

EG-SPIKEFORMER: EYE-GAZE GUIDED TRANSFORMER ON SPIKING NEURAL NETWORKS FOR MEDICAL IMAGE ANALYSIS

Yi Pan^{1,*}, Hanqi Jiang^{1,*}, Junhao Chen¹, Yiwei Li¹, Huaqin Zhao¹, Yifan Zhou¹, Peng Shu¹
Zihao Wu¹, Zhengliang Liu¹, Dajiang Zhu², Xiang Li³, Yohannes Abate⁴, Tianming Liu^{1,†}

¹School of Computing, University of Georgia, Athens, GA, USA

²Department of Computer Science and Engineering, University of Texas at Arlington, TX, USA

³Department of Radiology, Massachusetts General Hospital and Harvard Medical School, MA, USA

⁴Department of Physics and Astronomy, University of Georgia, Athens, GA, USA

ABSTRACT

Neuromorphic computing has emerged as a promising energy-efficient alternative to traditional artificial intelligence, predominantly utilizing spiking neural networks (SNNs) implemented on neuromorphic hardware. Significant advancements have been made in SNN-based convolutional neural networks (CNNs) and Transformer architectures. However, neuromorphic computing for the medical imaging domain remains underexplored. In this study, we introduce EG-SpikeFormer, an SNN architecture tailored for clinical tasks that incorporates eye-gaze data to guide the model’s attention to the diagnostically relevant regions in medical images. Our developed approach effectively addresses shortcut learning issues commonly observed in conventional models, especially in scenarios with limited clinical data and high demands for model reliability, generalizability, and transparency. Our EG-SpikeFormer not only demonstrates superior energy efficiency and performance in medical image prediction tasks but also enhances clinical relevance through multi-modal information alignment. By incorporating eye-gaze data, the model improves interpretability and generalization, opening new directions for applying neuromorphic computing in healthcare.

Index Terms— Neuromorphic computing, spiking neural networks, eye-gaze, medical image analysis, healthcare applications.

1. INTRODUCTION

Spiking Neural Networks (SNNs) have garnered significant attention for their bio-inspired properties and low power consumption [1, 2]. In computer vision, SNNs offer enhanced energy efficiency and interpretability compared to traditional Convolutional Neural Networks (CNNs) and Transformer architectures. However, their adoption is often hindered by lower accuracy. To overcome this limitation, researchers have

combined SNNs with Artificial Neural Networks (ANNs) to leverage the strengths of both model types. Notably, integrations of SNNs with traditional CNNs [3, 4] and advanced Transformer models [5, 6] have been explored. In many Transformer-based approaches, spike neurons replace standard neurons within the architecture, which can impede the full realization of SNNs’ low-power advantages. Recent studies have proposed spike-driven transformers employing Spike-Driven Self-Attention, reducing energy consumption by utilizing linear computations at token and channel levels [7, 8].

Despite these advancements, the application of SNNs in medical image processing remains limited. Some exceptions include the use of reservoir SNNs combined with salient feature extraction and time encoding for breast cancer image recognition, achieving high accuracy across multiple datasets [9]. Another study introduced a spiking convolutional neural network (SCNN) for photon-based imaging, converting time-of-flight data into high-resolution 3D images with superior accuracy and energy efficiency compared to conventional methods [10]. Additionally, SNNs have been applied to image segmentation using adaptive synaptic weights to effectively process noisy medical images [11]. Nevertheless, the integration of SNNs in medical imaging is constrained by challenges in training and scalability relative to traditional deep learning techniques.

In this paper, we propose a novel gaze-guided spike-driven hybrid model EG-SpikeFormer for medical diagnosis for the first time, which integrates the low-power computation benefits of SNN with the powerful feature extraction capabilities of Transformers. The model incorporates radiologists’ eye-gaze data as prior information during training, effectively guiding the model’s attention. Our experiments on two public medical datasets demonstrate the model’s superior performance in both energy efficiency and diagnostic accuracy. The main contributions of our work are as follows:

- To the best of our knowledge, this is the first attempt to introduce a hybrid SNN model combining CNN and Transformer in the field of medical diagnosis, leveraging both

*Equal Contribution.

†Corresponding Author.

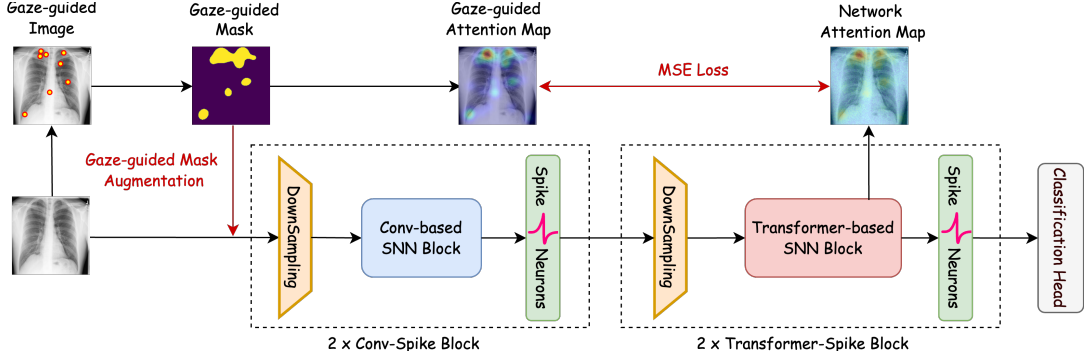


Fig. 1: Overall Architecture of the proposed framework.

low power consumption and strong interpretability.

- By incorporating radiologists’ prior information during training, the model learns to focus on disease-relevant regions, reducing irrelevant information and significantly improving performance and interpretability in medical image diagnosis tasks.
- We firstly propose a hardware-aware co-design framework that integrates neuromorphic computing and eye-gaze guidance, addressing medical challenges such as shortcut learning and data scarcity. This collaborative design not only enhances diagnostic accuracy and energy efficiency but also underscores the significance of such synergies in advancing practical healthcare applications.

2. METHOD

The overall architecture of the proposed framework is illustrated in Figure 1. This framework integrates eye-gaze data with a hybrid SNN to optimize both spatial and temporal feature extraction.

2.1. Blocks with Spike Neurons

Our EG-SpikeFormer utilize Leaky Integrate-and-Fire (LIF) neurons [12] as fundamental units in both convolution-based and Transformer-based SNN blocks. The LIF model maintains a membrane potential that accumulates incoming signals and decays over time. When this potential reaches a pre-defined threshold, the neuron emits a spike and resets. The dynamics of our LIF neurons are mathematically described by:

$$V_s = V_{t-1} + \omega x, \quad (1)$$

$$S_t = \begin{cases} 1, & \text{if } V_s \geq V_{th} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$V_t = S_t V_{reset} + (1 - S_t)(1 - \lambda)V_s \quad (3)$$

Here, V_{t-1} is the membrane potential from the previous time step, ω is the synaptic weight, x is the input signal, V_s is

the membrane potential after receiving input at time t , and S_t is the binary output spike. Upon reaching or exceeding the threshold V_{th} , the neuron fires ($S_t = 1$) and resets its potential to V_{reset} ; otherwise, the potential decays with leakage factor λ . The convolution-based SNN block, illustrated

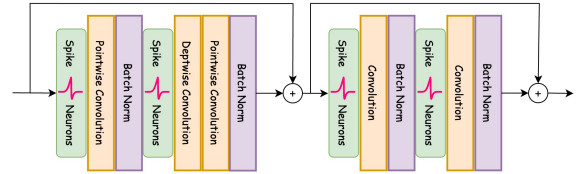


Fig. 2: Architecture of the convolution-based SNN block.

in Figure 2, integrates LIF neurons with convolutional operations to enhance feature extraction and energy efficiency. It consists of spike neurons (SN), pointwise convolutions (Conv_p), depthwise convolutions (Conv_d), and batch normalization (BN) layers. Spiking activations introduce sparsity, reducing computational overhead while effectively representing spatio-temporal features. The membrane shortcut allows direct input propagation and is formulated as:

$$x' = x + \text{BN}(\text{Conv}_p(\text{Conv}_d(\text{SN}(x)))) \quad (4)$$

The Transformer-based SNN block, depicted in Figure 3, integrates the attention mechanisms of traditional Transformers with the spiking neurons. In this block, spiking neurons generate spike-based queries Q_s , keys K_s , and values V_s , which are essential components of the self-attention mechanism. These spike-based representations are obtained by processing incoming spike signals through re-parameterized convolution layers that culminate with spiking neurons. Specially, for a input X to the transformer-based SNN block, the queries, keys, and values are computed as

$$\{Q_s, K_s, V_s\} = \text{SN}(\text{RepCon}(X, W)) \quad (5)$$

The $W \in \{W_q, W_k, W_v\}$ are learnable weight matrices for the respective query, key, and value convolutions. Each spiking neuron $\text{SN}(\cdot)$ accumulates input over time and fires when

its membrane potential crosses a threshold, creating sparse, event-driven representations for Q , K , and V . The attention output is then computed using the spike attention mechanism with a shortcut.

$$X' = X + \text{RepCon}\left(\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right) V_s\right) \quad (6)$$

By combining these elements, the block effectively leverages both spatial and temporal processing capabilities, making it well-suited for dynamic or event-driven data tasks.

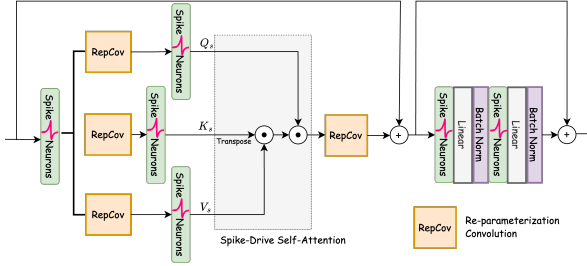


Fig. 3: Architecture of the Transformer-based SNN block.

2.2. Eye-gaze Guided Spike Vision Transformer

Although SNNs optimize the energy efficiency of network architectures, the accuracy of SNN-based Transformers is generally lower than that of traditional Vision Transformer (ViT). SNNs inherently possess temporal and spatial characteristics, but current research predominantly focuses on natural images. In small-scale medical datasets, eye-gaze data, which contains temporal information, can serve as valuable prior knowledge for SNNs, guiding network convergence. Therefore, we designed two strategies to leverage this insight.

We augment the original images by integrating eye-gaze data to highlight key regions without losing information from other areas through the gaze mask (GM) method. Let $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ denote the original image, and let $\mathbf{M} \in \mathbb{R}^{H \times W}$ be the normalized eye-gaze mask with values between 0 and 1. The enhanced image \mathbf{I}' is computed as:

$$\mathbf{I}' = \mathbf{I} \odot (\mathbf{1} + \alpha \mathbf{M}) \quad (7)$$

where \odot denotes element-wise multiplication, $\mathbf{1}$ is an all-ones matrix of compatible dimensions, and α is a hyperparameter controlling the emphasis on eye-gaze regions. This operation amplifies pixel intensities in areas with higher eye-gaze values, effectively guiding the network's focus. To align the model's attention with human gaze patterns, we introduce an attention alignment (ALH) loss. Let $\mathbf{A}_t \in \mathbb{R}^{N \times N}$ be the model's attention map from the last layer of the final Transformer block, and $\mathbf{A}_g \in \mathbb{R}^{N \times N}$ be the attention map derived from eye-gaze data, where N is the number of tokens. The attention alignment loss $\mathcal{L}_{\text{align}}$ is defined as:

$$\mathcal{L}_{\text{align}} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left(\mathbf{A}_t^{(i,j)} - \mathbf{A}_g^{(i,j)} \right)^2 \quad (8)$$

This loss encourages the Transformer module to focus on diagnostically relevant regions, improving convergence speed and performance. The total loss function combines the standard classification loss \mathcal{L}_{cls} with the attention alignment loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{align}} \quad (9)$$

where λ is a weighting factor balancing the two loss components.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1. Datasets

To evaluate the proposed model, we utilized two public clinical datasets: INbreast [13] and SIIM-ACR [14, 15]. The INbreast dataset comprises 410 full-field digital mammography images from low-dose X-ray breast examinations, classified into normal (302 cases), benign (37 cases), and malignant (71 cases) based on BI-RADS assessments [16]. Eye movement data were collected from diagnoses made by a radiologist with 10 years of experience. The SIIM-ACR dataset includes 1,170 images, with 268 cases of pneumothorax, and eye gaze data were obtained from three experienced radiologists [14, 15]. For the INbreast dataset, patients were randomly split into 80% for training and 20% for testing. To balance and diversify the training set, random cropping and contrast-related augmentations were applied, resulting in 482 normal, 512 benign mass, and 472 malignant mass samples. During testing, images from the remaining 20% were processed using a sliding window with a size of 1,024 and a stride of 512. In the SIIM-ACR dataset, all images were 1,024×1,024 pixels and were directly input into the model without additional cropping.

3.2. Results and Analysis

We extensively evaluated our proposed model on the INbreast [13] and SIIM-ACR [15] datasets, comparing it with various existing methods. To assess how well the model's attention aligns with human gaze patterns, we calculated the Structural Similarity Index (SSIM) between the model's attention maps and the gaze data. Following Chen et al. [22], we also computed the energy consumption for each model, demonstrating that our approach maintains competitive performance while being energy-efficient. The total energy consumption is expressed as a combination of the energy required for both multiply-accumulate (MAC) and accumulate (AC) operations [23]:

$$\text{Energy} = E_{\text{MAC}} \times \text{FLOPs}(c) + E_{\text{AC}} \times \text{SOPs}(s) \quad (10)$$

where $E_{\text{MAC}} = 4.6$ pJ and $E_{\text{AC}} = 0.9$ pJ represent the energy per MAC and AC operation implemented on the 45nm hardware [24], respectively; FLOPs and SOPs denote the number of floating-point and spike-based operations at the CNN layers c and SNN layers s .

Method	Spike	Power (mJ)	Param (M)	INbreast [13]				SIIM-ACR [15]			
				Acc. \uparrow	AUC \uparrow	F1 \uparrow	SSIM \uparrow	Acc. \uparrow	AUC \uparrow	F1 \uparrow	SSIM \uparrow
ResNet-50 [17]	\times	19.0	26	89.19	86.62	80.51	0.276	84.65	71.97	83.83	0.153
ResNet-101 [17]	\times	36.1	45	90.54	87.84	88.14	0.302	84.80	75.55	84.75	0.158
EfficientNet [18]	\times	56.6	119	91.46	86.62	88.86	0.352	82.66	74.37	79.56	0.197
Swin-T [19]	\times	13.8	28	91.80	88.10	90.27	0.227	84.62	74.82	84.67	0.159
ViT [20]	\times	81.0	86	92.07	87.13	86.96	0.395	84.00	70.76	83.03	0.205
MS-Res-SNN [21]	\checkmark	10.2	77	80.68	71.92	69.74	0.178	83.59	69.92	81.37	0.202
ResNet-50+Gaze	\times	19.0	26	90.54	86.93	86.00	0.208	84.80	75.00	83.71	0.153
ResNet-101+Gaze	\times	36.1	45	91.88	88.89	88.19	0.244	84.80	72.68	82.89	0.254
ViT+Gaze	\times	81.0	86	93.12	92.30	91.83	0.402	85.60	75.30	85.35	0.280
Ours-T2	\checkmark	26.7	55	93.18	91.76	92.20	0.398	87.89	75.17	85.39	0.273
Ours-T4	\checkmark	52.4	55	94.32	93.00	95.53	0.427	88.28	76.20	86.33	0.285

Table 1: Performance Comparison on INbreast and SIIM-ACR Datasets. The Accuracy (Acc.), ROC AUC (AUC), and F1-score (F1) metrics are reported. **Red** and **blue** denote the best and second-best results.

As shown in Table 1, our eye-gaze guided SNN architecture, EG-SpikeFormer achieves state-of-the-art performance across all metrics, outperforming classical ViTs, eye-gaze augmented ViTs, and CNNs. Besides lower energy consumption, our model enhances performance, demonstrating efficiency and superior capability in complex medical imaging tasks. Leveraging the spatio-temporal event-driven computation of SNNs, our architecture balances high performance with low energy consumption without compromising accuracy. It surpasses traditional deep learning architectures, highlighting the potential of SNNs in resource-constrained medical environments. Integrating eye-gaze data effectively guides the model’s attention, improving accuracy and convergence speed without significantly increasing computational resources. We evaluated two training processes: a two-step and a four-step. Even with two-step training, our model matches or exceeds other state-of-the-art methods. The four-step training further enhances capabilities, indicating that additional steps can unlock more potential without substantially increasing energy consumption.

3.3. Ablation Study

Ablation			INbreast [13]			SIIM-ACR [15]		
GM	ALH	T4	Acc. \uparrow	AUC \uparrow	F1 \uparrow	Acc. \uparrow	AUC \uparrow	F1 \uparrow
\checkmark		\checkmark	92.05	86.34	85.35	85.55	71.28	82.02
	\checkmark	\checkmark	93.08	86.57	92.15	87.11	74.39	84.60
\checkmark	\checkmark		93.18	91.76	92.20	87.89	75.17	85.39
\checkmark	\checkmark	\checkmark	94.32	93.00	95.53	88.28	76.20	86.33

Table 2: Ablation study on the INbreast and SIIM-ACR datasets. The ablation experiments test the impact of removing GM, AM, and T4 modules. Accuracy (Acc.), AUC and F1-score (F1) are reported. **Bold** indicates the best result, and \uparrow means higher values are better.

We conducted an ablation study to assess the impact of GM, ALH, and the four-step training process (T4) on model performance. The detailed results are presented in Table 2.

The model without T4 also showed lower performance metrics. The highest performance was achieved when all components were included, confirming that GM, ALH, and T4 each contribute significantly to the model’s effectiveness.

4. DISCUSSION

Neuromorphic chips, leveraging SNNs as the most accessible approach, offer transformative potential for clinical diagnostics by addressing key challenges such as energy efficiency, data scarcity, and interpretability [25, 26]. EG-SpikeFormer tackles shortcut learning and focuses on diagnostically relevant features, making it ideal for real-time medical applications in resource-constrained settings and offering a forward-thinking framework for medical neuromorphic chip design. Additionally, its eye-gaze-driven, spatio-temporal event processing optimizes neuromorphic systems for clinical tasks. By enhancing diagnostic accuracy and promoting algorithm-hardware co-design, EG-SpikeFormer facilitates the development of efficient, high-performance medical chips. Furthermore, its ability to reduce data demands and computational overhead accelerates energy-efficient and scalable chip design, advancing accessible clinical diagnostics and expanding the role of neuromorphic technologies in healthcare.

5. CONCLUSION

EG-SpikeFormer integrates eye-gaze data with SNN to enhance medical image analysis, utilizing the energy efficiency and spatio-temporal dynamics of SNN alongside the feature extraction strengths of CNN and Transformers. By directing attention on clinically relevant regions, the model achieves superior diagnostic accuracy, interpretability, and generalizability, and addresses the limitations of conventional models. These results highlight neuromorphic computing’s potential in healthcare, providing a high-performance, low-energy solution, particularly well-suited for medical environments with critical data constraints and accuracy requirements.

References

- [1] Tavanaei et al., “Deep learning in spiking neural networks,” *Neural networks*, vol. 111, pp. 47–63, 2019.
- [2] Yamazaki et al., “Spiking neural networks and their applications: A review,” *Brain Sciences*, vol. 12, no. 7, pp. 863, 2022.
- [3] Fang et al., “Deep residual learning in spiking neural networks,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 21056–21069, 2021.
- [4] Zheng et al., “Going deeper with directly-trained larger spiking neural networks,” in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 11062–11070.
- [5] Zhou et al., “Spikformer: When spiking neural network meets transformer,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [6] Yao et al., “Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [7] Zhou et al., “Spikformer: When spiking neural network meets transformer,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [8] Yao et al., “Spike-driven transformer,” *Advances in neural information processing systems*, vol. 36, 2024.
- [9] Fu et al., “Breast cancer recognition using saliency-based spiking neural network,” *Wirel. Commun. Mob. Comput.*, vol. 2022, Jan. 2022.
- [10] Kirkland et al., “Imaging from temporal data via spiking convolutional neural networks,” in *Emerging Imaging and Sensing Technologies for Security and Defence V; and Advanced Manufacturing Technologies for Micro- and Nanosystems in Security and Defence III*. SPIE, 2020, vol. 11540, pp. 66–85.
- [11] Zheng et al., “Image segmentation method based on spiking neural network with adaptive synaptic weights,” in *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2019, pp. 1043–1049.
- [12] Maass, “Networks of spiking neurons: The third generation of neural network models,” *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, September 1997.
- [13] Moreira et al., “Inbreast: toward a full-field digital mammographic database,” *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.
- [14] Saab et al., “Observational supervision for medical image classification using gaze data,” in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 603–614.
- [15] Zawacki et al., “Siim-acr pneumothorax segmentation,” 2019.
- [16] Liberman and Menell, “Breast imaging reporting and data system (bi-rads),” *Radiologic Clinics*, vol. 40, no. 3, pp. 409–430, 2002.
- [17] He et al., “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [18] Tan and Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [19] Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.
- [20] Vaswani, “Attention is all you need,” *Advances in Neural Information Processing Systems*, 2017.
- [21] Yao et al., “Attention spiking neural networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 8, pp. 9393–9410, 2023.
- [22] Chen et al., “Training full spike neural networks via auxiliary accumulation pathway,” *arXiv preprint arXiv:2301.11929*, 2023.
- [23] Yao et al., “Attention spiking neural networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9393–9410, 2023.
- [24] Horowitz, “1.1 computing’s energy problem (and what we can do about it),” in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 10–14.
- [25] Aboumerhi et al., “Neuromorphic applications in medicine,” *Journal of Neural Engineering*, vol. 20, no. 4, pp. 041004, aug 2023.
- [26] Schuman et al., “Opportunities for neuromorphic computing algorithms and applications,” *Nature Computational Science*, vol. 2, no. 1, pp. 10–19, 2022.