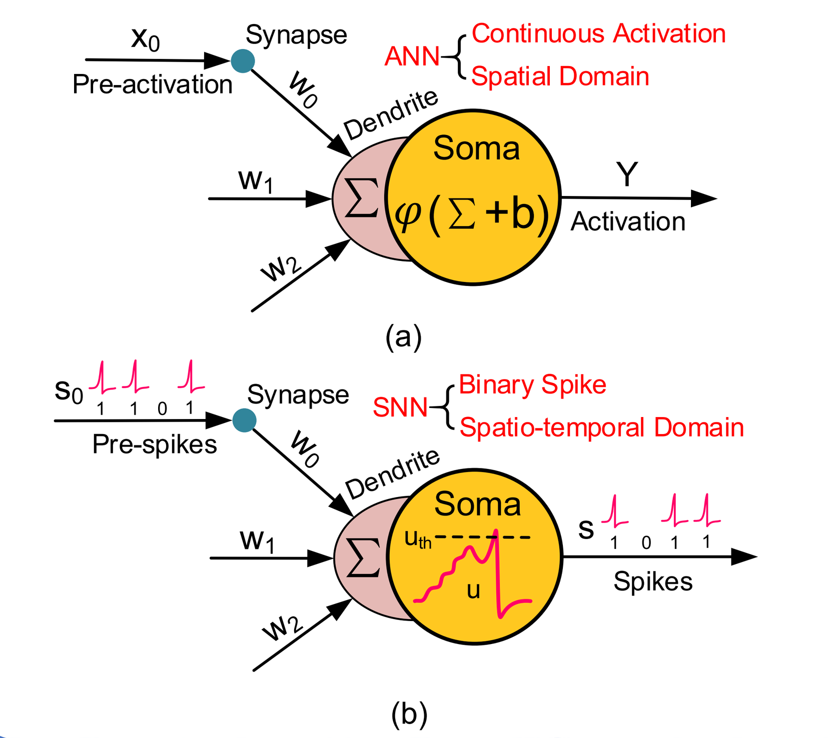
Unveiling the Spatial-temporal Effective Receptive Fields of Spiking Neural Networks

Jieyuan Zhang, Xiaolong Zhou, Shuai Wang, Wenjie Wei, Hanwen Liu, Qian Sun, Malu Zhang*, Yang Yang, Haizhou Li

电子科技大学 University of Electronic Science and Technology of China
深圳河套學院 Shenzhen Loop Area Institute
香港中文大學(深圳) The Chinese University of Hong Kong, Shenzhen

NEURAL INFORMATION PROCESSING SYSTEMS



Motivation

Compared to traditional artificial neural networks(ANNs), spiking neural networks(SNNs) has complex spatial-temporal interactions. In ANN field, the effective receptive field (ERF) serves as a valuable tool for analyzing feature extraction capabilities in visual long-sequence modeling. We hope to extend the framework to be capable for SNNs so that we could analyze the learning behaviors inside a SNN model.

Spatial-temporal Effective Receptive Field (ST-ERF)

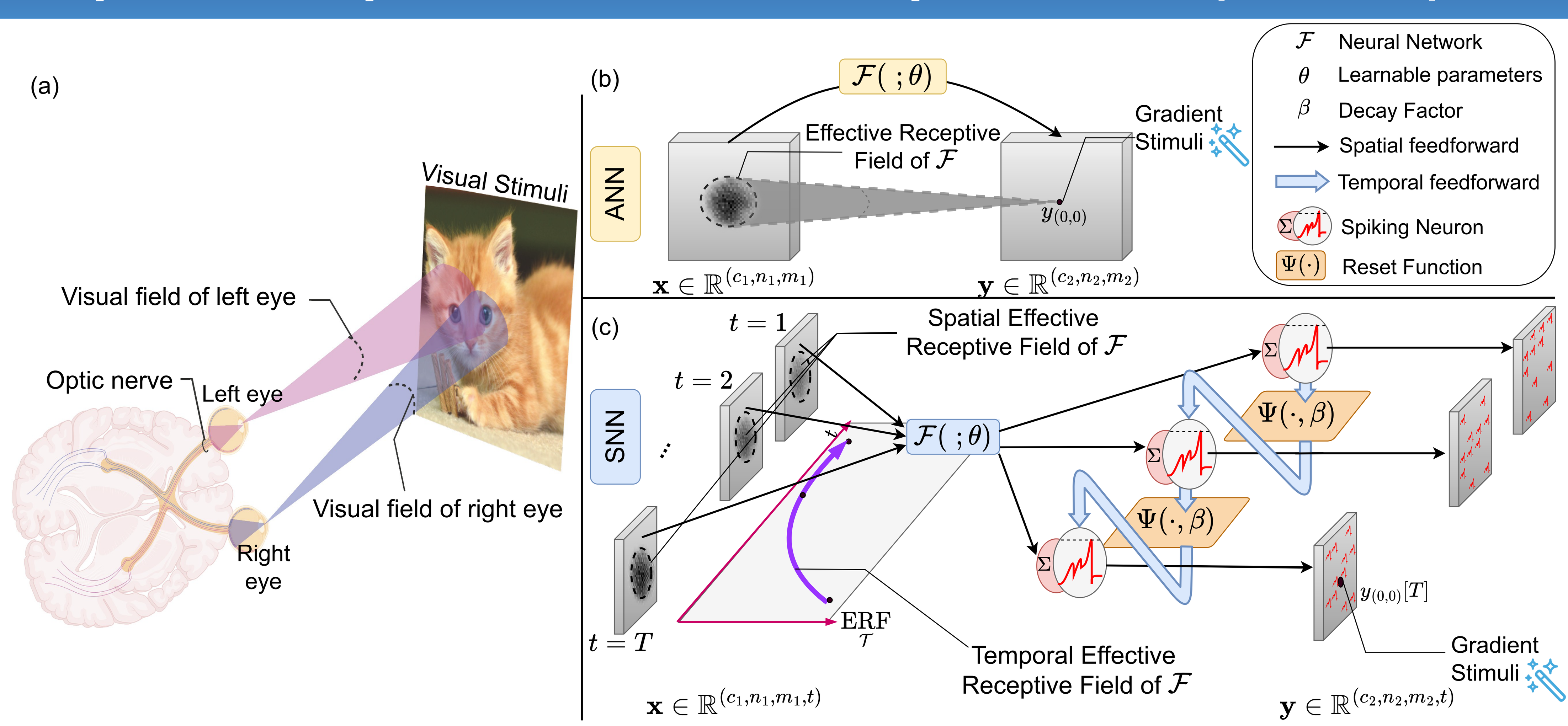


Figure 1: (a): Human visual field. (b): ERF in ANNs. (c): ST-ERF in SNNs

ANN ERF

$$\mathrm{ERF}_{(i,j)}[y_{(m,n)};\mathbf{x}] = \frac{\partial y_{(m,n)}}{\partial x_{(i,j)}}$$

SNN ST-ERF

$$\mathrm{ERF}^{(\mathcal{S},\mathcal{T})}_{(i,j)}[y_{(m,n)}[t],\tau;\mathbf{x}] = \frac{\partial y_{(m,n)}[t]}{\partial x_{(i,j)}[t-\tau]}, \quad 1 \le t \le T, 0 \le \tau \le t-1.$$

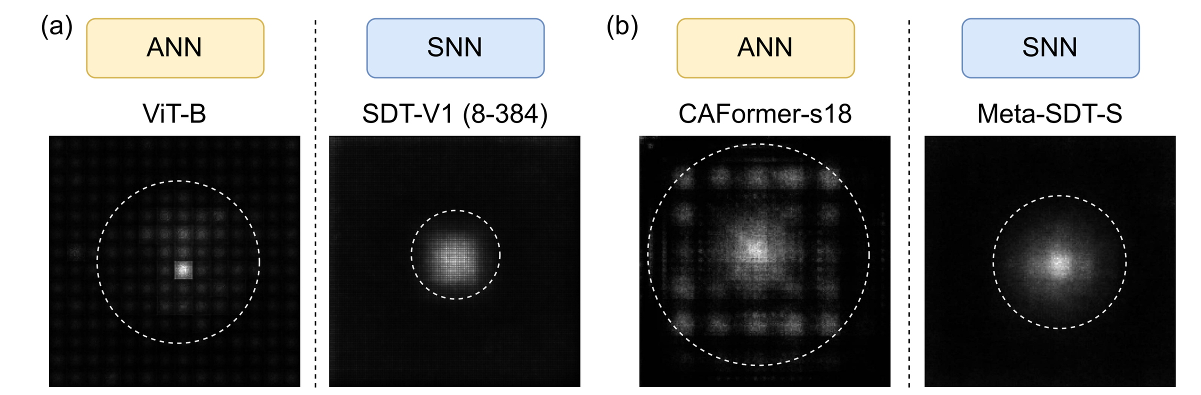
SNN Spatial-ERF

$$\mathrm{ERF}^{(\mathcal{S})}_{(i,j)}[y_{(m,n)};\mathbf{x}] = \frac{1}{T}\sum_{t=1}^{T}\sum_{\tau=0}^{t-1} w(t,\tau)\cdot \mathrm{ERF}^{(\mathcal{S},\mathcal{T})}_{(i,j)}[y_{(m,n)}[t];\mathbf{x},\tau]$$

SNN Temporal-ERF

$$\mathrm{ERF}^{(\mathcal{T})}[\tau;\mathbf{x}] = \sum_{i,j}\sum_{m,n} \mathrm{ERF}^{(\mathcal{S},\mathcal{T})}_{(i,j)}[y_{(m,n)}[T];\mathbf{x},\tau].$$

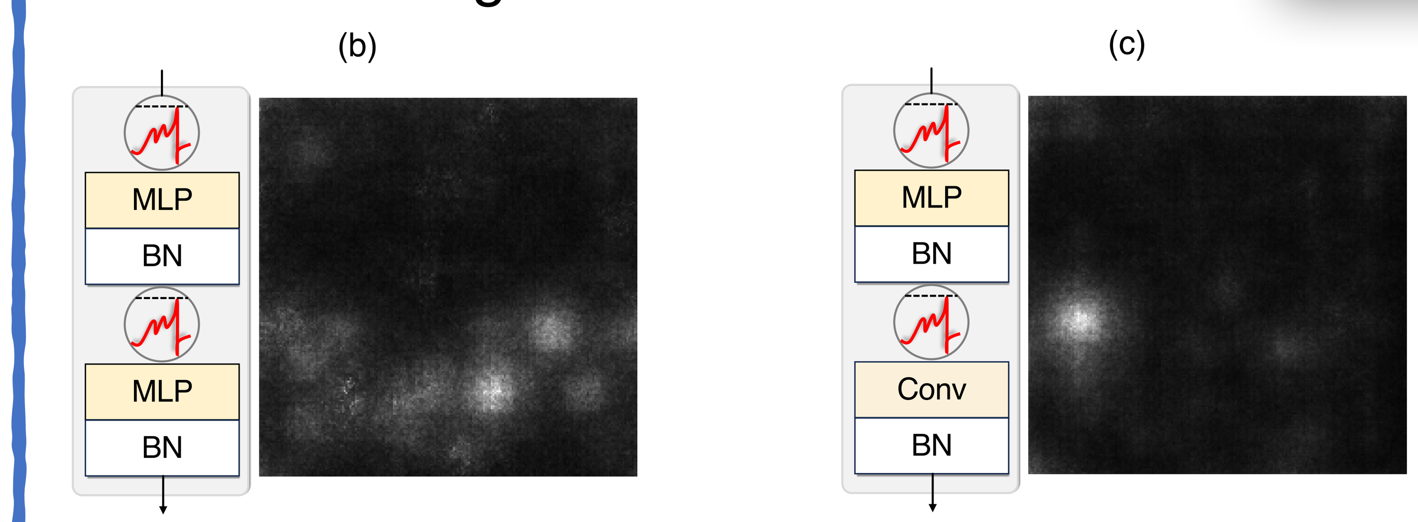
Problem Analysis



Mainstream Transformer-based SNNs incorporates multiple convolutional layers at the early stage of the network, facilitating low-level spatial features extraction from input images. This design enhances local feature extraction, yet it inherently constrains the model's capacity to aggregate information across distant spatial regions. Together, these findings suggest that the convolutional operations enhances local feature sensitivity but poses challenges for maintaining long-range spatial coherence in Transformer-based SNNs.

Method

Re-design of Channel Mixer Block

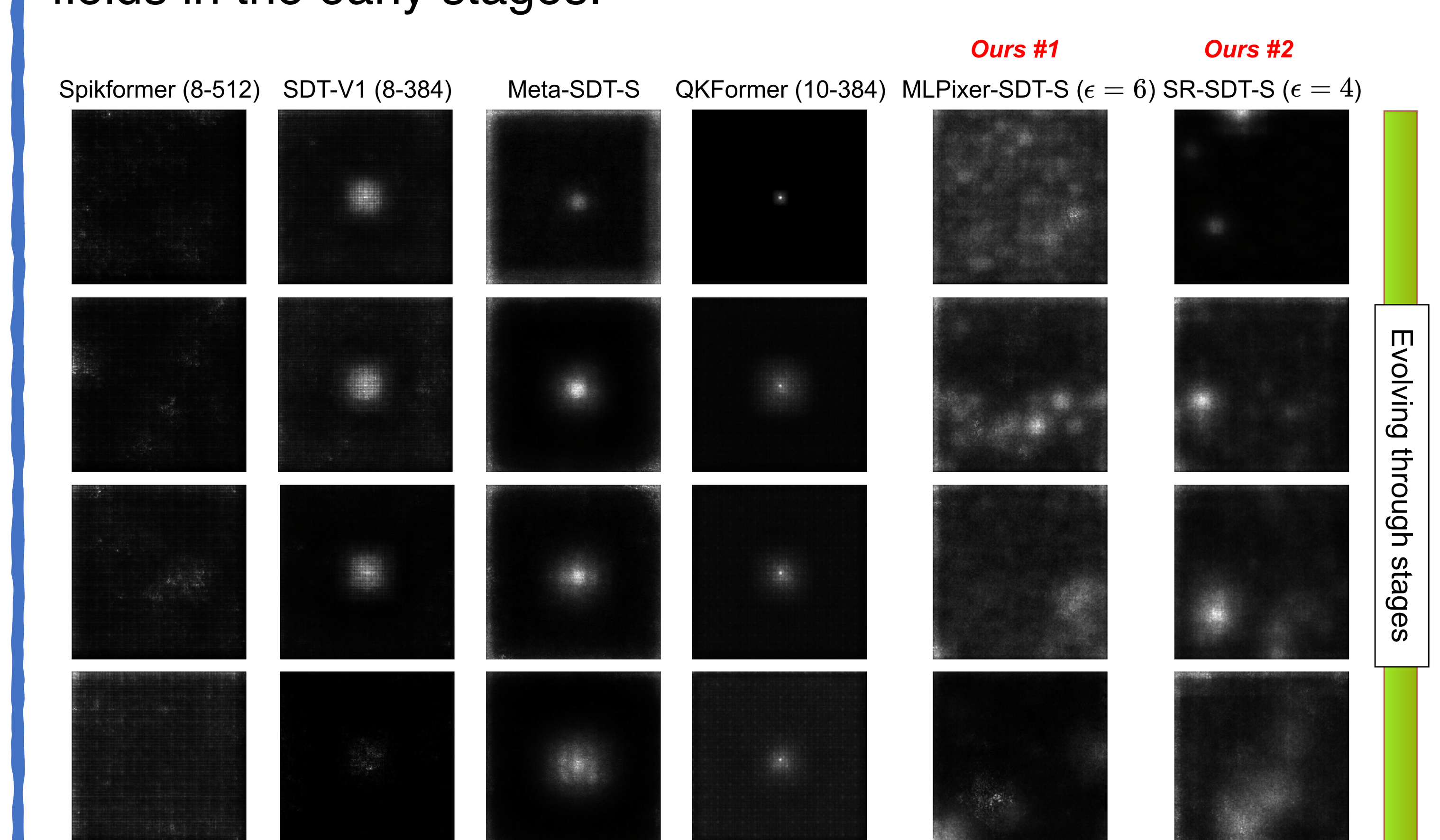


MLPixer mitigates the ERF's bias toward a Gaussian-like central concentration and enabling SNNs to capture long-range dependencies more effectively.

SRB module replaces only the second convolution in the channel mixer with a single-layer MLP operation. In this manner, SRB module reduces additional parameters while maintaining performance.

Experiments & Conclusion

Spikformer shows diffuse receptive fields across all stages. SDT-V1, Meta-SDT, and QKFormer exhibit more centered spatial distributions that gradually expand as depth increases. Our two Meta-SDT variants establish global spatial receptive fields in the early stages.



COCO 2017 Performance

Arch.	#T	#P	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
SDTv3-T[33]	4	25M	15.2	35.5	10.2	15.2	33.0	12.3
MLPixer(ϵ4)	4	24M	16.2	37.0	11.5	15.2	32.9	12.5
MLPixer(ϵ6)	4	25M	17.5	38.5	13.2	16.2	34.5	13.5
SRB(ϵ4)	4	25M	**18.2**	**39.2**	**13.8**	**17.5**	**34.8**	**14.3**
SDTv3-B[33]	4	39M	21.7	46.9	17.0	20.1	41.8	17.5
MLPixer(ϵ4)	4	36M	22.9	47.6	19.2	21.0	43.4	18.3
MLPixer(ϵ6)	4	39M	25.1	48.8	22.5	21.9	43.5	19.6
SRB(ϵ4)	4	37M	**25.8**	**48.9**	**22.8**	**22.5**	**43.9**	**20.4**

ADE20K Performance

Arch.	Ch. Mixer	#T	Param.(M)	mIoU(%)
SDTv3 -T[33]	C2d-k3(ϵ4)	4	6.5 BASE	34.9 BASE
	MLPix.(ϵ4)	4	5.9 (↓0.6)	34.9 (↑0.0)
	MLPix.(ϵ6)	4	6.6 (↑0.1)	35.9 (↑1.0)
	SRB(ϵ4)	4	6.2 (↓0.3)	**38.2** (↑3.3)
SDTv3 -B[33]	C2d-k3(ϵ4)	4	20.4 BASE	41.1 BASE
	MLPix.(ϵ4)	4	18.0 (↓2.4)	42.0 (↑0.9)
	MLPix.(ϵ6)	4	20.7 (↑0.3)	43.4 (↑2.3)
	SRB(ϵ4)	4	19.2 (↓1.2)	**43.7** (↑2.6)

Event-based Tracking Performance

Architecture	Timesteps	Param. (M)	FE108 [69]		VisEvent [70]	
			AUC(%)	PR(%)	AUC(%)	PR(%)
SD-Track(Tiny) [18]	4×1	19.61	56.7	89.1	**35.4**	48.7
+MLPixer ($\epsilon=4$)	4×1	20.21	57.1	89.2	33.7	47.3
+MLPixer ($\epsilon=6$)	4×1	22.99	57.9	**90.1**	34.5	**48.9**
+SRB ($\epsilon=4$)	4×1	21.43	**58.2**	88.5	33.8	48.0

This paper presents ST-ERF as a novel framework for analyzing the spatial-temporal modeling behaviors in SNNs from a new perspective. Through this analysis, an inherent limitation in current Transformer-based SNN models is identified when applied to visual long-sequence modeling tasks. To address this limitation, two channel-mixer architectures, MLPixer and SRB, are proposed. Overall, the proposed ST-ERF framework offers valuable insights for the design and optimization of SNN architectures across a wide range of tasks.