation of sensory information by neural populations," *Nature neuroscience*, vol. 9, no. 5, pp. 690–696, 2006.
- [2] 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, apr 2020.
- [3] M. C.-K. Wu, S. V. David, and J. L. Gallant, "Complete functional characterization of sensory neurons by system identification," *Annu. Rev. Neurosci.*, vol. 29, pp. 477–505, 2006.
- [4] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.
- [5] J. K. Liu and D. V. Buonomano, "Embedding multiple trajectories in simulated recurrent neural networks in a self-organizing manner," *Journal of Neuroscience*, vol. 29, no. 42, pp. 13 172–81, 2009.
- [6] H. Tang, K. C. Tan, and Z. Yi, *Neural networks: computational models and applications*. Springer Science & Business Media, 2007, vol. 53.
- [7] Q. Xu, J. Peng, J. Shen, H. Tang, and G. Pan, "Deep CovDenseSNN: A hierarchical event-driven dynamic framework with spiking neurons in noisy environment," *Neural Networks*, vol. 121, pp. 512–519, 2020.
- [8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [9] 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.
- [10] T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani, "Neural decoding of visual imagery during sleep," *Science*, vol. 340, no. 6132, pp. 639–642, 2013.
- [11] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, pp. 352–355, 2008.
- [12] 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.
- [13] D. L. K. Yamins and J. J. Dicarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, no. 3, p. 356, 2016.
- [14] 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*, pp. 1–12, 2020.
- [15] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS Comput Biol*, vol. 10, no. 12, p. e1003963, 2014.
- [16] Q. Xu, Y. Qi, H. Yu, J. Shen, H. Tang, and G. Pan, "CSNN: An Augmented Spiking based Framework with Perceptron-Inception." in *IJCAI*, 2018, pp. 1646–1652.
- [17] 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.
- [18] 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.
- [19] 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, pp. 1–14, 2017.
- [20] A. Onken, J. K. Liu, 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, nov 2016.
- [21] R. L. Jenison, J. W. Schnupp, R. A. Reale, and J. F. Brugge, "Auditory space-time receptive field dynamics revealed by spherical white-noise analysis," *Journal of Neuroscience*, vol. 21, no. 12, pp. 4408–4415, 2001.
- [22] W. J. Speechley, J. L. Hogsden, and H. C. Dringenberg, "Continuous white noise exposure during and after auditory critical period differentially alters bidirectional thalamocortical plasticity in rat auditory cortex in vivo," *European Journal of Neuroscience*, vol. 26, no. 9, pp. 2576–2584, 2007.
- [23] Y. LeCun, C. Cortes, and C. J. Burges, "The MNIST database of handwritten digits, 1998," *URL http://yann.lecun.com/exdb/mnist*, vol. 10, p. 34, 1998.
- [24] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 473–480.
- [25] S. Schoenmakers, M. Barth, T. Heskes, and M. Van Gerven, "Linear reconstruction of perceived images from human brain activity," *NeuroImage*, vol. 83, pp. 951–961, 2013.
- [26] D. Roy, P. Panda, and K. Roy, "Synthesizing Images From Spatio-Temporal Representations Using Spike-Based Backpropagation," *Frontiers in neuroscience*, vol. 13, p. 621, 2019.
- [27] M. Liberman, R. Amsler, K. Church, E. Fox, C. Hafner, J. Klavans, M. Marcus, B. Mercer, J. Pedersen, P. Roossin *et al.*, "Ti 46-word," *Philadelphia (Pennsylvania): Linguistic Data Consortium*, 1993.
- [28] M. Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep*, vol. 10, no. 1998, 1998.
- [29] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.
- [30] A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017.
- [31] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1488–1499, 2011.
- [32] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Handbook of Systemic Autoimmune Diseases*, vol. 1, no. 4, 2009.
- [33] Y.-T. Lin and G. D. Finlayson, "Physically Plausible Spectral Reconstruction from RGB Images," *arXiv preprint arXiv:2001.00558*, 2020.
- [34] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [35] 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.
- [36] C. P. Hung, G. Kreiman, T. Poggio, and J. J. Dicarlo, "Fast Readout of Object Identity from Macaque Inferior Temporal Cortex," *Science*, vol. 310, no. 5749, pp. 863–866, 2005.
- [37] J. Hu, H. Tang, K. C. Tan, and H. Li, "How the Brain Formulates Memory: A Spatio-Temporal Model Research Frontier," *IEEE Computational Intelligence Magazine*, vol. 11, no. 2, pp. 56–68, 2016.
- [38] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 15, pp. 6424–6429, 2007.
- [39] T. Serre and T. Poggio, "A Neuromorphic Approach to Computer Vision," *Communications of the Acm*, vol. 53, no. 10, pp. 54–61, 2010.
- [40] O. Peter, N. Daniel, S. C. Liu, D. Tobi, and P. Michael, "Real-time classification and sensor fusion with a spiking deep belief network," *Frontiers in Neuroscience*, vol. 7, p. 178, 2013.
- [41] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. S. Modha, "A digital neurosynaptic core using embedded crossbar memory with 45pj per spike in 45nm," in *Custom Integrated Circuits Conference*, 2011, pp. 1–4.

- 887 [42] L. Perrinet, M. Samuelides, and S. Thorpe, "Sparse spike coding in an
888 asynchronous feed-forward multi-layer neural network using matching
889 pursuit," *Neurocomputing*, vol. 57, pp. 125–134, 2004.
- 890 [43] M. Wang, X. Liao, R. Li, S. Liang, R. Ding, J. Li, J. Zhang, W. He,
891 K. Liu, J. Pan, Z. Zhao, T. Li, K. Zhang, X. Li, J. Lyu, Z. Zhou, Z. Varga,
892 Y. Mi, Y. Zhou, J. Yan, S. Zeng, J. K. Liu, A. Konnerth, I. Nelken, H. Jia,
893 and X. Chen, "Single-neuron representation of learned complex sounds
894 in the auditory cortex," *Nature Communications*, vol. 11, no. 1, aug
895 2020.
- 896 [44] M. J. Berry and M. Meister, "Refractoriness and neural precision."
897 *Journal of Neuroscience*, vol. 18, no. 6, p. 2200, 1998.
- 898 [45] V. J. Uzzell and E. J. Chichilnisky, "Precision of spike trains in primate
899 retinal ganglion cells." *Journal of Neurophysiology*, vol. 92, no. 2, pp.
900 780–9, 2004.
- 901 [46] A. A. Lazar and Y. Zhou, "Reconstructing natural visual scenes from
902 spike times," *Proceedings of the IEEE*, vol. 102, no. 10, pp. 1500–1519,
903 2014.
- 904 [47] J. Hawkins and D. George, "Hierarchical temporal memory," *Alphascript
905 Publishing*, vol. suppl, no. 5, pp. 1 – 10, 2011.
- 906 [48] M. Yang, S. C. Liu, and T. Delbruck, "Comparison of spike encoding
907 schemes in asynchronous vision sensors: Modeling and design," in *IEEE
908 International Symposium on Circuits & Systems*, 2014.
- 909 [49] Minhao, Yang, Shih-Chii, Liu, Tobi, and Delbruck, "Analysis of encod-
910 ing degradation in spiking sensors due to spike delay variation," *IEEE
911 Transactions on Circuits & Systems I Regular Papers*, 2017.
- 912 [50] F. Corradi, C. Elias-Smith, and G. Indiveri, "Mapping arbitrary mathemat-
913 ical functions and dynamical systems to neuromorphic vlsi circuits for
914 spike-based neural computation," in *IEEE International Symposium on
915 Circuits & Systems*, 2014.
- 916 [51] D. Ma, J. Shen, Z. Gu, M. Zhang, X. Zhu, X. Xu, Q. Xu, Y. Shen,
917 and G. Pan, "Darwin: A neuromorphic hardware co-processor based on
918 spiking neural networks," *Journal of Systems Architecture*, vol. 77, pp.
919 43–51, 2017.
- 920 [52] J.-D. Haynes and G. Rees, "Predicting the orientation of invisible stimuli
921 from activity in human primary visual cortex," *Nature neuroscience*,
922 vol. 8, no. 5, pp. 686–691, 2005.
- 923 [53] E. Chong, A. M. Familiar, and W. M. Shim, "Reconstructing represen-
924 tations of dynamic visual objects in early visual cortex," *Proceedings
925 of the National Academy of Sciences*, vol. 113, no. 5, pp. 1453–1458,
926 2016.
- 927 [54] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani, "Modular encoding and
928 decoding models derived from bayesian canonical correlation analysis,"
929 *Neural computation*, vol. 25, no. 4, pp. 979–1005, 2013.
- 930 [55] Y. Rivenson, Y. Zhang, H. Günaydin, D. Teng, and A. Ozcan, "Phase
931 recovery and holographic image reconstruction using deep learning in
932 neural networks," *Light: Science & Applications*, vol. 7, no. 2, pp.
933 17 141–17 141, 2018.
- 934 [56] Y. Güçlütürk, U. Güçlü, K. Seeliger, S. Bosch, R. van Lier, and M. A.
935 van Gerven, in *Advances in Neural Information Processing Systems*,
936 2017, pp. 4246–4257.
- 937 [57] S. Martin Arjovsky and L. Bottou, "Wasserstein Generative Adversarial
938 Networks," in *Proceedings of the 34 th International Conference on
939 Machine Learning, Sydney, Australia*, 2017.
- 940 [58] M. Yang, Z. Zhou, J. Zhang, S. Jia, T. Li, J. Guan, X. Liao, B. Leng,
941 J. Lyu, K. Zhang, M. Li, Y. Gong, Z. Zhu, J. Yan, Y. Zhou, J. K.
942 Liu, Z. Varga, A. Konnerth, Y. Tang, J. Gao, X. Chen, and H. Jia,
943 "MATRIEX imaging: multiarea two-photon real-time in vivo explorer,"
944 *Light: Science & Applications*, vol. 8, no. 1, nov 2019.