Simultaneous Neural Spike Encoding and Decoding Based on Cross-modal Dual Deep Generative Model

Qiongyi Zhou[1,2], Changde Du[1,2,3], Dan Li[1,2], Haibao Wang[1,2], Jian K. Liu[5] and Huiguang He[1,2,4,*]

[1]Research Center for Brain-inspired Intelligence and National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Huawei Cloud BU EI Innovation Lab, Beijing 100085, China
[4]Center for Excellence in Brain Science and Intelligence Technology,
Chinese Academy of Sciences, Shanghai 200031, China
[5]Centre for Systems Neuroscience, Department of Neuroscience, Psychology and Behaviour,
University of Leicester, Leicester LE1 7RH, U.K
Email: {zhouqiongyi2018, lidan2017, huiguang.he}@ia.ac.cn,
duchangde@gmail.com, haibaow@hotmail.com, jian.liu@leicester.ac.uk

***Abstract*—Neural encoding and decoding of retinal ganglion cells (RGCs) have been attached great importance in the research work of brain-machine interfaces. Much effort has been invested to mimic RGC and get insight into RGC signals to reconstruct stimuli. However, there remain two challenges. On the one hand, complex nonlinear processes in retinal neural circuits hinder encoding models from enhancing their ability to fit the natural stimuli and modelling RGCs accurately. On the other hand, current research of the decoding process is separate from that of the encoding process, in which the liaison of mutual promotion between them is neglected. In order to alleviate the above problems, we propose a cross-modal dual deep generative model (CDDG) in this paper. CDDG treats the RGC spike signals and the stimuli as two modalities, which learns a shared latent representation for the concatenated modality and two modal-specific latent representations. Then, it imposes distribution consistency restriction on different latent space, cross-consistency and cycle-consistency constraints on the generated variables. Thus, our model ensures cross-modal generation from RGC spike signals to stimuli and vice versa. In our framework, the generation from stimuli to RGC spike signals is equivalent to neural encoding while the inverse process is equivalent to neural decoding. Hence, the proposed method integrates neural encoding and decoding and exploits the reciprocity between them. The experimental results demonstrate that our proposed method can achieve excellent encoding and decoding performance compared with the state-of-the-art methods on three salamander RGC spike datasets with natural stimuli.**

***Index Terms*—dual learning, cross-modal generation, retinal ganglion cells, neural encoding, neural decoding**

I. Introduction

Visual pathway starts from retina where the light energy is transferred into neuronal signal, goes through lateral geniculate nucleus (LGN) and terminates in the visual cortex. Research has mainly focused on neural encoding and decoding of LGN and primary visual cortex and has made significant progress to date [1]–[7].

However, retinal ganglion cells (RGCs) are the only output neurons of retinas given visual stimuli. RGCs represent stimuli as spike trains in a very compact form and can be taken as independent information processors [8]. The encoding mechanism of RGCs can be utilized to build up retinal prostheses that perceive visual stimuli and generate simulated spike trains. It can be applied to vision recovery, even virtual and augmented reality through neural activity control [9]. Oppositely, neural decoding of RGCs can assess the performance of neuroprostheses, get deep insight into information compressed in spike trains and then be applied to brain-machine interfaces [8].

Till now, there have been lots of researches on RGC spike encoding. The existing methods contain the linear nonlinear model (LN) and its cascaded version LN-LN, the generalized linear model (GLM) taking spike history as feedback [10] and kinds of machine learning techniques [11]. However, the above methods only fit well on stimuli with simple artificial stimuli and are easy to overfitting with natural scenes which have more complicated distribution. These deficiencies are attributed to the complex nonlinear processes in neural circuits of retinas but relatively simple encoding models. To solve this problem, several methods based on deep neural networks (DNN) have been attempted, such as the convolutional neural network (CNN) [12] [13] or the recurrent neural network (RNN) [14] which have strong abilities to fit nonlinearity. These novel studies prove that deep learning is a brand-new and feasible way to mimic RGCs.

There has been some effort on RGC spike decoding. [15] provided a nonlinear decoder but can only execute pixel-by-pixel reconstruction of simple artificial stimuli. [16] used simulated spike data to generate coarse intermediate images firstly and refined them via a convolutional autoencoder. The experiments were conducted on simulated spike data but not experimental spike data. Due to imperfect encoding techniques especially when applied to natural stimuli, experimental spike data are more appropriate to assess the decoding method. [8] proposed a simple but efficient decoding algorithm and applied

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

it on experimental data. The idea was similar with [16] but it had no constraints on intermediate images. However, it is a pure decoding model and doesn't have the ability to encode stimuli. Accordingly, the liaison of mutual promotion between encoding and decoding is overlooked.

To our own knowledge, researches of RGC spike encoding and decoding have been isolated to date. However, encoding and decoding are dual processes. Simultaneous training can make use of the reciprocity between them.

Considering the above relationship and inspired by cross-modal generation, we propose a method called cross-modal dual deep generative model (CDDG) to compensate for the deficiencies of the current research. That's to say, visual stimuli and RGC spike signals are considered as two modalities. The method learns latent representations not only for the concatenated modality but also for two modalities specifically. And then, by forcing the distributions of three kinds of latent representations to be close, it establishes the relationship between image and spike modalities. Furthermore, cross-consistency and cycle-consistency constraints which are inspired by the concept of dual learning are forced onto generated variables to ensure higher ability for cross-modal generation. Thus, our model can achieve the generation from spike signals to stimuli and the inverse process. Generating RGC spikes given visual stimuli and generating visual stimuli given RGC spikes, are equivalent to encoding and decoding, respectively. RGC spike encoding and decoding are transformed into the bi-directional cross-modal generation issue. The cross-modal generation capability of CDDG supports the synchronic optimization and the mutual promotion of RGC encoding and decoding.

Experimental results demonstrate that our method accomplishes simultaneous neural encoding and decoding ideally. On three salamander RGC spike datasets with natural stimuli, it shows that our method achieves great encoding and decoding results compared with the state-of-the-art CNN-based RGC population spike encoder [12] and the state-of-the-art spike decoder [8].

In short, the main contributions of the paper are as follows.

- Inspired by the truth that our brains are bi-directional information-processing devices, we deploy a dual deep generative network to do simultaneous RGC spike encoding and decoding.
- We impose cross-consistency and cycle-consistency constraints on generated variables to obtain excellent cross-modal-generation capacity.
- The experimental results demonstrate that our approach can achieve excellent encoding and decoding performances in comparison with the state-of-the-art methods on three datasets with natural stimuli.
- Our work provides a new perspective and will inspire more work on RGC population spike encoding.

II. Related Work

A. Cross-modal generation

There has been a plenty of work on cross-modal generation. Automatic caption generation from images and the inverse processes have been achieved [17]–[20]. The cross-modal generation has also made progress in images, audio and so on [21]–[23]. Deep canonically correlated autoencoders (DCCAE) propounded in [24] can learn a shared representation through the correlation-based optimization and then reconstruct each modal. However, it only preserves the correlated information and abandons the uncorrelated one. Thus, it is inappropriate for cross-modal generation. [25] proposed JMVAE to do bi-directional cross-modal generation. It learned modal-specific latent representations and a modal-shared latent representations whose distributions were forced to be close. However, it had no constraints on the cross-modal generated variables and is flawed to do cross-modal generation. In contrast, the merit of our model is that it takes full-scale consistency constraints into consideration to acquire better cross-modal-generation performance.

Synchronic mutual generation of two modalities can improve the generation results of both directions but has not been attached much importance. A crucial issue is how to drive the performance of two generators to coevolve. An implicit form of constraints called cycle consistency is utilized to achieve this. Cycle consistency was first employed for dual learning of machine translation [26]. It enables the use of unpaired data, exploits the reciprocity of two generating processes and thus, has widespread applications in cross-modal generation [27]–[29]. In this paper, we resort to this idea for bi-directional cross-modal generation.

B. Neural spike encoding and decoding

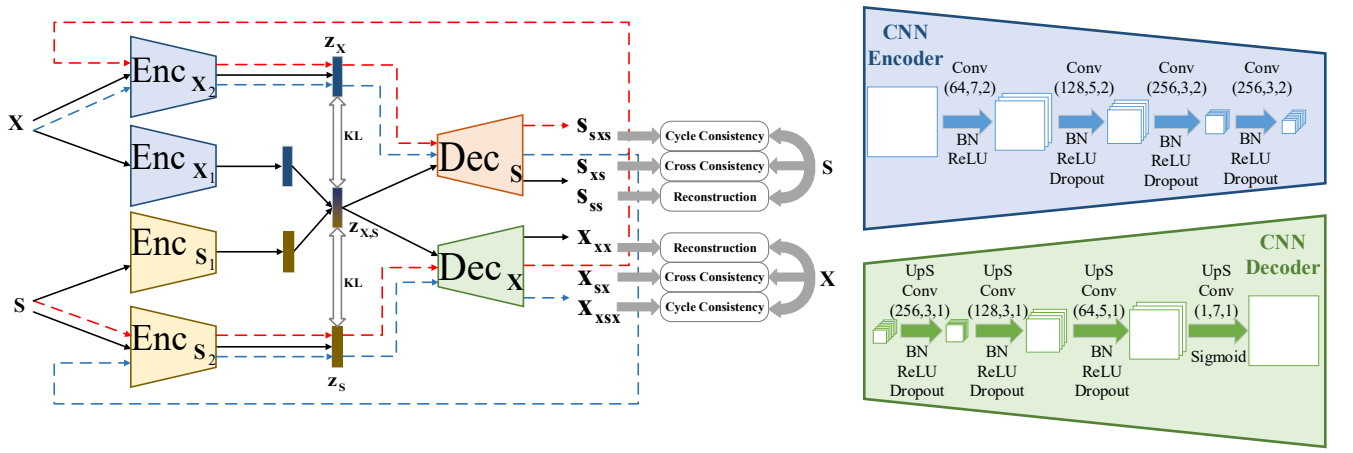
Neural spike encoding and decoding can be seen as processes of translation between stimuli and spike signals in different directions. There have been many classic spike encoding methods such as LN, LN-LN, GLM [10]. These methods use receptive fields of RGCs as spatiotemporal filters. However, the fitting abilities of them are limited especially on natural images. More complex models based on DNN learn the receptive fields automatically and have higher encoding performance [12]–[14]. The DNN-based encoding model inspires the CNN-based encoding part of CDDG.

Researches on RGC spike decoding have shown good results [8], [15], [16]; however, they have been isolated from researches of spike encoding. In fact, it's essential to integrate spike encoding and decoding into one framework. Taking the retinal neuroprosthesis as an example, the synchronic training of encoding and decoding models can promote the performance and the performance evaluation at the same time and is helpful to obtain perfect neuroprostheses [30]. Our proposed model CDDG builds up a closed-loop computation of spike encoding and decoding which is the main difference compared with the state-of-the-art spike encoding and decoding methods.

III. Methodology

A. Overview

As for the multi-modal generation issue, two modalities, visual stimuli and RGC spike signals, are represented as $\mathbf{x} \in \mathbb{R}^{N\times P\times P}$ and $\mathbf{s} \in \mathbb{R}^{N\times M}$, respectively. N, P and M



(a) (b)

Fig. 1. The schematic diagram of CDDG and related CNN architecture. (a) It depicts the network structure and constraints of CDDG. It's made up of four encoders and two decoders. And it has six outputs and four kinds of constraints. The black solid arrows denote the self-reconstruction processes and the dotted arrows in red and blue denote the cycle-generation processes. The abbreviation KL means KL-divergence constraints. The wide gray arrows in the right of the subfigure denote three constraints imposed on six outputs. See Section III for more details. (b) CNN architecture and specific parameter settings. The architecture in upper blue trapezoid is used to build an image encoder. Oppositely, the architecture in upper green trapezoid is used to build an image decoder.

denote the sample size, the image resolution, the number of RGCs. The generated variables are represented as $\mathbf{x}$ or $\mathbf{s}$ with a specific subscript meaning the generation routine. For example, $\mathbf{x}_{\mathbf{sx}}$ is the stimuli $\mathbf{x}$ generated from the original spike signals $\mathbf{s}$.

The following section is divided into three parts with the first two parts introducing the skeleton models that CDDG in the third part is based on. The most basic one is called the multi-modal deep generative model (MDG). The black arrows in Fig. 1 (a) describe the generation of $\mathbf{x}_{\mathbf{xx}}, \mathbf{s}_{\mathbf{ss}}$ with reconstruction and KL-convergence constraints. The polished model called cross-modal deep generative model (CDG) adds cross-modal-generation constraints for the sake of domain alignments. It has additional outputs $\mathbf{x}_{\mathbf{sx}}, \mathbf{s}_{\mathbf{xs}}$. Inspired by dual learning, we ameliorate CDG into cross-modal dual deep generative model (CDDG) with additional cycle-consistency constraints. The generation routines of $\mathbf{x}_{\mathbf{xx}}, \mathbf{s}_{\mathbf{ss}}, \mathbf{x}_{\mathbf{sx}}, \mathbf{s}_{\mathbf{xs}}, \mathbf{x}_{\mathbf{xsx}}, \mathbf{s}_{\mathbf{sxs}}$ are presented by the black solid arrows, the blue and red dotted arrows in Fig. 1 (a). Three kinds of constraints imposed on the generated variables are shown on the right of Fig. 1 (a) by thick gray arrows.

B. Multi-modal Deep Generative Model

The variational autoencoder (VAE) algorithm focuses on single modality learning [31]. It's made up of one encoder $q_\phi(\cdot)$ and one decoder $p_\theta(\cdot)$, with ϕ and θ as their respective network parameters. Given the i.i.d. observation variables $\mathbf{x}$ and continuous latent variables $\mathbf{z}$, the loss function of VAE is

$$\mathcal{L}_{\text{VAE}} = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - E_{q_{\phi(\mathbf{z}|\mathbf{x})}}[\log p_\theta(\mathbf{x}|\mathbf{z})] \quad (1)$$

where $\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. The first term in (1) represents a regularization on $q_\phi(\mathbf{z}|\mathbf{x})$ and the second term in (1) represents reconstruction constraints on $\mathbf{x}$. The VAE also uses the reparameterization trick to change $\mathbf{z}$ into $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The applications of VAE include reconstruction, unsupervised representation learning of single modality.

VAE can be extended to multi-modal version with multiple inputs and outputs. The inputs are the i.i.d. multi-modal dataset $(\mathbf{x}, \mathbf{s}) = \{(\mathbf{x}_1, \mathbf{s}_1), (\mathbf{x}_2, \mathbf{s}_2), \cdots, (\mathbf{x}_N, \mathbf{s}_N)\}$, where $\mathbf{x}, \mathbf{s}, N$ are told in Section III-A. The generating processes are represented as $\mathbf{x}, \mathbf{s} \sim p(\mathbf{x}, \mathbf{s}|\mathbf{z}) = p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}) p_{\theta_\mathbf{s}}(\mathbf{s}|\mathbf{z})$. $p_{\theta_\mathbf{x}}(\cdot), p_{\theta_\mathbf{s}}(\cdot)$ are decoders of $\mathbf{x}$ and $\mathbf{s}$ corresponding to $\text{Dec}_\mathbf{x}, \text{Dec}_\mathbf{s}$ in Fig. 1 (a) with $\theta_\mathbf{x}$ and $\theta_\mathbf{s}$ as their respective parameters. Inversely, the inference processes are $\mathbf{z}_{\mathbf{x},\mathbf{s}} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x},\mathbf{s}}, \boldsymbol{\sigma}^2_{\mathbf{x},\mathbf{s}})$, $\mathbf{z}_\mathbf{x} \sim q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\sigma}^2_\mathbf{x})$ and $\mathbf{z}_\mathbf{s} \sim q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{s}, \boldsymbol{\sigma}^2_\mathbf{s})$, where $q_{\phi_\mathbf{x}}(\cdot), q_{\phi_\mathbf{s}}(\cdot), q_\phi(\cdot)$ are encoders of $\mathbf{x}$, $\mathbf{s}$ and their concatenated modality $(\mathbf{x}, \mathbf{s})$ with $\phi_\mathbf{x}$, $\phi_\mathbf{s}$ and ϕ as their respective parameters. The encoders $q_{\phi_\mathbf{x}}(\cdot), q_{\phi_\mathbf{s}}(\cdot)$ learn modal-specific latent representations while $q_\phi(\cdot)$ learns a modal-shared latent representation. The networks $\text{Enc}_{\mathbf{x}_2}, \text{Enc}_{\mathbf{s}_2}$ in Fig. 1 (a) are corresponding to $q_{\phi_\mathbf{x}}(\cdot), q_{\phi_\mathbf{s}}(\cdot)$, respectively. $(\text{Enc}_{\mathbf{x}_1}, \text{Enc}_{\mathbf{s}_1})$ are taken as a whole encoding network that is equivalent to $q_\phi(\cdot)$. The loss function of MDG is

$$\begin{aligned}\mathcal{L}_{\text{MDG}} = & D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s})||p(\mathbf{z})) + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s})||q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})) \\ & + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s})||q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s})) \\ & - \alpha(E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})}[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})] + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})}[\log p_{\theta_\mathbf{s}}(\mathbf{s}|\mathbf{z})]) \quad (2)\end{aligned}$$

where $\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and α is the trade-off parameter. The first three terms are intended to close the distributions of $\mathbf{z}, \mathbf{z}_\mathbf{x}, \mathbf{z}_\mathbf{s}, \mathbf{z}_{\mathbf{x},\mathbf{s}}$. The last two terms in (2) are reconstruction constraints on $\mathbf{x}$ and $\mathbf{s}$ respectively, which are similar to the second term in (1).

In brief, MDG consists of three encoders and two decoders. The encoder $q_\phi(\cdot)$ is disabled during testing because the testing procedure generates $\mathbf{x}$ only from $\mathbf{s}$ and vice versa. The inference processes during the testing stage only contains $\mathbf{z}_\mathbf{x} \sim q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\sigma}^2_\mathbf{x})$ and $\mathbf{z}_\mathbf{s} \sim q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{s}, \boldsymbol{\sigma}^2_\mathbf{s})$. This model has several advantages. The gap between two modalities is easier to close using the concatenated modality

as an intermediate variable. Besides, the modal-specific encoders make the model supportive for datasets with incomplete modalities.

C. Cross-modal Deep Generative Model

Nonetheless, the MDG model is flawed for cross-modal generation. The reason is that MDG only imposes reconstruction constraints on $\mathbf{x_{xx}}$ and $\mathbf{s_{ss}}$ which are generated from the modal-shared latent representation $\mathbf{z_{x,s}} \sim q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})$ but overlooks constraints on $\mathbf{x_{sx}}$ and $\mathbf{s_{xs}}$ generated from $\mathbf{z_s} \sim q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s})$ and $\mathbf{z_x} \sim q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})$, respectively. Therefore, the above model is only suitable for modality reconstruction, not for cross-modal generation.

We introduce the cross-modal deep generative model (CDG) with the following loss function to relieve the problem,

$$\begin{aligned}\mathcal{L}_{\text{CDG}} &= D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})||p(\mathbf{z})) + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})||q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})) \\ &+ D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})||q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s})) \\ &- \alpha(E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})}[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})] + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})}[\log p_{\theta_\mathbf{s}}(\mathbf{s}|\mathbf{z})]) \\ &- \beta(E_{q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s})}[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})] + E_{q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})}[\log p_{\theta_\mathbf{s}}(\mathbf{s}|\mathbf{z})]) \end{aligned} \quad (3)$$

where β is the trade-off parameter. The last two additional terms play the role of minimizing the reconstruction errors of $\mathbf{x_{sx}}$ and $\mathbf{s_{xs}}$. The terms put constraints on cross-modal-generation consistency and we call them cross-consistency constraints.

In order to analysis the necessity of the added terms, we consider the case in (2) and take the generation of $\mathbf{x_{sx}}$ as an example. Similar distributions are not equivalent to similar decoding results. That is to say, sampling from the modal-specific and the modal-shared latent representations $\mathbf{z_s}$ and $\mathbf{z_{x,s}}$ which have similar distributions, the generated variables $\mathbf{x_{sx}}$ and $\mathbf{x_{xx}}$ are not bound to be consistent even both generated by $p_{\theta_\mathbf{x}}(\cdot)$. Therefore, there need extra constraints on generated variables and the additional terms in (3) play the role. As a result, the cross-consistency can help the domain alignment.

D. Cross-modal Dual Deep Generative Model

Inspired by dual learning, we input the generated variables $\mathbf{x_{sx}}$ and $\mathbf{s_{xs}}$ into one more cross-modal generation and obtain $\mathbf{s_{sxs}}$ and $\mathbf{x_{xsx}}$ with additional cycle-consistency constraints.

We further modify CDG model into CDDG model whose second D represents the abbreviation of dual,

$$\begin{aligned}\mathcal{L}_{\text{CDDG}} &= D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})||p(\mathbf{z})) + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})||q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})) \\ &+ D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})||q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s})) \\ &- \alpha(E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})}[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})] + E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})}[\log p_{\theta_\mathbf{s}}(\mathbf{s}|\mathbf{z})]) \\ &- \beta(E_{q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s})}[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})] + E_{q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})}[\log p_{\theta_\mathbf{s}}(\mathbf{s}|\mathbf{z})]) \\ &- \gamma(E_{q_{\phi_\mathbf{s}}(\mathbf{z}|\mathbf{s_{xs}})}[\log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z})] + E_{q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x_{sx}})}[\log p_{\theta_\mathbf{s}}(\mathbf{s}|\mathbf{z})]) \end{aligned} \quad (4)$$

where γ is the trade-off parameter.

Taking $\mathbf{x_{sx}}$ generated from $\mathbf{s}$ as an example, the added constraints encourage $\mathbf{x_{sx}}$ to generate $\mathbf{s_{sxs}}$ aligned with the original variables $\mathbf{s}$ through one domain-cycle (image $\rightarrow$ spike $\rightarrow$ image). The supplementary terms have lots of benefits. On the one hand, some modal-specific features will be p-reserved during the cross-modal generation for the purpose of minimizing the corresponding reconstruction error in the closed loop. On the other hand, as inspired by [32], the cycle-consistency terms can mitigate the underconstrained cross-domain generation issue and then, enable the learning of cross-modal generation on modal missing datasets which are common for neural data.

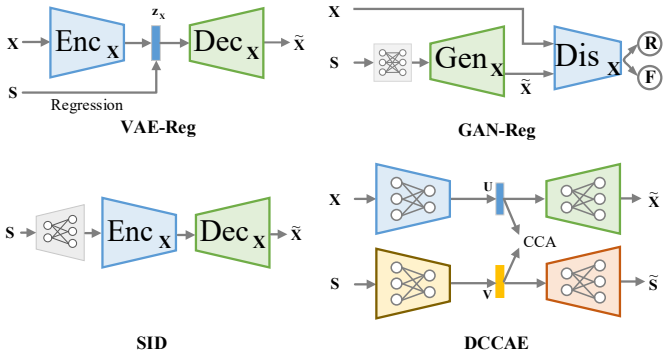


Fig. 2. Diagrams of the compared methods. See Section IV-A for more details.

IV. Experimental Results

In this section, we give an introduction on the compared method, datasets, experimental settings and evaluation metrics used in experiments. We also show experimental results via figures and charts.

A. Compared Methods

The compared methods include both unidirectional and bi-directional cross-modal generators. The diagram of each neural decoding method is shown in Fig. 2.

- **VAE-based regression (VAE-Reg)** [33]: The model combines the vanilla VAE and regression together and can only generate images from spike singals. It learns the image representation using VAE firstly. Then, it does regression from the neural spike signals to $\mathbf{z_x}$. Finally, $\mathbf{z_x}$ reconstruct images through $\text{Dec}_\mathbf{x}$.
- **GAN-based regression (GAN-Reg)** [6]: The GAN is trained to generate images from random noise. After that, parameters of GAN are fixed and a fully-connected (FC) network is trained to estimate the latent space from neural spike signals.
- **Spike-Image Decoder (SID)** [8]: The SID contains two parts: spike-image converter and image-image autoencoder. The converter turns RGC spike signals into intermediate images from which the autoencoder reconstructs images next.
- **DCCAE** [24]: This is a deep multi-view algorithm extended from canonical correlation analysis (CCA) [34]. It designs one autoencoder for each modality and applies CCA on the learnt latent representations.
- **CNN-based** [12]: It can only generate spike signals from images. Images are encoded into spike signals by CNNs and FC networks. [12] has proved that this neural encoding method outperforms LN and GLM on natural stimuli. The diagram can be found in [12].

B. Datasets

We use three datasets publicized in [35]. The RGC spike signals were collected from isolated salamander retinas with natural stimuli which contain static images and dynamic movie clips. They are named by the type of the stimuli.

1) Natural Image: The stimuli contain 300 different gray natural images. Each stimulus was shown to the retina for 200ms. Neural spike trains of 80 RGCs were recorded and summed up to spike counts in bins of 200ms.

2) Natural Movie-I: The movie clip is 60s long and at a frame rate of 30Hz. So, there are 1800 gray natural frames in total. Neural spike trains of 90 RGCs were recorded and binned in bins of $1000/30$ ms.

3) Natural Movie-II: The movie clip is at a frame rate of 30Hz and has 1600 gray natural frames in total. Neural spike trains of 49 RGCs were recorded and binned in bins of $1000/30$ ms.

Binning spike trains in bins can transfer spike trains into spike counts which don't have the temporal structure so that to remove noise. All experimental processes resized stimuli into 64×64 resolution except the DCCAE method in which each image was flatten as a vector before fed into the network. For the sake of fast converging, we normalized the pixels of all stimuli to $[0, 1]$. As for every sample of $\mathbf{s}$, it contains average counts of each RGC over all trials to each stimulus. In experiments, 90% data is used for training and the remaining 10% is for model testing.

C. Experimental Settings

All models used in the experiments except the DCCAE exploit CNN and deconvolutional neural networks (De-CNN) as image encoders and decoders, respectively. All image encoder modules and the discriminator in GAN-Reg model share the same parameters of the layer design, kernel size, stride and filter number. The same to all De-CNN decoder modules and the generator in GAN-Reg model. The architecture of the image encoder and decoder is shown in Fig. 1 (b). The triples in each pair of parentheses represent the filter number, kernel size and stride of each convolution operation. The abbreviations BN and UpS denote batch normalization and upsampling, respectively. The parameter settings and training mode of all methods are written as follows.

- **CDDG**: For the sake of dimension consistency, an FC network was used to reduce the dimension of $\mathbf{z_{x,s}}$ to be the same as the one of $\mathbf{z_x}$ and $\mathbf{z_s}$. The number of layers of the FC networks of spike encoder and decoder was 2 for static images and 3 for dynamic movie clips. The training of CDDG model was end-to-end. In practice, the trade-off parameters (α, β, γ) were set to $(100, 100, 1)$ and $(1000, 1000, 10)$ for the static image dataset and the dynamic movie datasets respectively. The dimension of latent variables and the learning rate were set to 256 and 0.001.
- **VAE-Reg**: The dimension of latent variables and learning rate were set to 256 and 0.001. Regression algorithms Lasso and k-nearest neighbor (k-NN) were adopted for static images and dynamic movies, respectively.
- **GAN-Reg**: The dimension of the latent space of the GAN was set to 90. The learning rates when training the GAN and the FC network were set to 0.0002 and 0.001, respectively.
- **SID**: The SID adopted end-to-end training with reconstruction constraints. The learning rate was set to 0.001.
- **DCCAE**: It used FC but not CNN networks for computation acceleration. The latent space dimension was set to 16, 32, 16 for natural image, natural movie-I and movie-II datasets respectively. The learning rate was 0.001.
- **CNN-based**: The number of layers of the FC networks was 2. The learning rate was set to 1×10^{-5}.

D. Performance Evaluation

Here we denote the metrics used for performance evaluation on our model and the compared methods.

1) Neural Encoding Quality Metrics: We encode the spike signals into spike counts in our paper. Considering this data property, we use mean square error (MSE) to evaluate the encoding performance of our model, DCCAE and the CNN-based RGC encoding model. The metrics reflect the level of spike counting bias averaged on all cells of all samples.

2) Image Quality Metrics:

- Mean Square Error (MSE): MSE denotes the expectation of the squared error between predicted and original pixel values. The calculation of MSE for a pair of images $\langle \mathbf{I}_1, \mathbf{I}_1 \rangle$ with the resolution of $H \times W$ is

$$\text{MSE} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (\mathbf{I}_1(i,j) - \mathbf{I}_2(i,j))^2 \quad (5)$$

  Generally, the lower the MSE metric is, the better the image quality is.
- Structural-Similarity Metric (SSIM): SSIM can conduct structure comparison between two images. It was proposed in [36] under the assumption that human vision perceives image distortion by extracting structural information changes. The calculation of SSIM of images $\langle \mathbf{I}_1, \mathbf{I}_2 \rangle$ is

$$\text{SSIM} = \frac{(2\mu_{\mathbf{I}_1}\mu_{\mathbf{I}_2} + c_1)(2\sigma_{\mathbf{I}_1\mathbf{I}_2} + c_2)}{(\mu_{\mathbf{I}_1}^2 + \mu_{\mathbf{I}_2}^2 + c_1)(\sigma_{\mathbf{I}_1}^2 + \sigma_{\mathbf{I}_2}^2 + c_2)} \quad (6)$$

  where $\mu_{\mathbf{I}_1}, \mu_{\mathbf{I}_2}$ are mean of $\mathbf{I}_1, \mathbf{I}_2$, $\sigma_{\mathbf{I}_1}^2, \sigma_{\mathbf{I}_2}^2$ are variance of $\mathbf{I}_1, \mathbf{I}_2$, $\sigma_{\mathbf{I}_1\mathbf{I}_2}$ is covariance of $\mathbf{I}_1, \mathbf{I}_2$, c_1, c_2 are constants for computational stability.
  SSIM metrics have a roughly positive relation with image quality. There is another image quality metric called Peak Signal to Noise Ratio (PSNR). It has approximately opposite changes to MSE and so it's a little redundant to use. Here we use only MSE and SSIM as evaluation metrics.

TABLE I
EVALUATION ON NEURAL ENCODING PERFORMANCE ON TEST SETS OF THREE DATASETS WITH DIFFERENT METHODS. THE OPTIMAL VALUE ON EACH METRIC IS HIGHLIGHTED.

Encoding Method	Natural Image	Natural Movie-I	Natural Movie-II
	MSE	MSE	MSE
DCCAE	1.915	0.050	0.006
CNN-based	0.529	**0.030**	0.006
CDDG	**0.527**	**0.030**	**0.004**

TABLE II
EVALUATION OF NEURAL DECODING PERFORMANCE ON TEST SETS OF THREE DATASETS WITH DIFFERENT METHODS. THE OPTIMAL VALUE ON EACH DATASET AND EACH METRIC IS HIGHLIGHTED.

Decoding Method	Natural Image		Natural Movie-I		Natural Movie-II	
	MSE	SSIM	MSE	SSIM	MSE	SSIM
VAE-Reg	0.031	**0.493**	0.029	0.637	0.060	0.280
GAN-Reg	0.038	0.322	0.025	0.480	0.049	0.189
SID	0.029	0.391	**0.008**	**0.763**	**0.031**	0.408
DCCAE	0.030	0.331	0.027	0.546	0.048	0.246
CDDG	**0.027**	0.385	0.012	0.706	0.034	**0.421**

TABLE III
EVALUATION OF NEURAL ENCODING AND DECODING PERFORMANCE ON TEST SETS OF THREE DATASETS WITH ABLATION EXPERIMENTAL METHODS. THE OPTIMAL VALUE ON EACH METRIC IS HIGHLIGHTED.

	Natural Image			Natural Movie-I			Natural Movie-II		
	MSE_s	MSE	SSIM	MSE_s	MSE	SSIM	MSE_s	MSE	SSIM
MDG	0.671	0.031	0.343	0.039	0.020	0.612	0.006	0.045	0.331
CDG	0.558	0.031	0.350	**0.030**	0.015	0.687	0.005	**0.032**	**0.430**
CDDG	**0.527**	**0.027**	**0.385**	**0.030**	**0.012**	**0.706**	**0.004**	0.034	0.421

E. Encoding Performance

The performances of three encoding methods are shown in Table I. It can be seen that CDDG surpasses DCCAE and the CNN-based neural encoding model on three datasets. Our model achieves better encoding performance even in compared with the CNN-based method which has much better encoding ability than LN and GLM [12]. The CNN-based neural encoding model can be taken as a single-modal generative model while CDDG is a multi-modal generative model. Our model achieves simultaneous neural encoding and decoding and then utilizes the reciprocity of the dual processes. That's the reason why our model has superiority.

F. Decoding Performance

Examples of the decoding results on three datasets, Natural Image, Natural Movie-I and Natural Movie-II, are shown in Fig. 3. Images in the first row of each subfigure are original images. Other rows list decoding results with their method name marked on the left. Among these approaches, CDDG and SID work best because they draw the outline of scenes in Fig. 3 (a), reconstruct the eyes of the swimming salamanders in Fig. 3 (b) and the face of the tiger in Fig. 3 (c) clearly. Compared to SID, CDDG reconstructs the images in a more sharp way with less blur especially in Fig. 3 (b) and (c). The VAE-Reg depicts the light and shade parts of images in Fig. 3 (a) but the images look messy in Fig. 3 (b) and (c). Results of GAN-Reg and DCCAE also have a lot of noise.

Table II shows the objective metrics on image quality. As for Natural Image dataset, CDDG achieves the lowest MSE and SID achieves the highest SSIM except for the VAE-Reg method. However, it can be seen that the results of CDDG and SID in Fig. 3 (a) are visibly better and more legible than those of VAE-Reg. The phenomenon reflects that SSIM has weakness in distinguishing the distortion level between a blurred image and a low-noise image that has been discussed in [37]. On the other two datasets, CDDG obtains the lowest MSE and the highest SSIM among all methods except for SID. It matches the state-of-the-art decoding method SID on every dataset. In general, CDDG and SID are well-matched in both the subjective perception and objective assessment.

In short, the proposed method CDDG is far superior to the bi-directional cross-modal-generation method DCCAE in term of spike decoding on all datasets. In addition, CDDG has similar performance with SID which is the greatest unidirectional cross-modal-generation approach among the compared methods and the state-of-the-art RGC decoding method.

G. Ablation Experiments

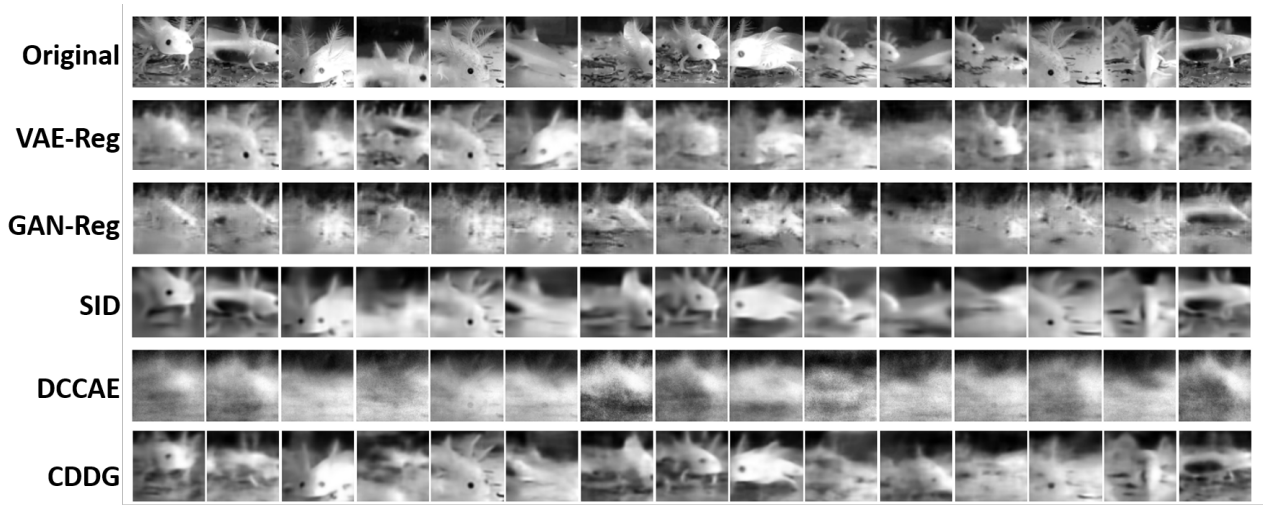
We conducted series of ablation experiments on three datasets to prove that every additional term in (4), cross-consistency and cycle-consistency, makes positive effect on our model performance. First of all, we used the function in (2), the standard loss of MDG. Secondly, we used Equation (3) of CDG model as the loss function that has two added cross-consistency constraints to encourage domain alignment. Finally, we added two cycle-consistency constraints to encourage that a sample from one modality could still keep consistency with itself after two rounds of cross-modal generation. Equation (4) of CDDG was used as the objective function.

The results are shown in Table III. To distinguish it from the MSE of images, we denote the MSE of spike signals as $\mathbf{MSE_s}$ As for both encoding and decoding metrics on all datasets, CDG outperforms MDG. It is obvious that the cross-consistency can help the improvement of the model performance. CDDG surpasses or is equal to CDG on almost all metrics. The improvement from CDG to CDDG is not very significant. However, the advantages of our model for semi-supervised learning with modal missing data cannot be ignored.

V. CONCLUSION

Inspired by the state-of-the-art DNN-based neural spike encoding and decoding methods, we propose a cross-modal dual deep cross-generative model into consideration to do bi-directional cross-modal generation between visual stimuli and spike signals of retinal ganglion cells. It's the first attempt to integrate the RGC encoding and decoding processes into one framework for reciprocity. CDDG performs well compared with other neural decoding methods on different datasets with natural stimuli. It matches with the state-of-the-art RGC spike decoder well. Meanwhile, it has higher ability to encode RGC spike compared with the state-of-the-art RGC spike encoding method. To summarize, our method achieves the promotion

(a)



(b)

(c)

Fig. 3. Examples of decoding results on three datasets with CDDG and compared methods. (a) Results on the Natural Image dataset. (b) Results on the Natural Movie-I dataset. (c) Results on the Natural Movie-II dataset.

on neural spike encoding while keeping excellent decoding performance as a result of simultaneous training.

In the future, we will extend our method into semi-supervised learning on datasets with incomplete modalities. Besides, we expect to utilize RNN to reconstruct video stimuli. As for concrete applications, we plan to use our model to develop and evaluate retinal prostheses with the support of more neuroscience knowledge.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 61976209, 61906188), CAS International Collaboration Key Project, and the Strategic Priority Research Program of CAS (XDB32040200).

REFERENCES

[1] G. B. Stanley, F. F. Li, and Y. Dan, "Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus," *Journal of Neuroscience*, vol. 19, no. 18, pp. 8036–8042, 1999.

[2] S. Ling, M. S. Pratte, and F. Tong, "Attention alters orientation processing in the human lateral geniculate nucleus," *Nature neuroscience*, vol. 18, no. 4, p. 496, 2015.

[3] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fmri," *Neuroimage*, vol. 56, no. 2, pp. 400–410, 2011.

[4] Y. Miyawaki, H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.

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

[6] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. A. van Gerven, "Generative adversarial networks for reconstructing natural images from brain activity," *NeuroImage*, vol. 181, pp. 775–785, 2018.

[7] 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.

[8] 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.

[9] T. Siriborvornratanakul, "Through the realities of augmented reality," in *International Conference on Human-Computer Interaction*, pp. 253–264, Springer, 2019.

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

[11] G. P. Das, P. J. Vance, D. Kerr, S. A. Coleman, T. M. McGinnity, and J. K. Liu, "Computational modelling of salamander retinal ganglion cells using machine learning approaches," *Neurocomputing*, vol. 325, pp. 101–112, 2019.

[12] L. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. Baccus, "Deep learning models of the retinal response to natural scenes," in *Advances in neural information processing systems*, pp. 1369–1377, 2016.

[13] 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," *arXiv preprint arXiv:1811.02290*, 2018.

[14] E. Batty, J. Merel, N. Brackbill, A. Heitman, A. Sher, A. Litke, E. Chichilnisky, and L. Paninski, "Multilayer recurrent network models of primate retinal ganglion cell responses," in *International Conference on Learning Representations*, 2017.

[15] 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.

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

[17] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

[18] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[19] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, pp. 2048–2057, 2015.

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[21] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 349–357, ACM, 2017.

[22] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3550–3558, 2018.

[23] Y. Wen, B. Raj, and R. Singh, "Face reconstruction from voice using generative adversarial networks," in *Advances in Neural Information Processing Systems*, pp. 5266–5275, 2019.

[24] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning: objectives and optimization," *arXiv preprint arXiv:1602.01024*, 2016.

[25] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," *arXiv preprint arXiv:1611.01891*, 2016.

[26] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," in *Advances in Neural Information Processing Systems*, pp. 820–828, 2016.

[27] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[28] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, pp. 2849–2857, 2017.

[29] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "It takes (only) two: Adversarial generator-encoder networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[30] Z. Yu, J. K. Liu, S. Jia, Y. Zhang, Y. Zheng, Y. Tian, and T. Huang, "Towards the next generation of retinal neuroprosthesis: Visual computation with spikes," *Engineering*, 2020.

[31] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[32] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1602–1612, 2018.

[33] Y. Yoo, S. Yun, H. Jin Chang, Y. Demiris, and J. Young Choi, "Variational autoencoded regression: high dimensional regression of visual data on complex manifold," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2017.

[34] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*, pp. 162–190, Springer, 1992.

[35] 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, 2016.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[37] X. Fei, L. Xiao, Y. Sun, and Z. Wei, "Perceptual image quality assessment based on structural similarity and visual masking," *Signal Processing: Image Communication*, vol. 27, no. 7, pp. 772–783, 2012.