Depth Perception with Interocular Blur Differences based on a Spiking Network

Weisi Liu¹ and Xinsheng Liu¹*

¹ State Key Laboratory of Mechanics and Control of Mechanical Structures, Institute of Nano Science and Department of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

*Corresponding author: Xinsheng Liu (E-mail address: xsliu@nuaa.edu.cn).

This work was supported in part by National Natural Science Foundation of China (61374183, 51535005), National Key Research and Development Program of China (2019YFA0705400), the Research Fund of State Key Laboratory of Mechanics and Control of Mechanical Structures (MCMS-I-0419K01), the Fundamental Research Funds for the Central Universities (NJ2020003, NZ2020001), A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

ABSTRACT Visual depth perception is the basic function of the visual nervous system. To a stimulus in the stereo space, the visual nervous system could generate a perception about depth of its position. Experimental data have demonstrated that interocular blur differences could lead to illusory perceptions about depths of moving stimuli. However, to stimuli with interocular blur differences, influences of different factors on illusory depth perceptions are still unclear. To explore these influences, this paper constructs a plastic two-layer k-winner-take-all (k-WTA) spiking network, simulating primary visual cortical responses. With incompatible stimuli presented into two eyes in experiments, binocular rivalry could occur in the primary visual cortex and interact with depth perception. To simulate binocular rivalry, the network consists of two parallel visual channels driven by left-eye and right-eye stimuli and competing with each other through mutual inhibition. In simulations, the horizontally moving stimulus is filtered with different Gaussian filters to generate paired monocular stimuli with interocular blur differences. The blurry strength, the moving direction and the moving speed perform as varying factors of moving stimuli. The network updates its dynamics through probabilistic inference, reflecting impacts of each factor on both neural responses and binocular rivalry. The modified responses could simulate illusory depth perceptions of stimuli as observed in experiments. To stimuli with interocular blur differences, varying factors could modify binocular rivalry in the network, inducing distinguishing illusory depth perceptions. Based on probabilistic inference, our model could provide possible explanations to illusory depth perceptions with interocular blur differences.

INDEX TERMS Depth perception, Binocular rivalry, Winner-take-all, Spiking neural networks.

I. INTRODUCTION

Visual motion perception is an important function of our brains and may be influenced by many possible factors, such as the saccadic eye movements, contexts, contrasts, the spatiotemporal frequency and so on [1]–[8]. To a stimulus in the three-dimension space, the visual nervous system could generate a perception about the depth of its stereo location. Particularly, interocular blur differences caused by a prescription lens correction have been found to induce the illusory perception about the depth of the moving stimulus [3]. These illusions may impact public safety. However, the influence of different factors in the depth misperception induced by interocular blur differences are still unclear.

People observe the external environment with two eyes. Previous experimental studies have demonstrated the effect of binocular vision on visual motor perception. Biologically, binocular vision affects the activity of the primary visual
cortex (V1) during visual perception [9]–[12]. Two eyes receive monocular images and transform visual information into sensory signals [13]–[15]. Starting from V1, these monocular signals rival and fuse into a stable view of the world with depth [16], [17]. Binocular visual mechanisms perform as the basic background in the exploration of stereo visual perceptions [18]–[20]. With incompatible stimuli presented to two eyes, binocular rivalry could occur in the primary visual cortex. For depth perception, previous experimental studies have demonstrated that binocular rivalry could coexist and interact with it [21]–[23]. Through the dichoptic view of random-dot patterns with various noisy contrasts, experimental results have found that rivalry and stereopsis occur parallelly and independently [21]. To visual plaid patterns in experiments, depth perception and rivalry could coexist in different spatial frequency and orientation bands [22]. Stereoscopic depth perception was evident even when incompatible monocular images engage in binocular rivalry, demonstrating binocular rivalry and stereopsis could coexist [23]. To explain illusory depth perceptions induced by interocular blur differences, binocular vision should be the important biological background and binocular rivalry might be the possible mechanism.

Neural responses in the visual cortex are associated with various visual functions, including the contour matching, contour detection, the associative processing, the orientation selectivity [24]–[28]. For the primary visual cortex, the biological background of binocular rivalry has been modeled by spiking networks [9], [17], [29]. Probabilistic inference has been used to explore the principle of neural coding, such as the phenomenon of visual perceptions [30]–[32]. In the probabilistic framework, the winner-take-all spiking networks could simulate the primary visual cortical activities [33]–[35]. With neural spikes, the learning rules of connective weights in these networks take the forms of the Hebbian spike-timing-dependent plasticity (STDP) [36].

Besides, the temporal structure of neural spiking trains has been proved conducive to the information transformation, compared to the rate-based codes [37]. Particularly, the k-winner-take-all (k-WTA) network could simulate the simultaneous spiking responses of multiple neurons which is the general experimental phenomenon [33].

In the exploration of primary visual cortical mechanisms, the winner-take-all networks are designed in the two-layer structures with simplification [33]–[35]. The first-layer neurons transform the visual images into spiking trains to the second-layer network. The second-layer network is designed following the ubiquitous cortical microcircuits in layers 2/3, consisting of pyramidal cells with lateral inhibition [38]. The second-layer neurons perform as decoders to outside images. With plastic connective weights following Hebbian STDP, the WTA network could simulate primary visual cortical sparse responses through probabilistic inference and decode the visual stimuli. Our previous work has explored how different eye movements induce the neural responding variability with the k-WTA spiking network [39]. Yet, these WTA networks have not explored visual perception with binocular vision in the primary visual cortex.

This paper constructs a two-layer k-winner-take-all spiking network with plastic connections, considering binocular rivalry. With the designed network, this paper would explore the influences of blurry strengths, moving directions and moving speeds on depth perception of moving stimuli with interocular blur differences. In the network, the first-layer modeling neurons transform the information of monocular visual stimuli into Poisson spiking trains towards...
the second layer. The second layer consists of several groups of excitatory neurons which simulate monocular and binocular neurons in the primary visual cortex, respectively. Binocular rivalry is simulated through the mutual inhibition between two groups of monocular neurons. The binocular neurons receive all the lateral inputs from monocular neurons, fusing the visual information. Based on this network, this paper first simulates binocular rivalry and responding modulation induced by the moving visual stimuli along opposite directions used in [3]. After that, depth perceptions of stimuli with interocular blur differences are explored. To obtain stimuli with interocular blur differences, the bar is filtered with different Gaussian filters to generate different blurs, as introduced in [3]. Controlling the variables, three kinds of simulations are designed respectively to explore the impacts of blurry strength, moving directions and moving speeds upon binocular rivalry and depth perception. With binocular rivalry, the network updates its dynamics through probabilistic inference. Then, the network simulates depth perception of the horizontal moving path with the monocular images and induced network dynamics. The obtained depth perceptions with illusions accord with experimental observations. Our simulations show that, to the moving stimuli with interocular blur differences, the blurry strength, moving directions and moving speeds could modify the responses and binocular rivalry from the network and induce bias depth perceptions. Factors lead to illusory depth perceptions by inducing different binocular rivalry in the network. Based on probabilistic inference, our network provides the possible framework with binocular rivalry to explore the depth perception with interocular blur differences.

The article is structured as follows. In section II, the structure of the spiking network and the corresponding probabilistic model are illustrated. In section III, depth perceptions of moving stimuli with interocular blur differences are investigated. The interocular blur difference, the moving direction and the moving speed perform as the factors on depth perception of moving stimuli. In simulations, each factor could modify network responses and illusory depth perceptions, with other factors fixed. With the binocular rivalry induced by interocular blur difference, the network can simulate illusory depth perceptions according with visual experiments. Our simulations show that stimuli with interocular blur differences could impact binocular rivalry in the network and could induce illusory depth perceptions compared to the normal stimuli. In section IV, a conclusion is given.

II. METHOD

A. THE STRUCTURE OF THE K-WTA NETWORK

In this section, a two-layer k-WTA spiking network is constructed as shown in Fig. 1. Because the depth perception of the moving trace of the visual stimulus is associated with the primary visual cortex, a k-WTA network is constructed to simulate primary visual cortical responses with excitatory and inhibitory neurons interconnected with each other. The network considers two populations of afferent neurons driven by monocular visual stimuli, respectively. The monocular and binocular excitatory neural activities in the primary visual cortex are simulated in the downstream network. There are two groups of monocular excitatory neurons receiving feedforward signals and competing with each other. To simulate binocular rivalry, the connective structure among monocular excitatory neurons is designed as introduced in [17]. Binocular excitatory neurons in the second layer have the lateral inter-connections with specific
connective probabilities as those in experiments and the previous network about the primary visual cortex [33], [38]. To build the network consistent with experimental observations, connections among neurons are designed randomly with these connective probabilities.

The first-layer network contains two populations of excitatory afferent neurons which can receive monocular visual information. Two populations consist of $N_L$ and $N_R$ afferent modeling neurons, respectively. The subscripts $L$ and $R$ stand for the left and right eyes. The received field of each afferent neuron is simulated as the Difference-of-Gaussian filter. Each group of afferent neurons would transfer the received visual information through Poisson spikes towards the corresponding second-layer excitatory monocular neurons. The received fields are located within the image on a grid to make the image covered, as introduced in the previous study [40].

The second-layer network comprises monocular excitatory neurons, binocular excitatory neurons and inhibitory neurons. There are two populations of monocular excitatory neurons. One group contains $K_L$ left-eye monocular neurons, receiving the feedforward inputs from the $N_L$ left-eye afferent neurons. Similarly, the other group contains $K_R$ right-eye monocular neurons receiving the inputs from $N_R$ right-eye afferent neurons. In our network, two groups contain the same number of monocular neurons ($N_L = N_R$) and no interconnections are designed within these two groups. With the same serial number, a right-eye monocular neuron and a left-eye monocular neuron are considered as a pair of monocular neurons. Two groups of monocular excitatory neurons generate lateral inputs towards the binocular excitatory neurons.

To simulate binocular rivalry, two groups of monocular neurons are assumed to inhibit each other by activating special inhibitory neurons in the network. For the left-eye monocular neurons, there are $J_L$ left-eye local inhibitory neurons which inhibit them and receive excitatory inputs from right-eye monocular neurons. Similarly, there are $J_R$ right-eye local inhibitory neurons. Each monocular neuron would excite the specific inhibitory neuron to inhibit the corresponding monocular neuron in the opposite group. Particularly, a left-eye monocular neuron inhibits the corresponding right-eye monocular neuron by activating a specific right-eye local inhibitory neuron. Besides the local inhibitory neurons, there are total inhibitory neurons. A left-eye monocular neuron inhibits all the right-eye monocular neurons by activating a special inhibitory right-eye total inhibitory neuron. In our network, each right-eye (left-eye) monocular neuron would activate a corresponding left-eye (right-eye) local inhibitory neuron and a total inhibitory neuron. In the second-layer network, there are $J_{RT}$ right-eye local inhibitory neurons, $J_{RT}$ right-eye total inhibitory neurons, $J_L$ left-eye local inhibitory neurons and $J_{LT}$ left-eye total inhibitory neurons. Due to the connective structure among monocular excitatory and inhibitory neurons, it is assumed that $J_L = J_{LT} = J_R = J_{RT} = N_L = N_R$ in the network.
In the network, the temporal membrane potential of the $k_L$ th left-eye monocular excitatory neuron is given as the sum of its received inputs:

$$u_{k_L}^t = \sum_{n_L} w_{k_L,n_L} \tilde{x}_{n_L} - \sum_{j_L = 1}^{L_L} v^{E_j} \tilde{y}_{j_L} - \sum_{j_L = 1}^{L_L} v^{E_j} \tilde{y}_{j_L} + b_{k_L}^t, \quad (1)$$

where $u_{k_L}^t$ is the membrane potential of the $k_L$ th left-eye monocular excitatory neuron at time $t$, $w_{k_L,n_L}$ denotes the instantaneous excitatory postsynaptic potentials (EPSP) coming from the $n_L$ th afferent neuron. $w_{k_L,n_L}$ is the corresponding plastic feedforward weight. In the network, all the excitatory feedforward weights are plastic and limited in $(0, 1)$.

The indices of all the left-eye inhibitory neurons connecting to the $k_L$ th left-eye monocular neuron are...
marked as the set $J_L$. $v^{EF}_{\beta_{ik}}$ is the instantaneous inhibitory postsynaptic potential (IPSP) coming from the $j_i$ th left-eye local inhibitory neuron among $J_L$. Similarly, the set $J_{LT}$ represent the all the left-eye total inhibitory neurons connecting to the $k_L$ th left-eye monocular neuron. $v^{EF}_{\gamma_{il}}$ is the IPSP from the left-eye total inhibitory neuron. In the network, it is assumed that all the weights from inhibitory neurons to excitatory neurons are fixed, denoted as the common parameter $v^{EF}$. With the mean strength of the excitatory weights, $v^{EF} = 0.5$ is used in our simulations. The parameter $b_k^z$, sampled from the uniform distribution of $(0,1)$, controls the neural excitability of this neuron. $\tilde{x}_{m_i}$, $\tilde{y}_{il}$ and $\tilde{y}_{il_{er}}$ are the temporal synaptic traces of corresponding neurons at time $t$, calculated with their spiking times. For instance, the $n_i$ th afferent neuron is assumed to generate its spikes at $\{t^{(1)}_n, t^{(2)}_n, t^{(3)}_n, \cdots\}$, the temporal value of $\tilde{x}_{m_i}$ at time $t$ is presented by the synaptic responding kernel $\epsilon$ as:

$$\tilde{x}_{m_i} = \sum_{s} \epsilon(t - t^{(s)}_n) \epsilon(s) = \begin{cases} e^{-\tau_k t} & \text{if } 0 \leq s < t, \\ 0 & \text{otherwise} \end{cases},$$

with $t^{(s)}_n$ is the time of the $s$ th spike of this neuron. The rise-time constant is $\tau_r = 1$ timestep and the fall-time constant is $\tau_f = 10$ timesteps. Using the same kernel $\epsilon$, $\tilde{y}_{il}$ and $\tilde{y}_{il_{er}}$ can be calculated, similarly.

In this paper, the temporal neural active state indicates whether a neuron generates a spike at the timestep $t$ or not. The temporal neural active state describes the neural spiking activities. The timestep, standing for 1 millisecond in experiments, is small enough that each neuron is assumed to emit only one spike at most within each timestep. With this assumption, the temporal active state of the $k_i$ th left-eye monocular excitatory neuron in the second layer is expressed as a variable $z_{d_k} \in \{0,1\}$.

The distribution of $z_{d_k}$ depends on its temporal membrane potential as:

$$p\left(z_{d_k} \mid u_{d_k}^z\right) = \frac{\exp\left(u_{d_k}^z z_{d_k}\right)}{\sum_{z_{d_k} \in \{0,1\}} \exp\left(u_{d_k}^z z_{d_k}'\right)},$$

where $z_{d_k}$ is the temporal neural active state, $u_{d_k}^z$ is the temporal membrane potential. $z_{d_k}'$ in the denominator takes all the possible values of the neural temporal active state for normalization. The distribution of the neural active state depends on the temporal membrane potential $u_{d_k}^z$. The larger value of $u_{d_k}^z$ leads to a larger spiking probability. With the temporal membrane potential of 0, the $k_L$ th left-eye monocular neuron has the probability of 0.5 to generate a spike. After generating a spike, the $k_L$ th left-eye monocular neuron has a refractory period of 10 timesteps, during which it can’t emit another spike again and $z_{d_k}$ is set to be 0 [33].

The temporal membrane potential and the active state of each right-eye monocular excitatory neuron can be given similarly.

The $K$ binocular excitatory neurons receive lateral inputs from two groups of monocular neurons, as well as another group of $J$ binocular inhibitory neurons. These binocular inhibitory neurons are interconnected with each other and receive lateral excitatory inputs from each other. The temporal membrane potential of the $k$ th binocular neuron is expressed as the sum of its received inputs:

$$u_{d_k} = \sum_{k_l \in J_L} v_{d_k d_{kl}} \tilde{z}_{d_{kl}} + \sum_{k_r \in J_{LT}} v_{d_k d_{kr}} \tilde{z}_{d_{kr}} + \sum_{k_r \in J_{LT}} v_{d_k d_{r}} \tilde{z}_{d_{r}} - \sum_{j_{kl} \in J_L} v_{d_k j_{kl}} \tilde{y}_{d_{kl}} + b_k^z$$

where $u_{d_k}$ is the temporal membrane potential of the $k$ th binocular neuron at time $t$. $v_{d_k d_{kl}}$, $v_{d_k d_{kr}}$, and $v_{d_k d_{l}}$ are the instantaneous lateral EPSPs coming from the $k_L$ th left-eye monocular neuron, the $k_R$ th right-eye monocular neuron and the $k'$ th binocular neuron, respectively. $v_{d_k j_{kl}}$ and $v_{d_k d_{l}}$ are plastic lateral connective weights among second-layer excitatory neurons. In the network, the plastic lateral
excitatory weights are limited in \((0,1)\). Indices of all the binocular inhibitory neurons connecting to the \(k\) th binocular neuron are marked as the set \(J\). With a special connective probability \(p_{ku} = 0.6\), an inhibitory neuron is randomly decided to have a connection towards each binocular excitatory neuron or not [33]. Through the random decision of connections, the structure of the network could be consistent with circuits in the primary visual cortex observed in experiments [38]. \(v^E_\theta\hat{y}_j\) is the instantaneous lateral IPSPs from the \(j\) th binocular inhibitory neuron among \(J\). The parameter \(h_\theta^j\), sampled from the uniform distribution of \((0,1)\), controls the neural excitability of this neuron. \(z_{\theta k}, z_{\theta k}^\prime\) and \(\hat{y}_j\) are the temporal synaptic traces of corresponding neurons at time \(t\), given by the synaptic responding kernel \(\epsilon\) in (2).

The temporal active state of the \(k\) th binocular neuron is also expressed as a variable \(z_{\theta k} \in \{0,1\}\). Its distribution depends on the neural membrane potential as:
\[
p(z_{\theta k} | u_{\theta k}) = \frac{\exp(u_{\theta k} z_{\theta k}^\prime)}{\sum_{z_{\theta k}^\prime \in \{0,1\}} \exp(u_{\theta k} z_{\theta k}^\prime)},
\]
where \(z_{\theta k}, z_{\theta k}^\prime\) are the temporal neural active states, \(u_{\theta k}\) is the temporal membrane potential. In the network, the binocular excitatory neurons have a given refractory period of 10 timesteps [33].

To the stimuli, the network could compare the vector of temporal second-layer excitatory neural active states \((z_{t1},\ldots,z_{tK}, z_{t1}^\prime,\ldots,z_{tK}^\prime, z_{t1},\ldots,z_{tK})\) with the clustering sets at each timestep to simulate the identification of visual stimuli. At the timestep \(t\), the temporal vector of excitatory neural active states reflects the neural spikes generated by left-eye monocular neurons, right-eye monocular neurons and binocular neurons. The basic method to identify visual stimuli is introduced briefly. If \(S\) is the number of stimuli, the identification is expressed by the variable as the temporal action \(a_t \in \{1,2,\ldots,S\}\) at time \(t\). For instance, with the given \(s\) th outside stimulus, the identification is correct if \(a_t = s\) or incorrect otherwise. The correct identification would give the network the temporal reward as \(r_t = 1\). The incorrect identification would give the network the temporal reward as \(r_t = 0\). To each stimulus, the clustering set consists of the second-layer excitatory neural active states which simulate the previous correct identifications. These neural active states from correct identifications are collected as the comparison for each novel vector of neural responses in the next. For all the stimuli, the clustering sets are collected as \(C = \{c_s, s = 1\ldots,S\}\). \(c_s\) is a matrix with the limited size, containing previous second-layer excitatory neural responses to the \(s\) th stimulus. The temporal rewards control the modifications of both plastic connective weights and clustering sets, which are introduced in section II-B and section II-C.

The second-layer inhibitory neurons can generate their spikes through the stochastic spiking model. The connections among monocular excitatory and inhibitory neurons are designed as introduced in [17]. Through a frequency-current curve, the \(j_L\) th left-eye local inhibitory neuron has its temporal firing rate as:
\[
\rho^L_{j_L} = \sigma_{\text{rect}}(u^L_{j_L}) = \sum_{k_L} v^{E}_{k_L} z_{k_L} \cdot \sigma(k^L_{k_L}, j_L),
\]
where \(\rho^L_{j_L}\) is the time-varying firing rate of this neuron and \(u^L_{j_L}\) is the temporal membrane potential. \(\sigma_{\text{rect}}(\cdot)\) is the linear rectifying function. \(\sigma_{\text{rect}}(u^L_{j_L}) = u^L_{j_L}\) for \(u^L_{j_L} \geq 0\) and \(\sigma_{\text{rect}}(u^L_{j_L}) = 0\) otherwise. The temporal membrane potential of the \(u^L_{j_L}\) th left-eye local inhibitory neuron is given as the sum of all its received excitatory signals from the corresponding right-eye monocular neuron. \(z_{k_L}\) is the
temporal synaptic trace at time $t$. $v^{IE}$ is the fixed and common weight from excitatory neurons to inhibitory neurons. $\sigma(\cdot)$ implies the corresponding relationship between the monocular neurons and local inhibitory neurons. $\sigma(k_R, j_L) = 1$ if $k_R = j_L$ and $\sigma(k_R, j_L) = 0$ otherwise. This local inhibitory neuron generates the spike trains through Poisson process with $\rho_{t_{L}}^{\upsilon}$ as the temporal responding strength. Its temporal active state at time $t$ is expressed as a variable $y_{t_{L}} \in \{0,1\}$. After a spike, the $j_L$ th left-eye local inhibitory neuron enters a refractory period of 3 timesteps, during which it can’t emit another spike again and is set as 0 [33]. The temporal membrane potentials and instantaneous firing rates of right-eye local inhibitory neurons are given similarly.

The $j_{LT}$ th left-eye total inhibitory neuron has its temporal firing rate as:

$$u_{t_{LT}}^{\upsilon} = \rho_{t_{LT}}^{\upsilon} = \sigma_{t_{LT}}^{\upsilon}(u_{t_{L}}^{\upsilon},) = \sum_{k_L} v^{IE} \tau_{L}z_{k_L} \cdot \sigma(k_R, j_{LT}) - \sum_{i_L} v^{IE} \gamma_{i_L} \cdot \sigma(j_R, j_{LT}),$$

where $\rho_{t_{LT}}^{\upsilon}$ is the time-varying firing rate of this neuron and $u_{t_{LT}}^{\upsilon}$ is the temporal membrane potential. The temporal membrane potential of the $j_{LT}$ th left-eye total inhibitory neuron is calculated as the sum of all its received lateral inputs. It receives excitatory signals from the corresponding right-eye monocular neuron and inhibitory signals from the specific right-eye local inhibitory neuron. The functions $\sigma_{t_{LT}}^{\upsilon}(\cdot)$ and $\sigma(\cdot)$ are introduced in (6). This inhibitory neuron generates the spike trains through Poisson process with $\rho_{t_{LT}}^{\upsilon}$ as the temporal responding strength. Its temporal active state is expressed as a variable $y_{t_{LT}} \in \{0,1\}$. The $j_{LT}$ th left-eye total inhibitory neuron has a refractory period of 3 timesteps [33]. The temporal membrane potentials and instantaneous firing rates of right-eye total inhibitory neurons are given similarly.

The instantaneous firing rate and the temporal membrane potential of the $j$ th binocular inhibitory neuron can be expressed as:

$$\rho_{t}^{\upsilon} = \sigma_{t_{BT}}^{\upsilon}(u_{t}^{\upsilon}, u_{t}^{\upsilon}) = \sum_{l_{LT}} v^{IE} \tau_{L}z_{l_{LT}} - \sum_{j_{LT}} v^{IE} \gamma_{j_{LT}} y_{j_{LT}},$$

where $\rho_{t}^{\upsilon}$ is the time-varying firing rate and $u_{t}^{\upsilon}$ is the temporal membrane potential. $z_{l_{LT}}$ and $y_{j_{LT}}$ are the temporal synaptic traces from corresponding neurons. In the network, the $j$ th inhibitory neuron is decided to have the connection from each lateral excitatory (inhibitory) neuron with the probability of $p^{IE} = 0.575 (p_{r} = 0.55)$ [33]. Indices of all the excitatory (inhibitory) neurons connecting to the $j$ th binocular inhibitory neuron are marked as the set $\varphi_{j}(\sigma_{j})$. $v^{IE}(\upsilon_{j})$ is the fixed weight from the lateral excitatory (inhibitory) neurons to the inhibitory ones. Similar with $v^{IE}$ in (1), $v^{IE}(\upsilon_{j})$ is set to be 0.5. The $j$ th binocular inhibitory neuron generates its spikes through Poisson process with temporal responding strength of $\rho_{t}^{\upsilon}$. Its temporal active state is expressed as $y_{t} \in \{0,1\}$. Its absolute refractory period is 3 timesteps [33].

The excitatory afferent neurons are designed to transmit visual information towards the second-layer network through Poisson spikes. The temporal stimulus within the receptive field determines the responses of an afferent neuron. In the network, the receptive field of the $n_{L}$ th left-eye afferent neuron is modeled with the Difference-of-Gaussians filter [41]:

$$f_{n_{L}}(i, i_{n_{L}}) = \left[ f_{n_{L}}^{\upsilon}(i, i_{n_{L}}) - \phi \cdot f_{n_{L}}(i, i_{n_{L}}) \right],$$

$$f_{n_{L}}^{\upsilon}(i, i_{n_{L}}) = \exp \left\{-\frac{\left[ (i - i_{n_{L,1}})^{2} + (i - i_{n_{L,2}})^{2} \right]}{\sigma_{n_{L}}^{2}} \right\},$$

$$f_{n_{L}}(i, i_{n_{L}}) = \exp \left\{-\frac{\left[ (i - i_{n_{L,1}})^{2} + (i - i_{n_{L,2}})^{2} \right]}{\sigma_{n_{L}}^{2}} \right\},$$

$$f_{n_{L}}(i, i_{n_{L}}) = \exp \left\{-\frac{\left[ (i - i_{n_{L,1}})^{2} + (i - i_{n_{L,2}})^{2} \right]}{\sigma_{n_{L}}^{2}} \right\}.$$
where \( \tilde{r} = (i_j, l_j) \) indicates the position of a pixel in the image, \( \hat{\mathbf{r}}_n = (i_{n,t}, l_{n,z}) \) is the center of the received field. The center and surround Gaussian functions of the \( n_j \) th left-eye afferent neuron are \( f_{n_j}^c(\cdot) \) and \( f_{n_j}^s(\cdot) \). The spatial radii of two Gaussian functions are \( \sigma_{n_j}^c \) and \( \sigma_{n_j}^s \). \( \phi \) is the ratio between two functions. To cover and scan the image cooperatively, the afferent neural received fields are positioned on a grid with a spacing distance of 2 pixels, similar to the way in [40].

In our simulations, a timestep stands for 1 millisecond in experiments. The network contains two groups of 100 afferent neurons, 80 left-eye monocular neurons, 80 right-eye monocular neurons and 80 binocular neurons. Besides, the downstream network has 80 left-eye local inhibitory neurons, 80 left-eye total inhibitory neurons, 80 right-eye local inhibitory neurons, 80 right-eye total inhibitory neurons and 40 binocular inhibitory neurons.

**B. The Probabilistic Model for the Network**

Because the second-layer excitatory neurons receive both feedforward and lateral signals, the second-layer neural responses in the current timestep depends on the second-layer neural responses in the previous timestep. It means that the distribution of current hidden state probabilistic model has to consider the influence of the previous hidden state. In this way, the stochastic second-layer neural responses are simulated through the Hidden Markov model (HMM) [34].

In this spiking network, the instantaneous feedforward traces from two eyes perform as the temporal observed variables at time \( t \), \( \mathbf{O}_t = (\tilde{x}_L^t, \tilde{x}_R^t) \) with \( \tilde{x}_L^t = (\tilde{x}_{i_1}^t, \ldots, \tilde{x}_{i_N}^t) \) and \( \tilde{x}_R^t = (\tilde{x}_{i,1}^t, \ldots, \tilde{x}_{i,N}^t) \). All the second-layer neural active states, the synaptic traces and all the temporal reward constitute the temporal hidden variables as \( \mathbf{h}_t = (x_{i_1}^L, x_{i_2}^L, \ldots, x_{i_{N_2}}^L, x_{i_1}^R, x_{i_2}^R, \ldots, x_{i_{N_2}}^R, \tilde{y}_{i_1}^L, \tilde{y}_{i_2}^L, \ldots, \tilde{y}_{i_{N_2}}^L, \tilde{y}_{i_1}^R, \tilde{y}_{i_2}^R, \ldots, \tilde{y}_{i_{N_2}}^R, r_{i_1}, r_{i_2}, \ldots, r_{i_{N_2}}) \).

The plastic parameters consists of feedforward weights \( \{w_{ki\ell}, k = 1, \ldots, N_L, i = 1, \ldots, N_L\} \), lateral weights \( \{v_{kki}, k = 1, \ldots, N_L, k = 1, \ldots, N_L\} \), lateral weights \( \{v_{kki}, k = 1, \ldots, N_L, k = 1, \ldots, N_L\} \) and the clustering sets to stimuli \( \{c_s, s = 1, \ldots, S\} \). These parameters requiring modifications are collected as \( \Theta \). If \( t \) is the number of occurred timesteps, \( \mathbf{O}_t = (\tilde{o}_{i_1}, \tilde{o}_{i_2}, \ldots, \tilde{o}_{i_{N_2}}) \) are the observations and \( \mathbf{h}_{t-1} = (\tilde{h}_{i_1}, \tilde{h}_{i_2}, \ldots, \tilde{h}_{i_{N_2}}) \) are the hidden states up to \( t - 1 \), the stochastic dynamics of the k-WTA network implements a forward sampling process and samples a new hidden active state \( \tilde{h}_i \) forward in time based on \( \tilde{O}_t \) and \( \tilde{H}_{t-1} \) [34]. In the sampling of the hidden active state \( \tilde{h}_i \), the second-layer excitatory and inhibitory neural spikes are generated according to (3-8), which are similar to our previous study [39]. With the temporal second-layer excitatory neural spikes, the network generates the temporal action \( a_t \) to identify stimuli and obtains a temporal reward as a binary random variable, \( r_t \in \{0, 1\} \). The detail generations of \( a_t \) and \( r_t \) will be introduced in the section II-C.

With the network dynamics updated at each timestep, modifications of plastic weights depend on the temporal reward. If \( r_t = 1 \) at time \( t \), it means the temporal identification of the stimulus is correct and plastic connective weights would be modified at this moment. Up to time \( t \), a \( t \)-step sequence of observed variables to the stimulus \( Sti \) is marked as \( \tilde{O}_t = (\tilde{o}_{i_1}, \tilde{o}_{i_2}, \ldots, \tilde{o}_{i_{N_2}}) \), a \( t \)-step sequence of hidden variables is \( \tilde{H}_t = (\tilde{h}_{i_1}, \tilde{h}_{i_2}, \ldots, \tilde{h}_{i_{N_2}}) \). The objective function is expressed with \( Sti \), \( \tilde{O}_t \), and \( \tilde{H}_t \) as
[42]:

\[ I(\tilde{H}_i,\text{Sti},r_i = l) = \int \frac{p(\text{Sti},\tilde{H}_i,\tilde{O}_i,|r_i = l)}{p(\tilde{H}_i)} \log \frac{p(\text{Sti},r_i = l|\tilde{H}_i)}{p(\text{Sti},r_i = l)} \]

\[ = \int [d\tilde{O}_i,d\tilde{H}_i] p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l) \cdot \log p(\text{Sti},r_i = l|\tilde{H}_i) \]

\[ - \log p(\tilde{H}_i) + log p(\text{Sti}|r_i = l) \]

\[ = \int [d\tilde{O}_i,d\tilde{H}_i] p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l) \cdot \log p(\text{Sti},r_i = l|\tilde{H}_i) \]

\[ - \log p(\tilde{H}_i) + log p(\text{Sti}|r_i = l) \]

\[ + \log p(\tilde{O}_i,\tilde{H}_i) \]

where \( p(\text{Sti},r_i = l|\tilde{H}_i) \) is the distribution of the stimulus based on the induced hidden variables, \( p(\tilde{O}_i,\tilde{H}_i) \) is the joint distribution of observed variables and hidden variables, \( p(\text{Sti}|r_i = l), p(\tilde{O}_i), p(\tilde{H}_i) \) are the parameter-independent distributions.

The desired distributions are the ones that minimize the objective function. In our simulations, the stimulus is inferred with the induced responses of the network. It means that \( p(\text{Sti},r_i = l|\tilde{H}_i) \) is assumed as the Dirac delta function which is independent of parameters in this paper.

The joint distribution \( p(\tilde{O}_i,\tilde{H}_i) \) describes the induced activities of our network and is controlled by the connective weights in the model. Considering the influence of plastic weights and clustering sets, the joint distribution is re-expressed as \( p(\tilde{O}_i,\tilde{H}_i;\Theta) \). Neglecting the parameter-independent terms, the objective function also can be approximated as \( \langle -\log p(\tilde{O}_i,\tilde{H}_i;\Theta) \rangle_{p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l)} \). Minimization of the objective function in (10) is equivalent to maximization of \( \langle \log p(\tilde{O}_i,\tilde{H}_i;\Theta) \rangle_{p(\tilde{O}_i,\tilde{H}_i,|r_i = l)} \). However, this term can not be calculated directly because the network can not generate its responses from the distribution of \( p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l;\Theta) \). To overcome this problem, \( \langle \log p(\tilde{O}_i,\tilde{H}_i;\Theta) \rangle_{p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l)} \) is expressed in another equivalent form as:

\[ \langle r_i \cdot \log p(\tilde{O}_i,\tilde{H}_i;\Theta) \rangle_{p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l)} \]

\[ = \int d\tilde{O}_i,d\tilde{H}_i \left[ 1 \cdot \log p(\tilde{O}_i,\tilde{H}_i;\Theta) \cdot p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l) \right] \]

\[ + \int d\tilde{O}_i,d\tilde{H}_i \left[ 0 \cdot \log p(\tilde{O}_i,\tilde{H}_i;\Theta) \cdot p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = 0) \right] \]

\[ = \int d\tilde{O}_i,d\tilde{H}_i \left[ \log p(\tilde{O}_i,\tilde{H}_i;\Theta) \cdot p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l) \right] \]

\[ = \langle \log p(\tilde{O}_i,\tilde{H}_i;\Theta) \rangle_{p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l)} \cdot \]

The re-expression in Equation (11) is similar to the method in [42]. It means that, for a sequence of dynamics up to time \( t \), the network modifies its parameters to maximize \( \langle r_i \cdot \log p(\tilde{O}_i,\tilde{H}_i;\Theta) \rangle_{p(\text{Sti},\tilde{O}_i,\tilde{H}_i,|r_i = l)} \). For \( r_i = l \), there are several sub-sequences of dynamics contributing to the temporal identification, with different lengths expressed as \( T \in \{1,\cdots,t\} \). With a given length of \( T \), the observed variables and hidden variables up to time \( t \) can be expressed as \( \tilde{O}_T = (\tilde{o}_{T-1},\cdots,\tilde{o}_1), \tilde{H}_T = (\tilde{h}_{T-1},\cdots,\tilde{h}_1) \). The
contribute of this pair of $T$-step observed variables and hidden variables to $r_i$ is assumed as a rate

$$\alpha(T) = (1 - \gamma)^{T-1}$$

with the discount factor $\gamma \in (0, 1)$. It should be noticed that $\sum_{T=1}^{\infty} \alpha(T) = 1$. It follows the assumption that, the longer sequence of dynamics could be affected by previous information and make less contribution to the temporary reward $r_i = 1$.

With all the sequences of dynamics considered, the likelihood function $L(\Theta)$ is designed based on the objective function as:

$$L(\Theta) = r_i \cdot \log p(\tilde{O}_r, \tilde{H}_r; \Theta) \prod_{t=1}^{T} p\left(r_i, S_t, \tilde{O}_r, \tilde{H}_r; \Theta\right)$$

The joint distributions $p\left(r_i, S_t, \tilde{O}_r, \tilde{H}_r; \Theta\right)$ and $p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right)$ are factorized under the assumption of Hidden Markov model:

$$p\left(r_i, S_t, \tilde{O}_r, \tilde{H}_r; \Theta\right) = p\left(r_i, S_t; \tilde{O}_r, \tilde{H}_r; \Theta\right) p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right)$$

$$= \prod_{t=1}^{T} p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) p\left(r_i, S_t; \tilde{O}_r, \tilde{H}_r; \Theta\right)$$

$$= \prod_{t=1}^{T} p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) p\left(r_i, S_t; \tilde{O}_r, \tilde{H}_r; \Theta\right)$$

With the factorization, the likelihood function $L(\Theta)$ is expressed as:

$$L(\Theta) = r_i \cdot \sum_{T=1}^{\infty} \alpha(T) \cdot \left[ \log p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) \right]$$

The network optimizes its parameters through the stochastic online variant of the Expectation-Maximization algorithm, in which the expectation is estimated during the E-step by sampling a finite set of samples. In our simulations, the E-step is to estimate the expectation with a single sample from $p\left(r_i, S_t, \tilde{O}_r, \tilde{H}_r; \Theta\right)$ [34]. The likelihood function $L(\Theta)$ is then approximated as:

$$L(\Theta) = \sum_{T=1}^{\infty} \alpha(T) \cdot r_i \cdot \left[ \sum_{t=1}^{T} \log p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) \right]$$

For a given $t'$, the term $\log p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right)$ multiplies different rates $\alpha$ in the expectation. The approximation of the likelihood function $L(\Theta)$ is rearranged as:

$$L(\Theta) = r_i \cdot \sum_{T=1}^{\infty} \log \left[ p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) \right] \cdot \sum_{T=1}^{\infty} \alpha(T)$$

$$= r_i \cdot \sum_{T=1}^{\infty} \log \left[ p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) \right] \cdot \sum_{T=1}^{\infty} \alpha(T)$$

$$= r_i \cdot \sum_{T=1}^{\infty} \log \left[ p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) p\left(\tilde{O}_r, \tilde{H}_r; \Theta\right) \right] \cdot \sum_{T=1}^{\infty} \alpha(T)$$

In the M-step, the network optimizes its parameters towards the directions which are given by the partial derivatives of $L(\Theta)$. Specially, the partial derivative respects to the feedforward weight $w_{k,n}$ is:
The distribution \( p(\tilde{o}, \tilde{h}_i; \Theta) \) can be expressed as:

\[
p(\tilde{o}, \tilde{h}_i; \Theta) = p(\tilde{x}_i^L, \tilde{x}_i^R | z_i^L, z_i^R; \Theta) = \prod_{n_z} p(\tilde{x}_{i,n_z} | z_i^L; \Theta) \prod_{n_g} p(\tilde{x}_{i,n_g} | z_i^R; \Theta)
\]

\[
p(\tilde{x}_{i,n_z} | z_i^L; \Theta) = \frac{\exp\left(\sum_{k_z} w_{k_z,n_z} \tilde{x}_{i,n_z} z_{i,k_z}\right)}{\int \exp\left(\sum_{k_z} w_{k_z,n_z} \tilde{x}_{i,n_z} z_{i,k_z}\right) dx_{i,k_z}}
\]

\[
p(\tilde{x}_{i,n_g} | z_i^R; \Theta) = \frac{\exp\left(\sum_{k_g} w_{k_g,n_g} \tilde{x}_{i,n_g} z_{i,k_g}\right)}{\int \exp\left(\sum_{k_g} w_{k_g,n_g} \tilde{x}_{i,n_g} z_{i,k_g}\right) dx_{i,k_g}}
\]

where \( \tilde{x}_i^L = (\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,n_z})^T \), \( \tilde{x}_i^R = (\tilde{x}_{i,1}, \ldots, \tilde{x}_{i,n_g})^T \). In order to make these learning rules more biologically plausible, this paper makes the approximations similarly to a previous study [43]. In this approximation, the modification of each weight is assumed to depend on its current value, the pre- and postsynaptic responses and the temporal reward. For instance, the partial derivative with respect to the feedforward weight \( w_{k,n_z} \) can be approximated as:

\[
\frac{\partial}{\partial w_{k,n_z}} L(\Theta) = r \sum_{i=1}^{p^{(t)}} \frac{\partial}{\partial w_{k,n_z}} \log p(\tilde{x}_{i,n_z} | z_i^L; \Theta)
\]

\[
= r \sum_{i=1}^{p^{(t)}} \frac{\partial}{\partial w_{k,n_z}} \left[ I - \frac{1}{w_{k,n_z}} + \frac{1}{\exp(w_{k,n_z}) - 1} \right]
\]

Similarly, the modification of \( w_{k,n_g} \) can be expressed and approximated as:

\[
\frac{\partial}{\partial w_{k,n_g}} L(\Theta) = r \sum_{i=1}^{p^{(t)}} \frac{\partial}{\partial w_{k,n_g}} \log p(\tilde{x}_{i,n_g} | z_i^R; \Theta)
\]

\[
= r \sum_{i=1}^{p^{(t)}} \frac{\partial}{\partial w_{k,n_g}} \left[ I - \frac{1}{w_{k,n_g}} + \frac{1}{\exp(w_{k,n_g}) - 1} \right]
\]

The modifications of the plastic lateral weights are based on the distribution of hidden variables in the second layer. The distribution \( p(\tilde{h}_i | \tilde{h}_{i-1}; \Theta) \) can be written as:

\[
p(\tilde{h}_i | \tilde{h}_{i-1}; \Theta) = p(\tilde{x}_i^L, \tilde{x}_i^R | z_i^L, z_i^R; \Theta)
\]

\[
= p(\tilde{x}_i^L | z_i^L; \Theta) p(\tilde{x}_i^R | z_i^R; \Theta)
\]

\[
= \prod_{n_z} p(\tilde{x}_{i,n_z} | z_i^L; \Theta) \prod_{n_g} p(\tilde{x}_{i,n_g} | z_i^R; \Theta)
\]

\[
p(\tilde{x}_{i,n_z} | z_i^L; \Theta) = \frac{\exp\left(\sum_{k_z} w_{k_z,n_z} \tilde{x}_{i,n_z} z_{i,k_z}\right)}{\int \exp\left(\sum_{k_z} w_{k_z,n_z} \tilde{x}_{i,n_z} z_{i,k_z}\right) dx_{i,k_z}}
\]

\[
p(\tilde{x}_{i,n_g} | z_i^R; \Theta) = \frac{\exp\left(\sum_{k_g} w_{k_g,n_g} \tilde{x}_{i,n_g} z_{i,k_g}\right)}{\int \exp\left(\sum_{k_g} w_{k_g,n_g} \tilde{x}_{i,n_g} z_{i,k_g}\right) dx_{i,k_g}}
\]

where \( z_i^L, z_i^R, z_i^L, z_i^L, z_i^R, z_i^RT, z_i, \tilde{z}_i^L, \tilde{z}_i^R, \tilde{z}_i^RT, \tilde{z}_i, \tilde{z}_i \) and \( \tilde{y}_i \) are vectors of active states of different second-layer neural populations, \( \tilde{x}_i^L, \tilde{x}_i^R, \tilde{x}_i^L, \tilde{x}_i^L, \tilde{x}_i^R, \tilde{x}_i^RT, \tilde{x}_i, \tilde{x}_i \) and \( \tilde{y}_i \) are vectors of second-layer neural synaptic traces. In the right side of Equation (21), the first term describes the joint distribution of neural responses, and the other eight terms determinately describe the neural synaptic traces in the second layers. The neural synaptic traces could be determined by the neural spiking trains and neural responding kernel in (2). The joint distribution of neural synaptic traces is assumed as parameter-independent in this paper. The joint distribution of the second-layer neural active states can be expressed as:

\[
p(\tilde{x}_i^L, \tilde{x}_i^R, \tilde{x}_i^L, \tilde{x}_i^L, \tilde{x}_i^R, \tilde{x}_i^RT, \tilde{x}_i, \tilde{y}_i \mid \tilde{z}_i, \tilde{z}_i \Theta) = p(\tilde{x}_i^L | \tilde{z}_i^L, \tilde{z}_i; \Theta) p(\tilde{x}_i^R | \tilde{z}_i^R, \tilde{z}_i; \Theta)
\]

\[
= p(\tilde{y}_i^{RT} | \tilde{x}_i^R, \tilde{y}_i; \Theta) p(\tilde{y}_i^{L} | \tilde{x}_i^L, \tilde{y}_i; \Theta)
\]

\[
= p(\tilde{y}_i^{RT} | \tilde{x}_i^R, \tilde{y}_i; \Theta) p(\tilde{y}_i^{L} | \tilde{x}_i^L, \tilde{y}_i; \Theta)
\]

All the excitatory active states in the second layer are
sampled from the softmax distributions. The plastic lateral weights are included in the distribution

\[ p(z_i^L, z_i^R, \xi_i, \bar{\xi}_i; \Theta) \]

which can be expressed as:

\[ p(z_i^L, z_i^R, \xi_i, \bar{\xi}_i; \Theta) = \prod_k p(z_{ik}^L, z_{ik}^R, \xi_{ik}, \bar{\xi}_{ik}; \Theta) \]

\[ p(z_{ik}^L, z_{ik}^R, \xi_{ik}, \bar{\xi}_{ik}; \Theta) = \frac{\exp(u_{ik}^\dagger z_{ik})}{\sum_{i'k} \exp(u_{i'k}^\dagger z_{i'k})}, \]

with \( u_{ik}^\dagger \) given in (4). The partial derivatives of \( L(\Theta) \) respects to \( v_{ik}, v_{ik}, \) and \( v_{ik}' \) can be expressed and approximated as:

\[ \frac{\partial}{\partial v_{ik}} L(\Theta) = r_i \sum_{j=1}^{J} v_{jk} \log p(z_{ik}^L, z_{ik}^R, \xi_{ik}, \bar{\xi}_{ik}; \Theta) \]

\[ = r_i \sum_{j=1}^{J} v_{jk} \left[ \log \left( \frac{1}{1 + \exp(-v_{ik})} \right) \right] \]

\[ \frac{\partial}{\partial v_{ik}} L(\Theta) = r_i \sum_{j=1}^{J} v_{jk} \log p(z_{ik}^L, z_{ik}^R, \xi_{ik}, \bar{\xi}_{ik}; \Theta) \]

\[ = r_i \sum_{j=1}^{J} v_{jk} \left[ \log \left( \frac{1}{1 + \exp(-v_{ik})} \right) \right] \]

\[ \frac{\partial}{\partial v_{ik}} L(\Theta) = r_i \sum_{j=1}^{J} v_{jk} \log p(z_{ik}^L, z_{ik}^R, \xi_{ik}, \bar{\xi}_{ik}; \Theta) \]

\[ = r_i \sum_{j=1}^{J} v_{jk} \left[ \log \left( \frac{1}{1 + \exp(-v_{ik})} \right) \right] \]

In all the modifications, \( r_i \) is the temporary reward which depends on the identifications of stimuli. In (19), (20) and (24), \( r_i \) could control the connective modifications in the network. It means that, compared to the k-WTA network in [33], connective weights are updated only at the timestep with the correct identification and \( r_i = 1 \). Besides, the plastic weights are limited in \((0, I)\) during modifications. The discount factor \( \gamma = 0.9 \) is used in simulations.

C. UNSUPERVISED IDENTIFICATIONS OF MOVING STIMULI

This paper identifies the stimuli with second-layer excitatory spikes through an unsupervised online method [44]. This method has no constraint conditions for the dimension of neural spikes and is suitable for the network with different sizes. Using this method, the network could identify the stimuli at each timestep and receive the reward to control connective modifications.

The unsupervised identification depends on the distance between two sets of vectors. Generally speaking, for any two sets of vectors \( A = (a_1, \ldots, a_n) \) and \( B = (b_1, \ldots, b_n) \), the energy distance \( e_{a_i, b_j}(A, B) \) between them can be calculated as:

\[ e_{a_i, b_j}(A, B) = \frac{n_i n_j}{n_i + n_j} \left( \frac{2}{n_i n_j} \sum_{a_j} a_j b_j \right) - \frac{1}{n_i n_j} \sum_{a_j} a_j \left( \frac{1}{n_i n_j} \sum_{b_j} b_j \right) \]

(25)

where \( n_i, n_j \) are the sizes of \( A \) and \( B \), \( a_i, b_j \) indicate the elements in two sets. With the energy distance, the likelihood of the distributions of two sets being equal could be estimated.

Based on the energy distance in (25), the identification of the set \( B \) includes two steps.

In the first step, with a given significance level \( \alpha \), a test for equal distributions of two sample sets \( A \) and \( B \) can be implemented by nonparametric resamples [45]. Then, \( A \) and \( B \) are merged into a pooled set as \( U = (A, B) = (a_1, \ldots, a_n, b_1, \ldots, b_n) \). With elements denoted in the same form, this pooled set can be expressed as \( U = (\bar{a}_1, \ldots, \bar{a}_n, \bar{a}_{n+1}, \ldots, \bar{a}_{2n}) \). With the size of \( U \) expressed as \( n_i = n_a + n_b \), a resample is made without replacement from the pooled set and get a random sample \((\bar{a}_1^{(i)}, \ldots, \bar{a}_n^{(i)}, \bar{a}_{n+1}^{(i)}, \ldots, \bar{a}_{2n}^{(i)})\). The superscript represents the serial number of this resample. This sample is then separated into two subsets as \( A^{(i)} = (\bar{a}_1^{(i)}, \ldots, \bar{a}_n^{(i)}) \) and \( B^{(i)} = (\bar{a}_{n+1}^{(i)}, \ldots, \bar{a}_{2n}^{(i)}) \). The energy distance between two subsets of this resample can be calculated and denoted as \( e_{a_i, b_j}^{(i)} \). In this way, \( R \)
resamples are made from the pooled set and obtain energy distances as \( \{e_{s,A}^{(i)}\}, \ldots, e_{s,A}^{(R)} \). For A and B, the null hypothesis \( H_0 \) is that the distributions generating them are equal. For the significance level \( \alpha \), the null hypothesis would be accepted if \( e_{s,A}^{(r)}(A,B) \) does not exceed \( 100 \times (1 - \alpha) \% \) of \( \{e_{s,A}^{(i)}\}, \ldots, e_{s,A}^{(R)} \) \([45]\). Put it in another way, a parameter can be estimated for A and B from the resamples as \( e_{\text{reus}}(A,B,\alpha) = (1 - \alpha) \times \max \{e_{s,A}^{(r)}(A,B)\} \), where \( r = 1, \ldots, R \).

In the second step, for the given B and several compared sets \( \{A_s, s = 1, \ldots, S\} \), the identification of B requires the further comparison. For \( \{A_s, s = 1, \ldots, S\} \), the set of parameters \( \{e_{\text{reus}}(A_s,\alpha,B)\}, s = 1, \ldots, S \} \) can be estimated. Then, a common parameter is set for \( \{A_s, s = 1, \ldots, S\} \) as \( e_{\text{reus}}(B,\alpha) = \min \{e_{\text{reus}}(A_s,\alpha,B)\}, s = 1, \ldots, S \} \). With additional R resamples, the likelihood between B and each of \( \{A_s, s = 1, \ldots, S\} \) can be estimated through as:

\[
p(A_s,B,\alpha) = \frac{1}{R} \sum_{r=1}^{R} I\left(e_{s,A}^{(r)}(A_s,B,\alpha) \leq e_{\text{reus}}(B,\alpha)\right), \quad (26)
\]

where \( p(A_s,B,\alpha) \) is the likelihood between B and \( A_s \), \( e_{s,A}^{(r)}(A_s,B,\alpha) \) is the energy distance from the \( r \) th resample of the pooled set \( (A_s,B) \), \( I(\cdot) \) is the indicator function. If \( e_{s,A}^{(r)}(A_s,B,\alpha) \leq e_{\text{reus}}(B,\alpha) \), \( I(e_{s,A}^{(r)}(A_s,B,\alpha) \leq e_{\text{reus}}(B,\alpha)) = 1 \). Otherwise, \( I(e_{s,A}^{(r)}(A_s,B,\alpha) \leq e_{\text{reus}}(B,\alpha)) = 0 \). For the significance level \( \alpha \), the set \( A_s \) is assumed to have the maximum of \( \{p(A_s,B,\alpha), s = 1, \ldots, S\} \). If \( p(A_s,B,\alpha) \) is larger than \( 1 - \alpha \), the distribution generating B is considered to be equal with the distribution generating \( A_s \).

If there are more than one set having the largest likelihood, the identification can be made through sampling with the equal probabilities. If the maximum of \( \{p(A_s,B,\alpha), s = 1, \ldots, S\} \) is not larger than \( 1 - \alpha \), the likelihood between B and each of \( \{A_s, s = 1, \ldots, S\} \) is not large enough. The identification of B can be made through sampling. \( p_r \) stands for the probability to accept that the distribution generating B is equal with the distribution generating \( A_r \). With likelihoods \( \{p(A_r,B,\alpha), s = 1, \ldots, S\} \), estimated, \( p_r \) can be expressed as:

\[
p_r = \exp\left\{p(A_r,B,\alpha)\right\}/\sum_{s} \exp\left\{p(A_s,B,\alpha)\right\}. \quad (27)
\]

The identification of B can be made by sampling from the distribution in Equation (27).

In this way, not only the likelihoods between B and \( \{A_s, s = 1, \ldots, S\} \) but also the identification of B can be performed in the unsupervised method.

In this paper, the network uses the unsupervised method to identify the outside stimuli. In our simulations, the second-layer excitatory neural spikes perform as the sensory data for identifications. The temporal second-layer excitatory neural spikes represent the current sensory responses induced by the outside stimulus. These induced responses are used to identify the stimuli. As introduced in section II-A, the previous identified sensory data have been collected into a clustering set for the latter identifications. Each clustering set represents a distribution of the second-layer excitatory sensory responses from the correct identifications to a stimulus. To identify which stimulus inducing the current second-layer responses, the likelihood between the current second-layer responses and each clustering set could be calculated based on the energy distance \([44]\).

For the non-blank vector of second-layer excitatory spikes \( \{\xi_{t,1}, \ldots, \xi_{t,K}, \xi_{t,2}, \ldots, \xi_{t,K}, \xi_{t,3}, \ldots, \xi_{t,K}\} \) at time \( t \) and the common significance level \( \alpha \), the energy distances \( \{e(s,\alpha,t), s = 1, \ldots, S\} \) between second-layer excitatory spikes and clustering sets, as well as the parameters \( \{e_{\text{reus}}(s,\alpha,t), s = 1, \ldots, S\} \), could be estimated.
\{c_{s}, s = 1, \ldots, S\} are the clustering sets denoted in section II-A. In our simulations, a common received parameter is set for all the clustering sets as

\[ e_{\text{receive}}(\alpha, t) = \min \{e_{\text{receive}}(c_{s}, \alpha, t), s = 1, \ldots, S\}. \]

Then, \( p(c_{s}, \alpha, t) = p(e(c_{s}, \alpha, t) \leq e_{\text{receive}}(\alpha, t)), s = 1, \ldots, S\) could be estimated for the sets \( \{c_{s}, s = 1, \ldots, S\} \). With the common \( \alpha \) and \( e_{\text{receive}}(\alpha, t) \), these probabilities represent the likelihoods between the second-layer excitatory responding vector \( (z_{1}, \ldots, z_{K}, z_{2}, \ldots, z_{K}, z_{1}, \ldots, z_{K}) \) and all the clustering sets.

The temporal action \( a_{t} \) performs as the identification of the stimulus at time \( t \). If the maximal value of \( \{p(c_{s}, \alpha, t), s = 1, \ldots, S\} \) exceeds a predefined value \( 1 - \alpha \), the second-layer excitatory responding vector is considered to be quite similar to the special clustering set with the largest likelihood. If the maximal value of \( \{p(c_{s}, \alpha, t), s = 1, \ldots, S\} \) is smaller than \( 1 - \alpha \), the estimation is generated through sampling from the softmax distribution. The softmax distribution of \( a_{t} \) to choose the \( s' \) th clustering set is defined as:

\[ p(a_{t} = s') = \exp(p(c_{s}, \alpha, t))/\sum_{s} \exp(p(c_{s}, \alpha, t)). \]  

(28)

To update clustering sets, the temporal action \( a_{t} \) is compared with the stimulus and the temporal reward \( r_{t} \) is obtained. If \( r_{t} = 1 \), the corresponding clustering set is updated in a First-In-First-Out (FIFO) manner. In each update, the second-layer excitatory responding vector would be added as the novel component to the end of the clustering set. After adding, if the size of the clustering set is larger than \( n_{\text{cluster}} \), the redundant components are deleted from the beginning of the set.

Besides, at the end of each simulation, all the clustering sets will be compared with each other for merging. The two clustering sets will be merged into one set if the likelihood exceeds a predefined value \( p_{\text{max}} \), which is similar to a previous study [44]. To merge two clustering sets, the size of the merged set is designed as the larger one between the sizes of two clustering sets. Then, a raw merged set is generated through the random arrangement and combination of all the components of two clustering sets. With the size given, the merged set is derived from the raw merged set through the systematic sampling. After that, this merged set performs as the clustering set of one stimulus while the clustering set of the other stimulus is set to be blank. If a stimulus has the blank clustering set at the beginning of a simulation, the first temporal obtained sets of all the images will be merged into its clustering set. Particularly, the initial clustering sets for all the stimuli are blank at the beginning of the learning. These clustering sets are supplemented with the second-layer neural responses in the first learning simulation without any special conditions. They would have unsupervised modifications according to identifications in the next.

Each stimulus is presented to the network for \( T_{\text{pre}} \) timesteps. In each learning simulation, the action and the reward are set to be 0 at the beginning. If the previous reward is 0, it means the previous temporal identification is not correct and the temporal action will be sampled at the novel timestep. If the previous reward is 1, it means that the estimation is correct and the temporal action will be hold and not necessary to be resampled. At each timestep, the connective weights and clustering sets will be updated according to the temporal reward. At the end of each simulation, each pair of different clustering sets will compare with each other to decide whether to merge or not.

In our simulations, \( T_{\text{pre}} = 20 \) timestep, \( p_{\text{max}} = 0.8 \), \( n_{\text{cluster}} = T_{\text{pre}} \), \( \alpha = 0.05 \), \( R = 50 \).
D. Simulations of Depth Perception

For depth perception, the event-based method depends on the responses of modeling neurons and simulates the 3D reconstructions of objects [46]–[48]. This method uses the pair of retinal cameras to simulate binocular vision. The cameras detect the changes of illumination caused by the dynamic parts of a scene and mark the changes as events. These events are used to reconstruct the pair of monocular images at first. Then, the 3D reconstruction and depth perception of the object can be obtained with the pair of monocular images.

In this paper, the visual stimuli are images as used in [3]. The moving stimuli can be designed with MATLAB. At each timestep, the stimulus locates in a novel position. The obtained image is projected into the network as the visual stimulus. Two groups afferent neurons generate activities depending on the projected visual stimuli. Then, the feedforward inputs induce the second-layer responses. The reconstructions of monocular stimuli could be generated with the induced responses of the network.

With the image projected into two groups of afferent neurons, the monocular received images are different. If a point \( i = (i_x, i_y) \) in the image can be represented with the two monocular coordinates as \((i_x^l, i_y^l)\) and \((i_x^r, i_y^r)\), a new plane coordinate can be designed with these monocular coordinates as base vectors. It is assumed that the stimuli are projected to the two groups of afferent neurons without any slope. Then, the point has a common vertical value of \( i_y^l = i_y^r = i_y \) in two monocular coordinates. The disparity between two presentations of a point, \( d \), only depends on \( i_x^l \) and \( i_x^r \). In the simulations, the influences of \( i_x^l \) and \( i_x^r \) are reflected in the center of each afferent received field in (9). The distance between original points of two monocular coordinates is set as \( b \) pixels. The focal length is set as \( f_{\text{focal}} \) pixels.

The neural reconstructions of images depend on the both the afferent neural responses and downstream excitatory neural responses in repeated testing simulations. For a given point \((i_x, i_y)\) in the image, the \( n_x \) th left-eye afferent neuron can produce the reconstruction as:

\[
I^{\text{rec}, \text{retinal}}_{n_x}(i_x, i_y, t) = a^{\text{rec}}_{n_x} \cdot f\left(i_x - i^l_{n_x, i_x}, i_y - i^l_{n_x, i_y}\right)
\]

\[
a^{\text{rec}}_{n_x} = \sum_{(i_x, i_y)} I(i_x, i_y, t) \cdot f\left(i_x - i^l_{n_x, i_x}, i_y - i^l_{n_x, i_y}\right)
\]  (29)

where \((i_x, i_y)\) and \((i_x^l, i_y^l)\) are the points in the image, \((i^l_{n_x, i_x}, i^l_{n_x, i_y})\) is the center of the received field of the \( n_x \) th afferent neuron. As introduced in [48], the coding of received stimuli is given by the joint responses of neurons.

To generate the temporary left-eye reconstruction of the given point \((i_x, i_y)\), the instantaneous synaptic traces of both left-eye afferent neurons, left-eye monocular neurons and binocular neurons are considered. The left-eye temporary reconstruction of the given point \((i_x, i_y)\) in the image is given as:

\[
I^l_{\text{rec}}(i_x, i_y, t) = \sum_{k} z_k \cdot z_{k_n} \cdot I^{\text{rec}, \text{retinal}}_{n_x}(i_x, i_y, t)
\]  (30)

Similarly, the right-eye temporary reconstruction of the given point \((i_x, i_y)\) is given as:

\[
I^r_{\text{rec}}(i_x, i_y, t) = \sum_{k} z_k \cdot z_{k_n} \cdot I^{\text{rec}, \text{retinal}}_{n_x}(i_x, i_y, t)
\]  (31)

Compared to previous WTA networks for primary visual cortical responses [33], [34], the number of neurons in this network has no strict relationship with the size of the original image. With \( I^l_{\text{rec}} \) and \( I^r_{\text{rec}} \) as a pair of stereo images, the temporary disparity at each point can be calculated for each reconstructed image. The disparity error \( e_d \) describes the correspondence error between
disparities for two reconstructed images. As introduced in [47], the temporal depth error $\varepsilon_0(i_{1}, i_{2}, t)$ of a given triangulation point is estimated as:

$$\varepsilon_0(i_{1}, i_{2}, t) = \frac{b \cdot f_{focal}}{d(i_{1}, i_{2}, t)} - \frac{b \cdot f_{focal}}{d(i_{1}, i_{2}, t) + \varepsilon_s(i_{1}, i_{2}, t)}$$  \hspace{1cm} (32)

For a point $(i_{1}, i_{2})$, the relationship between $\varepsilon_s(i_{1}, i_{2}, t)$, $\varepsilon_x(i_{1}, i_{2}, t)$ and $z(i_{1}, i_{2}, t)$ is approximated as [47]:

$$z(i_{1}, i_{2}, t) = \frac{z^2(i_{1}, i_{2}, t)}{b \cdot f_{focal}} \cdot \varepsilon_x(i_{1}, i_{2}, t).$$  \hspace{1cm} (33)

The depth perception of each triangulation point $(i_{1}, i_{2})$ is simulated as:

$$z(i_{1}, i_{2}, t) = \sqrt{b \cdot f_{focal} \cdot \varepsilon_x(i_{1}, i_{2}, t) / \varepsilon_s(i_{1}, i_{2}, t)}$$  \hspace{1cm} (34)

In our simulations, the stimulus moves in the blank environment. In the grayscale image at each timestep, only the pixels corresponding to the stimulus are non-zero. Following the assumption in [48], the temporal depth error $\varepsilon_0(i_{1}, i_{2}, t)$ is approximated as the difference of temporary left-eye and right-eye reconstructions of the triangulation point $(i_{1}, i_{2})$. The temporal disparity error $\varepsilon_s(i_{1}, i_{2}, t)$ is the correspondence error between disparities for monocular reconstructions of point $(i_{1}, i_{2})$. The temporal depth perception of the moving stimulus is simulated as the average of the depths of all the triangulation points corresponding to the temporal location of the stimulus.

III. RESULTS

A. BINOCULAR RIVALRY AND RESPONDING MODULATION INDUCED BY MOVING STIMULI

In this section, the spiking network receives the opposite moving traces of a same moving stimulus. The network is trained and tested to simulate binocular rivalry to different moving stimuli.

The moving stimulus is that used in [3]. For the simulations in this section, the original background is a 20 pixel × 20 pixel blank square. The white-gray bar is designed as a 18 pixel × 4 pixel rectangle. The white part has the gray-scale value of 1, the gray part has the gray-scale value of 0.5, the outside frame has the gray-scale value of 0.5. The blank background has the gray-scale value of 0. Different moving stimuli are generated with opposite moving trace designed for the same white-gray bar. For the first moving trace, the bar moves from left to right, horizontally. For the second moving trace, the bar moves from right to left, horizontally. The common fixed moving speed is 1 pixel/timestep. The moving stimuli are generated through MATLAB. For each moving trace, the location of the white-gray bar in the 20 pixel × 20 pixel image is changed with varying time. At each timestep, the blank background and the white-gray bar at the novel location would constitute the temporal image projected into the network. In this way, the moving stimuli are generated and used in the simulations (Fig. 2(A)). All the simulations are performed with a timestep of 1 millisecond. Each moving stimulus is presented to the network for 20 timesteps.

Received fields of all the afferent neurons are modeled as Difference-of-Gaussian filters in (9). The received fields of each afferent neural population are distributed on the grid with the spacing distance of 2 pixels and cover the image cooperatively. The Poisson spikes of each afferent modeling neuron are inspired by the visual stimulus in its received field. It should be noticed that, there are no strict limits for the sizes of images and the number of afferent neurons. Compared to previous WTA networks, the uses of afferent received fields could reduce the size of the networks [33], [34].

Both learning simulations and testing simulations are designed for the network. The learning phase consists of
200 simulations. All the modifications of plastic feedforward and lateral connective weights depend on temporal actions and temporal rewards, with initial values of weights sampled from a uniform distribution of [0.001, 1], independently. In each learning simulation, the distribution of the second-layer excitatory neural active state depends on the temporal membrane potential ((3), (5)).

At the beginning of the learning phase, all the clustering sets are blank. They will be inherited and updated over all the learning simulations. In each learning simulation, the four designed moving traces are rearranged randomly as the outside stimuli. Based on the second-layer excitatory spikes at each timestep, the network generates a temporal action to identify the outside stimuli and receives a temporal reward. With the reward, all the plastic connective weights and clustering sets will be updated according to the learning rules.

The testing phase consists of 100 simulations. The connective weights and clustering sets are kept during the testing phase. With a given order in Fig. 2(A), the designed moving traces are presented to the network in each testing simulation. The modulation index (MI) is used to qualify the excitatory responding modulation induced by different moving traces. The MI for each neuron is defined as the difference divided by the sum of the responses to paired stimuli [49]. For instance, the MI of the \( k \) th second-layer binocular excitatory neuron to Sti. 1 and Sti. 2 is calculated as:

\[
MI_k = \frac{r_{k1} - r_{k2}}{r_{k1} + r_{k2}},
\]

where \( r_{k1} \) is the total spiking counts of the \( k \) th binocular excitatory neuron to Sti. 1 in testing simulations, \( r_{k2} \) is the total spiking counts to Sti. 2. If \( r_{k1} = r_{k2} \), the \( k \) th binocular neuron has the same responses to two stimuli and \( MI_k = 0 \). If \( r_{k1} \neq r_{k2} \), the \( k \) th binocular neuron has the different responses to two stimuli and \( MI_k \neq 0 \). The zero MI indicates the varying stimuli do not affect the neural responses, while the non-zero MI stands for the modulation on neural activities. The denominator in (35) limits MI within \([-1,1]\). The MI with the absolute value closer to 1 means the larger neural responding modulation. The modulations of the neural population are measured by the histogram of MIs. As shown in Fig. 2(B), for three second-layer excitatory neural populations, opposite moving traces of the same white-gray bar can induce the non-zero CMIs of most neurons. It means that with the given white-gray bar, the network can simulate the responding modulation induced by opposite moving traces.

To different moving traces, the binocular rivalry is measured through the comparisons of monocular spikes. As introduced in section II-A, each pair of second-layer monocular neurons can generate spikes and activate the inhibitory neurons to suppress each other. For a pair of the left-eye and right-eye monocular modeling neurons, left-eye and right-eye spiking counts are accumulated over time and simulations with a stimulus given. Then, the difference between left-eye monocular spiking count and the right-eye monocular spiking count is given to each moving trace. The difference of spiking count represents how the moving traces induce the distinguishing monocular spikes. As shown in Fig. 2(C), the color and length of columns could represent the difference of monocular spiking count. For a pair of monocular neurons, the red column means that the left-eye spiking count is larger than the right-eye spiking count. While the blue column stands for a larger right-eye spiking count. The short column means that left-eye and
right-eye spiking counts are almost equal to each other. The difference between monocular neural responses over timestep and simulations could reflect the binocular rivalry induced by each moving stimulus.

Besides the competition between each pair of monocular neurons, the temporal responses of monocular neural populations to each stimulus also could reflect the binocular rivalry. At each timestep, the spiking counts over a monocular neural population and simulations are accumulated. Then, the differences between temporal spiking counts of two monocular neural population are averaged over neural population and shown in Fig. 2(D). Similarly, the color and length represent the difference of responses. Our simulations show that, the moving traces can induce distinguishing temporal responses of monocular neural populations.

To different moving traces of the same bar, the network could simulate different binocular rivalry and reflect the modulations of neural responses.
FIGURE 2. The responses to opposite moving patterns. (A) Moving stimuli in simulations. The white-gray bar moves alone opposite directions. Each moving stimulus is presented of 20 timesteps. (B) MI of three second-layer excitatory neural populations induced by different moving stimuli. The non-zero MIs stand for the different neural responses to opposite moving patterns. The network could reflect neural responding modulations induced by different stimuli. (C) Difference of monocular neural spiking counts over timesteps and simulations. For each pair of monocular neurons, the neural competition is reflected by the color and length of the column. For a pair of monocular neurons, the red column means the larger left-eye neural spiking strength while the blue column stands for a larger right-eye neural spiking strength. The network could simulate different binocular rivalry to opposite moving patterns. (D) Difference of temporal spiking rates of neural populations over simulations. The temporal neural competition is reflected by the color and length of the column.
B. INTEROCULAR BLUR DIFFERENCES AFFECT DEPTH PERCEPTION OF MOVING STIMULI

In this section, the moving white-gray bar is presented to our network with binocular visual blur. For the moving stimuli with interocular blur differences, the influences of the varying blurring strength on depth perceptions are explored.

The moving stimuli are similar with those in section III-A. For the simulations in this section, the original image is a 20 pixel $\times$ 20 pixel square image. The white-gray bar is a 18 pixel $\times$ 4 pixel rectangle, shown in Fig. 3(A). When the bar moves in the environment, the non-stimulus part is blank with the gray-scale value of 0. At each timestep, the bar and the blank background constitute a new image.

The stimulus moves from left to right with moving speed of 1 pixel/timestep in all the simulations in this section. For this same moving stimulus, different blurry visual stimuli are designed for our spiking network. The original white-gray bar without any process is set as the norm type. The normal white-gray bar is filtered with the low-pass and high-pass Gaussian filters to make it sharp or blurry, respectively (Fig. 3(A)) [3]. The monocular moving stimuli could be normal, sharp or blurry. Different monocular moving stimuli could be presented to two groups of afferent neurons, respectively. In this way, the interocular blur differences are generated for our simulations.

The binocular stimuli are designed through different combinations of monocular stimuli. There are three kinds of binocular stimuli designed in our simulations. In the first kind, the normal bar without filtered is presented to two groups of afferent neurons. This kind of binocular stimuli means that two eyes can receive the normal inputs. It is designed for comparison with other binocular stimuli. In the second kind of binocular stimuli, the sharp bar is presented to the left-eye afferent neurons and the blurry bar is presented to the right-eye afferent neurons. Oppositely, the third kind of stimuli includes the left-eye blurry bar and the right-eye sharp bar. These different binocular stimuli with interocular blur differences are designed as introduced in [3] and shown in Fig. 3(B). For convenience, three kinds of binocular stimuli are marked as normal/normal, sharp/blurry and blurry/sharp, respectively. Both the low-pass and the high-pass Gaussian filters have the variance of 1. The moving stimuli and the environment are designed artificially in MATLAB, the distances and depths have no strict real-world units. The goal of this section is to compare depth perceptions to different kinds of interocular blur differences.

The training of the network includes 200 simulations as introduced in section III-A. The initial values of plastic weights are sampled from a uniform distribution of [0.001, 1]. In each learning simulation, the binocular stimuli will be rearranged randomly into a novel sequence and presented to the network.

For the trained network, the testing phase includes 100 simulations. To different interocular blur differences, the binocular rivalry could be measured through both the responses of paired monocular excitatory neurons and the temporal responses of monocular neural population. The measures are similar with those in section III-A. As the differences of monocular responses shown in Fig. 3(D), the color and length stand for the difference of spiking count. To the normal/normal stimulus, many pairs of monocular neurons have the responding difference close to 0. It means that the normal/normal stimulus can induce both left-eye
and right-eye monocular spikes which are almost equal to each other. Compared with monocular responses to the normal/normal stimulus, the sharp/blurry and blurry/sharp stimuli can induce obvious binocular rivalry from the network. Particularly, the monocular responding differences in Fig. 3(D) reflect the different binocular rivalry induced by the sharp/blurry and blurry/sharp stimuli, respectively. To the given pair of monocular neurons in the network, the different interocular blur differences can induce the varying competition. The temporal responses of monocular neural populations to each type of interocular blur difference are also measured to reflect binocular rivalry. To the given interocular blur difference, the differences between temporal spiking counts of two monocular neural population are averaged over neural population and shown in Fig. 3(E). Our simulations show that, the sharp/blurry and blurry/sharp stimuli can induce distinguishing temporal responses of neural populations.

To explore how the interocular blur differences affect depth perceptions, the temporal depth perceptions for each binocular stimulus is averaged over time and simulations according to (34). As shown in Fig. 3(B), compared to the normal/normal stimulus, the sharp/blurry stimulus can induce the larger perceptions of the depths while the blurry/sharp stimulus can lead to the smaller perceptions. The opposite kinds of interocular blur differences could induce opposite depth perceptions of the same moving trace from the network. The relationship between these perceptions are consistent with that introduced in [3]. It means that, with interocular blur differences, the network could simulate the illusory perception of the depths of the moving traces.

Besides, to explore how the blurry strength affects depth perception, a novel set of binocular stimuli is designed for our network. In the first stimulus, the normal moving bar without filtered performs as the input to two groups of afferent neurons. In the other four stimuli, all the left-eye bars are sharp and all the right-eye bars are blurry. Different from the stimuli in Fig. 3(B), to generate these four binocular stimuli, the original moving bar is processed by the Gaussian filters with the variance of 1, 1.1, 1.2 and 1.3, respectively. The bars move from left to right with the moving speed of 1 pixel/timestep.

The network is trained as introduced in section III-A. The initial values of weights sampled from a uniform distribution of [0.001, 1]. The simulated depth perception for each binocular stimulus over simulations are shown in Fig. 3(C). The variance of 0 stands for the moving bar without filtered and the depths are estimated for comparison. As shown in Fig. 3(C), the larger variance stands for the larger interocular blur differences between monocular images. To the stimuli with larger interocular blur differences, the network could simulate more biased depth perceptions.

The simulations in this section show that, to the stimuli with the given moving speed and moving direction, the network could simulate binocular rivalry to the interocular blur difference and generate biased depth perception of the moving trace. The stimuli with interocular blur difference could impact binocular rivalry in the network and lead to different illusory depth perceptions.
**FIGURE 3.** The simulated depth perceptions of the moving depths with interocular blur differences. (A) The bars with blur differences in our simulations. The normal white-gray bar becomes sharp or blurry after filtered by the different Gaussian filters. (B) Simulated depth perceptions for binocular stimuli with interocular blur differences. The bar moves from left to right. The simulated depth perceptions of each kind of binocular stimulus are averaged over time step and testing simulations. The interocular blur differences could lead to relationships between depth perceptions consistent with those in [3]. (C) Simulated depth perceptions for binocular stimulus with different variances of Gaussian filters. For the sharp/blurry stimulus, depth perceptions become larger with variances of Gaussian filters increasing. The increasing blurry strength could intensify the illusory depth perceptions. (D) Difference of spiking counts of paired monocular excitatory neurons. Similar with Fig. 2, the neural competition is reflected by the color and length of the column. (E) Difference of monocular populational spiking rates. All the monocular responses are measured over testing simulations. Different interocular blur differences can induce distinguishing monocular responses.

### C. MOVING DIRECTIONS AFFECT DEPTH PERCEPTION OF MOVING STIMULI

In this section, it is explored how the moving direction affects depth perception from our network. The moving bar has two opposite directions, from left to right or from right to left. The bar moves with the common moving speed of 1 pixel/timestep, which is shown in Fig. 4(A). At each time, the bar and the blank background constitute a new 20 pixel x 20 pixel square image into the network.

The binocular stimuli are the normal/normal stimulus and the blurry/sharp stimulus as introduced in section III-B. The variance of the Gaussian filters is 2. It means that there are four stimuli in this section, including the normal/normal stimulus from left to right, the normal/normal stimulus from right to left, the blurry/sharp stimulus from left to right and the blurry/sharp stimulus from right to left. For each direction, the normal/normal stimulus is designed for
comparison.

The training of the network includes 200 simulations as introduced in section III-A. For the trained network, the testing phase includes 100 simulations. To measure binocular rivalry, the monocular neural responses are measured over simulations. Similar with Fig. 2, the color and length represent the difference of spiking count. To each type of the moving stimulus, the responding difference of paired monocular neurons are shown in Fig. 4(C). The differences between temporal spiking counts of two monocular neural population are averaged over neural population in Fig. 4(D). Our simulations show that the moving direction can induce distinguishing neural responses and binocular rivalry from the network.

The simulated depth perception of each stimulus is averaged over time and testing simulations, as shown in Fig. 4(B). To the opposite moving directions, the network could generate responses and simulate opposite depth perceptions consistent with those in [3].

Our simulations shown that, for the given interocular blur difference and the moving speed, the moving direction could induce distinguishing monocular responses of our network. These induced network responses can simulate opposite depth perceptions consistent with the experimental results. The stimuli with interocular blur difference along different directions could modify binocular rivalry from the network and lead to opposite illusory depth perceptions.
FIGURE 4. The simulated depth perceptions of the moving depths with different directions. (A) Moving directions of the bar. With the common speed, the bar moves from left to right or from right to left. (B) Simulated depth perceptions for binocular stimuli. The binocular stimuli in this section including the normal/normal stimulus and the blurry/sharp stimulus. Depth perceptions are simulated as introduced in section III-B. Opposite directions could induce opposite depth perceptions from the network. For each direction, the relationship between depth perceptions is consistent with that in [3]. (C) Difference of monocular spiking counts. Similar with Fig. 2, the neural competition is reflected by the color and length of the column. (D) Difference of neural populational spiking rates. For the given blurry/sharp stimulus, the moving direction induces distinguishing monocular responses.

C. MOVING SPEEDS AFFECT DEPTH PERCEPTION OF MOVING STIMULI

In this section, it is explored how the moving speed affects depth perceptions from our network. For the stimulus with the interocular blur difference, the left-eye bar is designed to be blurry while the right-eye bar is sharp, which is shown in Fig. 5(A). In this way, the binocular stimuli are designed as the normal/normal stimulus and the blurry/sharp stimulus in this section. The variance of Gaussian filters is 0.5 for the blurry/sharp stimulus. The bar moves from left to right with the speeds of 0.8 pixel/timestep, 0.9 pixel/timestep and 1 pixel/timestep, respectively.

The training phase includes 200 simulations as introduced in section III-A. For the trained network, the testing phase includes 100 simulations. Over testing simulations, binocular rivalry is reflected by the comparisons of monocular neural responses. As shown in Fig. 5(C, D), the varying moving speed could induce distinguishing responses of the pairs of monocular neurons and temporal responses of monocular neural populations. With the modified responses, the network could simulate depth perceptions of each moving stimulus. Depth perceptions are averaged over timestep and testing simulations, as shown in Fig. 5(B). For the given interocular blur difference and the moving direction, the faster moving speed could induce the more biased depth perceptions compared to the normal stimuli with the same speed. The varying speed of moving stimuli with the interocular blur difference could induce distinguishing binocular rivalry from the network and lead to illusory depth perceptions. Our simulations could reflect the influence of the varying moving speed on neural responses, inducing different illusory depth perceptions.
FIGURE 5. The simulated depth perceptions of the moving depths with different speeds. (A) The moving direction and the stimulus with the interocular blur difference. (B) Simulated depth perceptions for binocular stimuli with different moving speeds. The binocular stimuli in this section including the normal/normal stimulus and the blurry/sharp stimulus. The depths are estimated as introduced in section III-B. The faster speed can intensify the misperceptions of depths. (C) Difference of monocular spiking counts. (D) Difference of population spiking rates. For the given blurry/sharp stimulus, the moving direction induces distinguishing monocular responses.

IV. CONCLUSION

This paper has provided a probabilistic framework based on binocular rivalry to explore depth perceptions of moving stimuli with the interocular blur differences. The stimuli with interocular blur differences could be caused by the prescription lens correction in daily life [3]. Interocular blur differences could cause misperceptions of motion, impacting public safety. While the binocular vision has been studied in the visual perception for a long time, how the interocular blur differences induce the visual motion misperception remains an open issue. To explore the
underlying mechanism, this paper constructs a stochastic spiking network with plastic connective weights, considering binocular rivalry. This network could reflect impacts of stimuli with interocular blur differences on neural responses and binocular rivalry, simulating illusory depth perceptions consistent with experimental observations.

Binocular rivalry and unsupervised online identification are considered into this k-WTA network. The network consists of two layers. Two populations of first-layer afferent neurons have their received fields as the Difference-of-Gaussians filters. Poisson spikes of each afferent neuron are induced by visual information projected into its received field. Compared to previous WTA networks, the uses of afferent received fields could reduce the size of the networks [25], [26]. The excitatory monocular neurons and binocular neurons in the second layer simulate the binocular rivalry and fusion in the primary visual cortex. Different groups of excitatory monocular neurons could compete with each other through the special inhibitory neural responses. Compared to previous WTA networks, our network could explore binocular visual perceptions in the primary visual cortex [25], [26]. The network generates its identifications of the outside stimuli with the multi-dimensional excitatory spikes in the second layer. Without limiting the size of the network, the unsupervised identification could provide online rewards to network and control the modifications of connective weights [37]. Deduced from the probabilistic model, the learning rules for the plastic connections depend on both the temporal rewards and the neural responses.

This paper designs binocular moving stimuli with various interocular blur differences, moving directions and moving speeds, exploring how these factors affect depth perception of the moving trace. For comparison, the normal moving stimuli are designed without any interocular blur differences. To reflect binocular rivalry, monocular responses of the trained network are measured over simulations in two aspects. For the given pair of monocular neurons, the difference between their spiking counts to various stimuli is calculated over timesteps and simulations. For the different groups of monocular neurons, the temporary responses of neural populations are compared at each timestep. Both two kinds of measures reflect that, with interocular blur differences, the network could simulate distinguishing binocular rivalry to various stimuli. It means that the considered factors could modify responses of the network. Then, to the moving stimuli, the network simulates the perception of its depth with the induced responses. To the stimuli with opposite types of interocular blur differences, the network could simulate opposite depth perceptions of the same moving trace. With the increasing blurry strength, the illusory depth perception from the network becomes more biased. The stimuli with opposite moving directions, could induce opposite depth perceptions from the network. The moving stimuli with faster speed could induce depth perceptions with larger bias. For each designed factor in this paper, the network could simulate modified depth perceptions consistent with experimental results [3]. The simulations show that, to stimuli with interocular blur differences, varying factors could modify binocular rivalry in the network and induce distinguishing illusory depth perceptions. To interocular blur differences caused by the prescription lens correction, the network provides the possible theoretical framework to explore the illusory depth perception. In the future, the k-WTA
network might be a possible method based on probabilistic inference for other explorations of binocular visual perception.

**REFERENCES AND FOOTNOTES**

[1] K. A. Turano and S. M. Heidenreich, “Eye movements affect the perceived speed of visual motion,” Vision Res., vol. 39, no. 6, pp. 1177–1187, 1999.

[2] A. Goettker, D. I. Braun, A. C. Schütz, and K. R. Gegenfurtner, “Execution of saccadic eye movements affects speed perception,” Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 9, pp. 2240–2245, 2018.

[3] J. Burge, V. Rodriguez-Lopez, and C. Dormorosoro, “Monovision and the misperception of motion,” Curr. Biol., vol. 29, no. 15, pp. 2586-2592.e4, 2019.

[4] R. F. Hess, C. V. Hutchinson, T. Ledgeway, and B. Mansouri, “Binocular influences on global motion processing in the human visual system,” Vision Res., vol. 47, no. 12, pp. 1682–1692, 2007.

[5] C. de S’perati and I. M. Thornton, “Motion prediction at low contrast,” Vision Res., vol. 154, pp. 85–96, 2019.

[6] J. M. Fernandez and B. Farell, “A neural model for the integration of stereopsis and motion parallax in structure from motion,” Neurocomputing, vol. 71, no. 7-9, pp. 1629–1641, 2008.

[7] M. Sizintsev and R. P. Wildes, “Spacetime stereo and 3D flow via binocular spatiotemporal orientation analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 11, pp. 2241–2254, 2014.

[8] C. Shi, S. Pundlik, and G. Luo, “Without low spatial frequencies, high resolution vision would be detrimental to motion perception,” J. Vis., vol. 20, no. 8, pp. 29, 1–13, 2020.

[9] D. Alais, “Binocular rivalry: competition and inhibition in visual perception,” Wiley Interdiscip. Rev. Cogn. Sci., vol. 3, no. 1, pp. 87–103, 2012.

[10] H. Ralio, J. Saastamoinen, S. Kylmälu, and A. Peltola, “Binocular disparity can augment the capacity of vision without affecting subjective experience of depth,” Sci. Rep., vol. 8, no. 1, pp. 15798, 2018.

[11] S. Dunn and M. Jones, “Binocular rivalry dynamics associated with high levels of self-reported autistic traits suggest an imbalance of cortical excitation and inhibition,” Behav. Brain Res., vol. 388, pp. 112603, 2020.

[12] K. Dougherty, M. A. Cox, J. A. Westerberg, and A. Maier, “Binocular modulation of monocular V1 neurons,” Curr. Biol., vol. 29, no. 3, pp. 381-391.e4, 2019.

[13] R. N. Raveendran, W. R. Bobier, and B. Thompson, “Binocular vision and fixational eye movements,” J. Vis., vol. 19, no. 4, pp. 9, 1–15, 2019.

[14] W. M. King, “Binocular coordination of eye movements - Hering’s Law of equal innervation or unicocular control?,” Eur. J. Neurosci., vol. 33, no. 11, pp. 2139–2146, 2011.

[15] L. Feng and L. Xiaqiang, “The dynamic model of lateral inhibition network and it is application,” Neural Comput. Appl., vol. 22, no. SUPPL.1, pp. 125–131, 2013.

[16] A. P. Morris and B. Krekelberg, “A stable visual world in primate primary visual cortex,” Curr. Biol., vol. 29, no. 9, pp. 1471-1480.e6, 2019.

[17] H. R. Wilson, “Binocular contrast, stereopsis, and rivalry: toward a dynamical synthesis,” Vision Res., vol. 140, pp. 89–95, 2017.

[18] G. Sun, Y. Ding, R. Deng, Y. Zhao, X. Chen, and A. S. Krylov, “Steroscopic image quality assessment by considering binocular visual mechanisms,” IEEE Access, vol. 6, pp. 51337–51347, 2018.

[19] G. Xu, J. Chen, and X. Li, “3-D reconstruction of binocular vision using distance objective generated from two pairs of skew projection lines,” IEEE Access, vol. 5, pp. 27272–27280, 2017.

[20] F. Shao, W. Chen, G. Jiang, and Y. S. Ho, “Modeling the perceptual quality of stereoscopic images in the primary visual cortex,” IEEE Access, vol. 5, pp. 15706–15716, 2017.

[21] R. Blake, Y. Yang, and H. R. Wilson, “On the coexistence of stereopsis and binocular rivalry,” Vision Res., vol. 31, no. 7-8, pp. 1191–1203, 1991.

[22] A. Buckthought and H. R. Wilson, “Interaction between binocular rivalry and depth in plaid patterns,” Vision Res., vol. 47, no. 19, pp. 2543–2550, 2007.

[23] T. J. Andrews and D. Holmes, “Steroscopic depth perception during binocular rivalry,” Front. Hum. Neurosci., vol. 5, pp. 99, 2011.

[24] B. Yu and L. Zhang, “Pulse-coupled neural networks for contour and motion matchings,” IEEE Trans. Neural Networks, vol. 15, no. 5, pp. 1186–1201, 2004.

[25] R. Eckhorn, A. M. Gail, A. Bruns, A. Gabriel, B. Al-Shaikhli, and M. Saam, “Different types of signal coupling in the visual cortex related to neural mechanisms of associative processing and perception,” IEEE Trans. Neural Networks, vol. 15, no. 5, pp. 1039–1052, 2004.

[26] K. Shimonomura and T. Yagi, “A multichip aVLSI system emulating orientation selectivity of primary visual cortical cells,” IEEE Trans. Neural Networks, vol. 16, no. 4, pp. 972–979, 2005.

[27] A. Rangan, L. Tao, G. Kovacic, and D. Cai, “Multiscale modeling of the primary visual cortex,” IEEE Eng. Med. Biol. Mag., vol. 28, no. 3, pp. 19–24, 2009.
[28] Z. Chen and R. Cai, “Contour detection by simulating the curvature cell in the visual cortex and its application to object classification,” IEEE Access, vol. 8, pp. 74472–74484, 2020.

[29] A. W. Freeman, “Multistage model for binocular rivalry,” J. Neurophysiol., vol. 94, no. 6, pp. 4412–4420, 2005.

[30] R. F. Murray, “A model of lightness perception guided by probabilistic assumptions about lighting and reflectance,” J. Vis., vol. 20, no. 7, pp. 28, 1–22, 2020.

[31] G. Orbán, P. Berkes, J. Fiser, and M. Lengyel, “Neural variability and sampling-based probabilistic representations in the visual cortex,” Neuron, vol. 92, no. 2, pp. 530–543, 2016.

[32] R. S. Zemel, “Probabilistic interpretation of population codes,” Neural Comput., vol. 10, no. 2, pp. 403–430, 1998.

[33] Z. Jonke, R. Legenstein, S. Habenschuss, and W. Maass, “Feedback inhibition shapes emergent computational properties of cortical microcircuit motifs,” J. Neurosci., vol. 37, no. 35, pp. 8511–8523, 2017.

[34] D. Kappel, B. Nessler, and W. Maass, “STDP installs in winner-take-all circuits an online approximation to Hidden Markov Model learning,” PLoS Comput. Biol., vol. 10, no. 3, pp. e1003511, 2014.

[35] S. Klampfl and W. Maass, “Emergence of dynamic memory traces in cortical microcircuit models through STDP,” J. Neurosci., vol. 33, no. 28, pp. 11515–11529, 2013.

[36] M. C. van Rossum, G. Q. Bi, and G. G. Turrigiano, “Stable Hebbian learning from spike timing-dependent plasticity,” J. Neurosci., vol. 20, no. 23, pp. 8812–8821, 2000.

[37] R. Van Rullen and S. J. Thorpe, “Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex,” Neural Comput., vol. 13, no. 6, pp. 1255–1283, 2001.

[38] M. Avermann, C. Tomm, C. Mateo, W. Gerstner, and C. C. H. Petersen, “Microcircuits of excitatory and inhibitory neurons in layer 2/3 of mouse barrel cortex,” J. Neurophysiol., vol. 107, no. 11, pp. 3116–3134, 2012.

[39] W. Liu and X. Liu, “The effects of eye movements on the visual cortical responding variability based on a spiking network,” Neurocomputing, vol. 436, pp. 58–73, 2021.

[40] I. Y. Segal, C. Giladi, M. Gedalin, M. Rucci, M. Ben-Tov, Y. Kushinsky, A. Mokeichev, R. Segev, “Decorrelation of retinal response to natural scenes by fixational eye movements,” Proc. Natl. Acad. Sci. U. S. A., vol. 112, no. 10, pp. 3110–3115, 2015.

[41] E. A. Benardete and E. Kaplan, “The receptive field of the primate P retina ganglion cell, I: Linear dynamics,” Vis. Neurosci., vol. 14, no. 1, pp. 169–185, 1997.

[42] E. Rueckert, D. Kappel, D. Tanneberg, D. Peceski, and J. Peters, “Recurrent spiking networks solve planning tasks,” Sci. Rep., vol. 6, pp. 21142, 2016.

[43] R. Legenstein, Z. Jonke, S. Habenschuss, and W. Maass, “A probabilistic model for learning in cortical microcircuit motifs with data-based divisive inhibition,” arXiv, 2017, https://arxiv.org/abs/1707.05182.

[44] J. Heierman, E. Haasdijk, and A. E. Eiben, “Unsupervised identification and recognition of situations for high-dimensional sensori-motor streams,” Neurocomputing, vol. 262, pp. 90–107, 2017.

[45] G. J. Székely and M. L. Rizzo, “Energy statistics: a class of statistics based on distances,” J. Stat. Plan. Inference, vol. 143, no. 8, pp. 1249–1272, 2013.

[46] J. Kogler, F. Eibensteiner, M. Humenberger, M. Gelautz, and J. Scharinger, “Ground truth evaluation for event-based silicon retina stereo data,” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., pp. 649–656, 2013.

[47] D. Gallup, J. Frahm, P. Mordohai, M. Pollefeys, and C. Hill, “Variable baseline / resolution stereo,” IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–8, 2008.

[48] M. Osswald, S. H. Ieng, R. Benosman, and G. Indiveri, “A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems,” Sci. Rep., vol. 7, pp. 40703, 2017.

[49] A. J. Keller, M. Dipoppa, M. M. Roth, M. S. Caudill, A. Ingrosso, M. D. Miller, M. Scanziani, “A disinhibitory circuit for contextual modulation in primary visual cortex,” Neuron, vol. 108, no. 6, pp. 1181-1193.e8, 2020.

Weisi Liu received the B.S. degree in Probability and Statistics from Qufu Normal University, P.R. China. Currently he is studying for Ph.D. degree in Nanjing University of Aeronautics and Astronautics, P.R. China. His research interests are in theoretical neuroscience, neural networks and applied probability and statistics.

Xinsheng Liu received the Ph.D. in Mathematical Statistics from Nanjing University, P.R. China. Currently he is a professor in Nanjing University of Aeronautics and Astronautics, P.R. China. His research interests are in theoretical neuroscience, neural networks, evolutionary dynamics and the applied statistics in biology.
