Spiking Meets Attention: Efficient Remote Sensing Image Super-Resolution with Attention Spiking Neural Networks

Yi Xiao, Qiangqiang Yuan, *Member, IEEE,* Kui Jiang, *Member, IEEE,* Wenke Huang, Qiang Zhang, Tingting Zheng, Chia-Wen Lin, *Fellow, IEEE,* and Liangpei Zhang, *Fellow, IEEE.*

***Abstract*—Spiking neural networks (SNNs) are emerging as a promising alternative to traditional artificial neural networks (ANNs), offering biological plausibility and energy efficiency. Despite these merits, SNNs are frequently hampered by limited capacity and insufficient representation power, yet remain underexplored in remote sensing super-resolution (SR) tasks. In this paper, we first observe that spiking signals exhibit drastic intensity variations across diverse textures, highlighting an active learning state of the neurons. This observation motivates us to apply SNNs for efficient SR of RSIs. Inspired by the success of attention mechanisms in representing salient information, we devise the spiking attention block (SAB), a concise yet effective component that optimizes membrane potentials through inferred attention weights, which, in turn, regulates spiking activity for superior feature representation. Our key contributions include: 1) we bridge the independent modulation between temporal and channel dimensions, facilitating joint feature correlation learning, and 2) we access the global self-similar patterns in large-scale remote sensing imagery to infer spatial attention weights, incorporating effective priors for realistic and faithful reconstruction. Building upon SAB, we proposed SpikeSR, which achieves state-of-the-art performance across various remote sensing benchmarks such as AID, DOTA, and DIOR, while maintaining high computational efficiency. The code of SpikeSR will be available upon paper acceptance.**

***Index Terms*—Image super-resolution, spiking neural network, attention mechanism, remote sensing.**

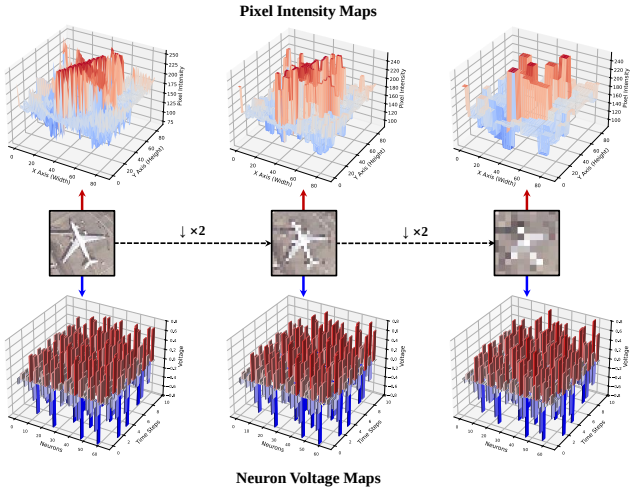

Fig. 1. The visualization of pixel intensity and neuron voltage in images under various degradation factors reveals important insights. The pixel intensity map illustrates that the high-frequency components of the image tend to be smooth, indicating a reduction in sharp details during progressive downsampling. Neuron intensity maps, derived from a LIF model, show that high-frequency details persist with drastic fluctuations, suggesting that the neurons remain in an active state.

I. Introduction

High-resolution remote sensing images (RSIs) contain fine-grained object structures and textures, which are critical for accurate interpretation in downstream tasks [1], [2], [3]. However, limited by the intrinsic resolution of airborne sensors, RSI can merely capture partial spatial details, resulting in sub-optimal scene representation and visual quality. Image super-resolution (SR) aims to alleviate this problem by reconstructing high-resolution (HR) images from low-resolution (LR) observations [4], [5]. Despite this, SR remains a challenging ill-posed issue, as a degraded input may correspond to multiple plausible outputs.

Early efforts rely on hand-crafted priors to tame the ill-posedness, *e.g.*, nonlocal mean [6], [7] and gradient profile [8], but they are often trapped in limited performance and scalability. Recent advances in artificial neural networks (ANNs), *e.g.*, CNNs and Transformers, have witnessed remarkable progress in SR with large-capacity models [9], [10], [11], [12], [13]. However, they often come with a trade-off of increased computational overhead and growing storage costs, making them less efficient in practical scenarios, particularly when reconstructing large-scale RSIs.

More recently, brain-inspired spiking neural networks (SNNs), as the third generation of neural networks, have emerged as a promising alternative for energy-efficient intelligence [14], [15], [16]. Different from ANNs that encode features as continuous values, SNN can emulate biological communication with discrete spiking signals and propagate them by neurons, thus enjoying lower power consumption. As depicted in Fig. 1, our experiments reveal a novel finding that spiking neurons maintain an active learning state across LR RSIs, even in severely damaged textures. Specifically, we observed that degraded RSIs exhibit smoothed pixel intensities and obscured sharp details, posing a significant challenge to characterize high-frequency representations. In contrast, spiking signals retain drastic responses and pronounced spike rates, highlighting that neurons remain in an active learning state. This naturally arises a question: *Can SNNs leverage their inherent properties to handle image degradation for efficient yet high-quality RSI SR?*

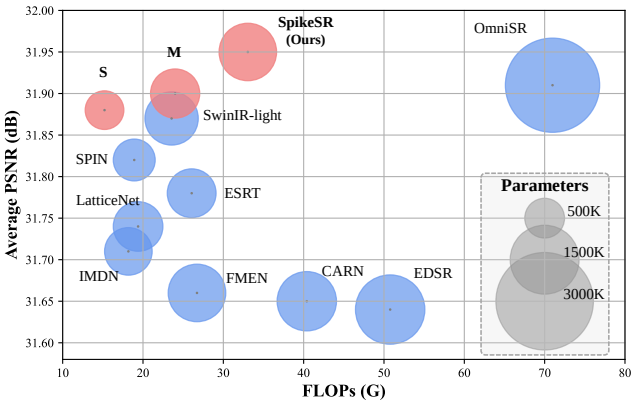

Fig. 2. FLOPs and PSNR performance comparison. The circle sizes represent the number of parameters. Our SpikeSR outperforms SOTA efficient SR methods with high efficiency. PSNR results are averaged on the AID, DOTA, and DIOR datasets.

In fact, to effectively grasp complex and diverse spatial details in RSIs, the network must possess adequate capacity and representation power. Unfortunately, there are two critical challenges when adapting SNNs for SR tasks. Firstly, **spiking activity in SNNs inevitably causes pixel-wise information loss,** which hampered the representation capacity of SNNs, especially when the network deepens. This stems from the discrete nature of binary spiking signals, leading to undesirable spiking degradation problems [17], [18]. Secondly, **SNNs remain constrained by suboptimal membrane potential dynamics,** restricting effective exploration of global context during spiking communications. This necessitates a customized strategy to optimize membrane potentials, but is barely explored before.

To address these limitations, we propose SpikeSR, an SNN-based framework inspired by human visual attention mechanisms, which can actively represent image degradation and modulate synaptic weights to focus on salient regions, which, in turn, regulate the spiking activity for improved capacity and representation power. Specifically, SpikeSR employs a concise yet effective spiking attention block (SAB) to optimize feature emphasis through spiking response dynamics, which integrates three key innovations: 1) the combination of CNN and SNN layers to mitigate information loss induced by discrete spiking activity; 2) introducing hybrid dimension attention (HDA) to recalibrate spiking response across both temporal and channel dimensions, facilitating a joint feature correlations learning; 3) accessing global self-similarity patterns in RSIs to infer spatial attention weights, incorporating effective priors for realistic and faithful reconstruction. Compared to state-of-the-art (SOTA) ANN-based efficient SR models, our SpikeSR demonstrates lower model complexity and superior performance, as shown in Fig. 2.

Our contributions are summarized as follows:

- We pioneer an attention spiking neural network for efficient SR of remote sensing imagery, providing a new perspective on developing efficient models in large-scale Earth observation scenarios.
- We devise a concise yet effective spiking attention block, which mitigates the information loss and regulates membrane potentials of spiking activity for improved representation of SNNs.
- Extensive experimental results on various remote sensing datasets demonstrate that our SpikeSR achieves competitive SR performance against state-of-the-art ANNs-based methods.

II. Related Work

Deep Networks for SR. Inspired by the pioneering SR-CNN [19], CNN-based SR methods have achieved remarkable progress, dominating the field for years. They mainly elaborated on the network design to tame the ill-posedness, with notable advances in residual connections [20], [21] and attention mechanisms [22], [23]. For instance, to grasp valuable self-similarity priors, self-similar attention [24] and non-local sparse attention [25] have been developed. However, these methods often suffer from high computational complexity due to exhaustive non-local modeling, making them less efficient in large-scale RSIs.

Recently, transformer-based SR models have demonstrated impressive performance, benefiting from their ability to model long-range dependencies. IPT [26] first introduces Transformers in SR field, but requires massive parameters and laborious pre-training processes. SwinIR [27] effectively reduces the model size by partitioning the image into smaller windows when applying multi-head attention mechanisms, while maintaining favorable performance. Despite these advancements, advanced SR models are often trapped by rising computational overhead and growing storage costs, posing significant concerns in real-world applications, particularly in remote sensing scenarios.

Efficient SR Models. To reduce computational budget, CARN [28] utilizes grouped convolutions and a cascading mechanism to improve the residual architecture. IMDN [29] progressively distills useful information during feature extraction and applies network pruning to further decrease complexity. FMEN [30] optimizes residual modules to accelerate inference. In Transformer-based SR methods, SPIN [31] enhances long-range modeling by combining self-attention with pixel clustering, facilitating interactions between superpixels. HiT-SR [32] expands the self-attention receptive field by applying different window sizes of hierarchal layers. Despite these successes, there is still room to further boost SR performance. Moreover, the potential of energy-efficient SNNs for SR tasks remains largely unexplored.

Spiking Neural Networks. Recent advances in neuromorphic computing have shown the great potential of SNNs in computational efficiency and power as CNNs. Currently, SNNs have been successfully applied to various tasks, such as image classification [14], [33], object detection and tracking [34], [15], optical flow estimation [35], [16], etc.

A common solution to build SNNs is converting pre-trained ANN models [36]. Li *et al.* [37] proposed a layer-wise calibration to minimize activation mismatch during conversion. Ding *et al.* [38] replaced ReLU with the rate norm layer, enabling direct conversion from a trained ANN to an

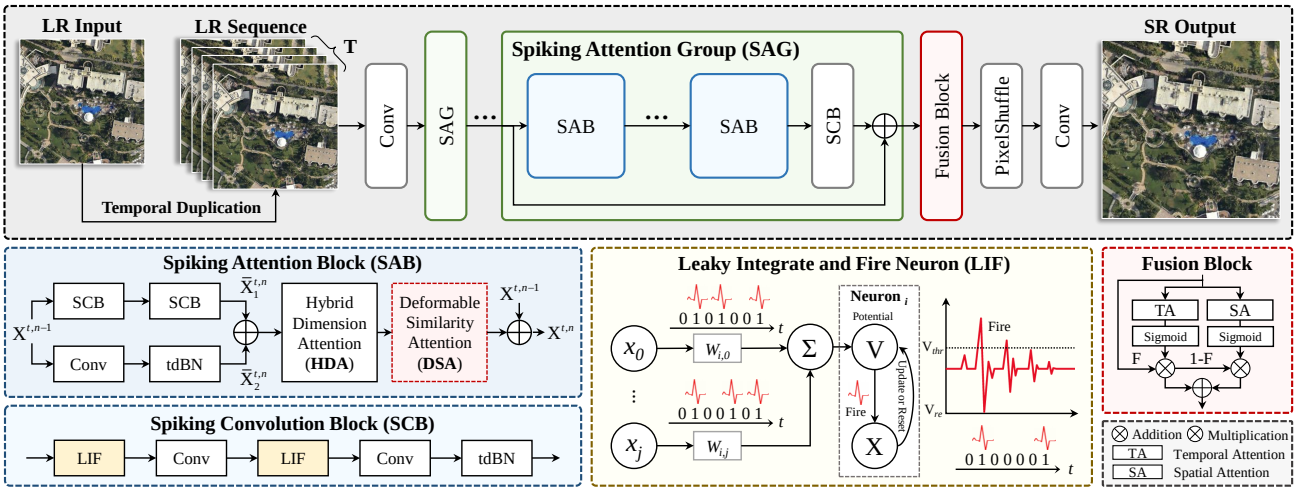

Fig. 3. Overall network architecture of SpikeSR. The LR input is replicated along the temporal dimension and then processed through a convolution to extract shallow features. The core module of SpikeSR is spiking attention groups (SAGs), which employ spiking attention blocks (SABs) to capture deep spiking representations. Each SAB contains three main components: (1) Spiking convolution block (SCB), (2) Hybrid dimension attention (HDA), and (3) Deformable similarity attention (DSA). The fusion block (FB) aggregates the spatial-temporal sequences, and pixelshuffle is used to reconstruct the SR output.

SNN. Stockl *et al.* [39] used time-varying multi-bit spikes to better approximate activation functions. However, conversion-based methods face accuracy gaps and high latency due to extensive time-step simulations, resulting in increased latency and energy consumption.

An alternative involves using agent gradient functions for continuous relaxation of non-smooth spike activities, enabling direct training via backpropagation through time. Lee *et al.* [40] treated membrane potential as a signal to overcome discontinuities, enabling direct training from spikes. Wang *et al.* [41] introduced an iterative LIF model and proposed spatiotemporal backpropagation (STBP) based on approximate peak activity derivatives. Later, Zheng *et al.* [42] proposed temporal delay batch normalization (tdBN), which significantly enhanced the depth of SNNs. To bridge the performance gap between ANNs and SNNs, some methods borrowed insights from CNNs, applying residual learning [18], [43] and attention mechanisms [44], [45] to SNNs. Nonetheless, there has been limited exploration of pixel-level regression tasks, such as SR.

III. Proposed Method

The architecture of the proposed SpikeSR is illustrated in Fig. 3, which mainly consists of spiking attention groups (SAGs). Before SAGs, we utilize a 3×3 convolution to extract high-dimensional features from the LR input. These features are then processed through m stacked SAGs to explore deeper representations. Each SAG includes n spiking attention blocks (SABs), a spiking convolution block (SCB), and a residual connection. In the SCB, leaky integrate-and-fire (LIF) neurons [46], [47] are used to convert the inputs into binary spike sequences (i.e., 0 or 1). As shown in Fig. 3, the output of the LIF neuron is 1 when the membrane potential exceeds the threshold, and 0 otherwise. To optimize the membrane potential, we introduce hybrid dimension attention (HDA), which refines the spiking activity using an efficient temporal-channel joint attention [48]. Furthermore, the proposed deformable similarity attention (DSA) is employed to introduce global context for accurate SR. After the terminal SAG, a fusion block (FB) is utilized to convert the spike sequence features into continuous values. Finally, SpikeSR generates super-resolved output from the fused features by applying pixel-shuffle [49] and a 3×3 convolutional layer.

A. Spiking Attention Block

As evidenced in Fig. 1, regions degraded by different factors exhibit noticeable fluctuations when encoded by LIF, highlighting pronounced firing spike rates of neurons. This provides robust and latent informative spiking cues from LR images. Unlike ANNs that encode images into continuous decimal values, SNNs use discrete binary spike values for neuronal communication, and thus demonstrated undesirable information loss [50], [44], resulting in limited capacity to represent degraded LR images. To address this, the design philosophy of the SAB is focused on leveraging CNNs and attention mechanisms to regulate membrane potentials, facilitating high-quality feature representation for SR, which in turn affects the spiking activity.

As shown in Fig. 3, in particular, the output of the n-th SAB at the t-th time step is denoted as $\mathbf{X}^{t,n}$, and can be obtained by the following:

$$\mathbf{X}^{t,n} = \mathbf{X}^{t,n-1} + \mathrm{DSA}(\mathrm{HDA}(\bar{\mathbf{X}}_1^{t,n} + \bar{\mathbf{X}}_2^{t,n})), \tag{1}$$

where $\bar{\mathbf{X}}_1^{t,n}$ and $\bar{\mathbf{X}}_2^{t,n}$ are two feature representations obtained from parallel branches, defined by:

$$\begin{aligned} \bar{\mathbf{X}}_1^{t,n} &= \mathrm{SCB}(\mathrm{SCB}(\mathbf{X}^{t,n-1})), \\ \bar{\mathbf{X}}_2^{t,n} &= \mathrm{tdBN}(\mathrm{Conv}(\mathbf{X}^{t,n-1})), \end{aligned} \tag{2}$$

where Conv represents a 3×3 convolution layer, and tdBN means the threshold-dependent batch normalization.

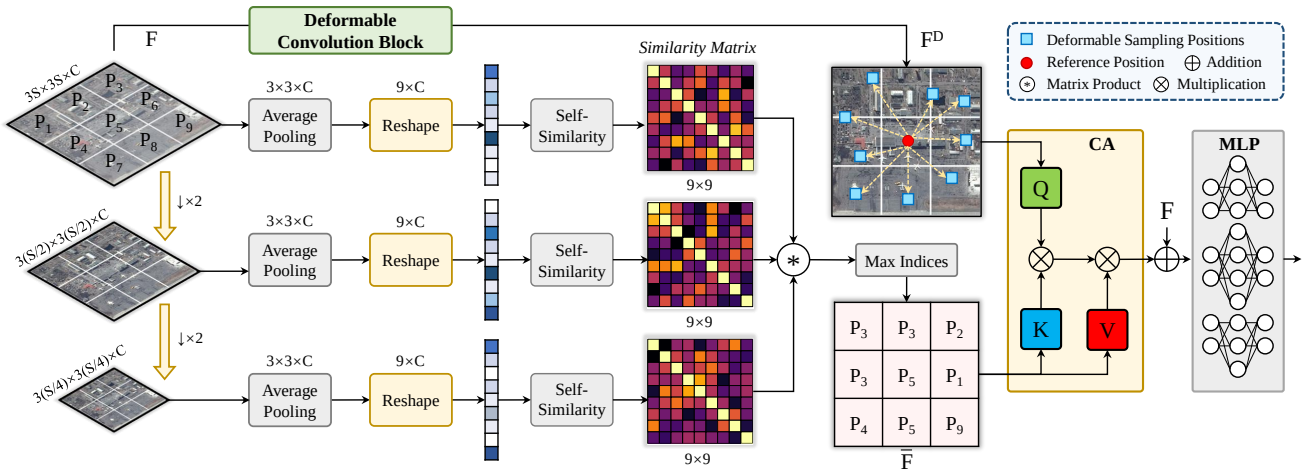

Fig. 4. The illustration of our DSA. Note that we set the diagonal elements of the similarity matrix to zero before selecting the indices of the highest scores. The deformable convolution operates at the patch level, alleviating the mismatch between the most similar patches.

Different from previous works that focus solely on separate temporal and channel modulation [17], [44], [51], SAB enables joint temporal-channel adjustment of the spike response with HDA, effectively achieving interdependencies between the temporal and channel scopes.

B. Deformable Similarity Attention

Non-local self-similarity has been recognized as an effective prior for SR tasks [52]. However, existing non-local attention mechanisms are computationally expensive due to exhaustive non-local operations, which impedes their efficiency in large-scale RSIs. In contrast, the proposed DSA efficiently grasps complex self-similar patterns in RSIs at the patch level to infer intricate spatial weights. Then, we utilize the cross-attention (CA) paradigm to enhance long-range communication, facilitating the fusion of useful context.

The details of DSA are shown in Fig. 4. Since the object scale exhibits explicit diversity in RSIs, the input feature F is downsampled using bilinear interpolation, forming a multi-scale feature pyramid. For clarity, we demonstrate this process by dividing the initial features into 9 patches. The final DSA exploits a cascaded patch division strategy [53]. Specifically, each patch is first average pooled to capture its spatial characteristics, then reshaped and subjected to self-similarity computation, yielding a similarity matrix. The final self-similarity scores are fused via matrix multiplication to enhance the multi-scale representation. The best-matching patch $\bar{\mathbf{P}}_i$ with $\mathbf{P}_i$ can be obtained by:

$$\bar{\mathbf{P}}_i = \underset{\mathbf{P}_j}{\operatorname{argmax}}\, E(\mathbf{P}_i)^{\mathrm{T}} E(\mathbf{P}_j), \quad j \neq i, \tag{3}$$

where we adopt the Gumbel-Softmax [54] to achieve the non-differentiable argmax function, and E means the operation of average pooling and feature reshaping.

Although matched patches contain highly relevant similarity, they are inevitably subject to mismatches and geometric transformations. Hence, we use deformable convolution (DConv) to reduce the generation of hallucinated textures. The deformable feature $\mathbf{F}^{\mathrm{D}}$ at location p_0 is computed as follows:

$$\mathbf{F}^{\mathrm{D}}(p_0) = \sum_{p_m \in \mathcal{R}} \omega(p_m) \cdot \mathbf{F}(p_0 + p_m + \Delta p_m), \tag{4}$$

where $\omega(p_m)$ is the convolution weight at relative location p_m, Δp_m is a 2D vector that represents the learnable offsets, $\mathcal{R}$ is a regular grid that determines the receptive field of the convolution kernel. For a 3×3 kernel, $\mathcal{R} = \{(-1,-1),(-1,0),\cdots,(1,1)\}$. To fuse the self-similar features $\mathbf{F}^{\mathrm{D}}$ with $\bar{\mathbf{F}}$, we embed $\mathbf{F}^{\mathrm{D}}$ to $\mathbf{Q}$, and $\bar{\mathbf{F}}$ to $\mathbf{K}$, $\mathbf{V}$, then perform aggregation by:

$$\bar{\mathbf{V}} = \operatorname{softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}}/\sqrt{d})\mathbf{V}, \tag{5}$$

Finally, the fused features are summarized with the original features $\mathbf{F}$ and fed into the multilayer perceptron (MLP), which contains two fully-connected layers, to obtain the final output:

$$\tilde{\mathbf{F}} = \mathrm{MLP}(\mathbf{F} + \bar{\mathbf{V}}). \tag{6}$$

C. Fusion Block

To transform discrete spiking sequences into continuous pixel values, a common approach is to apply mean sampling along the time dimension. However, this naive process may lead to the loss of crucial spatial details, potentially affecting the SR quality. Therefore, we introduce a fusion block that adaptively aggregates spiking sequences and mitigates information loss. Given an input spike input $\mathbf{Y}$, the computation process of FB can be formulated as:

$$\begin{aligned} \mathbf{Y}_1 &= \sigma(\mathrm{TA}(\mathbf{Y})) \otimes \mathbf{Y}, \\ \mathbf{Y}_2 &= \sigma(\mathrm{SA}(\mathbf{Y})) \otimes (1 - \mathbf{Y}_1), \end{aligned} \tag{7}$$

where TA and SA denote temporal and spatial attention [44], σ means a sigmoid function and $\otimes$ denotes feature multiplication. The final output of FB is obtained by summing $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

TABLE I. Quantitative comparison of SpikeSR with state-of-the-art methods on three remote sensing datasets. FLOPs are measured corresponding to an LR image of 160 × 160 pixels.

Methods	#Param.	FLOPs	AID [55]		DOTA [56]		DIOR [57]		Average	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	-	-	28.86	0.7382	31.16	0.7947	28.57	0.7432	29.53	0.7587
SRCNN [19]	20K	0.512G	29.70	0.7741	32.10	0.8264	29.49	0.7768	30.43	0.7924
VDSR [20]	667K	17.08G	30.44	0.8004	33.22	0.8569	30.36	0.8036	31.34	0.8203
EDSR [21]	1518K	50.77G	30.65	0.8086	33.64	0.8648	30.63	0.8116	31.64	0.8283
CARN [28]	1112K	40.39G	30.66	0.8068	33.66	0.8633	30.64	0.8102	31.65	0.8268
IMDN [29]	715K	18.18G	30.71	0.8076	33.70	0.8641	30.73	0.8115	31.71	0.8277
RFDN-L [58]	681K	16.49G	30.69	0.8074	33.73	0.8642	30.72	0.8114	31.71	0.8277
LatticeNet [59]	777K	19.39G	30.73	0.8089	33.75	0.8653	30.75	0.8126	31.74	0.8289
HNCT [60]	364K	8.48G	30.79	0.8104	33.83	0.8664	30.80	0.8136	31.81	0.8301
FMEN [30]	1046K	26.72G	30.65	0.8063	33.66	0.8631	30.66	0.8104	31.66	0.8266
RLFN [61]	544K	13.25G	30.70	0.8074	33.69	0.8636	30.70	0.8110	31.70	0.8273
ESRT [62]	752K	26.06G	30.77	0.8102	33.75	0.8668	30.81	0.8142	31.78	0.8304
SwinIR-light [27]	897K	23.56G	30.83	0.8114	33.94	0.8677	30.85	0.8149	31.87	0.8313
Omni-SR [63]	2803K	70.98G	30.89	**0.8142**	33.94	0.8695	30.89	0.8170	31.91	0.8336
NGswin [64]	995K	12.73G	30.79	0.8107	33.87	0.8667	30.79	0.8140	31.82	0.8305
SPIN [31]	555K	18.91G	30.78	0.8098	33.85	0.8673	30.82	0.8139	31.82	0.8303
HiT-SR [32]	792K	21.04G	30.87	0.8138	33.93	0.8689	30.89	0.8167	31.90	0.8331
SpikeSR-S (Ours)	472K	15.21G	30.86	0.8126	33.89	0.8687	30.89	0.8162	31.88	0.8325
SpikeSR-M (Ours)	763K	24.00G	30.88	0.8133	33.92	0.8689	30.90	0.8163	31.90	0.8328
SpikeSR (Ours)	1042K	33.05G	**30.91**	**0.8142**	**33.98**	**0.8700**	**30.95**	**0.8175**	**31.95**	**0.8339**

TABLE II. Quantitative comparison of SpikeSR with state-of-the-art methods on 30 scene types of AID datasets.

Scene types	EDSR [21]		CARN [28]		IMDN [29]		ESRT [62]		SwinIR [27]		SPIN [31]		HiT-SR [32]		SpikeSR	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Airport	29.93	0.8282	29.96	0.8264	30.00	0.8270	30.09	0.8292	30.14	0.8307	30.11	0.8295	30.21	0.8325	**30.27**	**0.8336**
Bare Land	36.94	0.8837	36.92	0.8829	36.93	0.8834	36.90	0.8840	36.99	0.8841	36.96	0.8843	**37.00**	**0.8846**	36.98	**0.8846**
Baseball Field	33.05	0.8765	33.06	0.8753	33.17	0.8763	33.21	0.8773	33.27	0.8782	33.14	0.8759	33.29	0.8791	**33.32**	**0.8794**
Beach	34.18	0.8727	34.27	0.8737	34.29	0.8739	34.35	0.8755	34.39	0.8756	34.36	0.8757	34.38	0.8762	**34.41**	**0.8764**
Bridge	32.93	0.8800	32.86	0.8774	32.93	0.8784	33.05	0.8803	33.15	0.8810	33.05	0.8803	33.14	0.8820	**33.27**	**0.8827**
Center	28.77	0.7921	28.71	0.7881	28.79	0.7892	28.88	0.7923	29.00	0.7954	28.88	0.7919	29.03	0.7974	**29.09**	**0.7985**
Church	26.30	0.7469	26.41	0.7449	26.46	0.7467	26.52	0.7489	26.59	0.7512	26.56	0.7507	26.64	0.7549	**26.70**	**0.7560**
Commercial	29.01	0.7940	29.11	0.7944	29.17	0.7958	29.22	0.7975	29.28	0.7996	29.27	0.7989	29.33	0.8021	**29.37**	**0.8033**
D-Residential	24.38	0.6839	24.56	0.6856	24.60	0.6864	24.67	0.6898	24.71	0.6924	24.63	0.6872	24.80	**0.6998**	**24.80**	0.6978
Desert	40.20	0.9268	40.22	0.9259	40.17	0.9264	40.06	0.9276	40.31	0.9275	40.24	0.9279	**40.27**	0.9282	40.25	**0.9283**
Farmland	35.00	0.8683	34.89	0.8656	34.94	0.8667	34.99	0.8679	35.02	0.8681	34.98	0.8678	35.09	0.8696	**35.09**	**0.8700**
Forest	29.85	0.7315	29.88	0.7304	29.90	0.7312	29.99	0.7365	29.98	0.7350	29.97	0.7350	30.05	**0.7395**	**30.05**	0.7380
Industrial	28.88	0.7931	28.84	0.7894	28.88	0.7904	28.98	0.7942	29.03	0.7959	28.98	0.7932	29.11	0.7998	**29.16**	**0.8007**
Meadow	34.63	0.7804	34.53	0.7769	34.55	0.7784	34.63	0.7814	34.66	0.7807	34.64	0.7813	34.58	0.7807	**34.68**	**0.7820**
M-Residential	28.34	0.7365	28.39	0.7347	28.42	0.7349	28.47	0.7370	28.56	0.7401	28.39	0.7334	**28.64**	**0.7443**	28.63	0.7436
Mountain	30.63	0.7885	30.70	0.7892	30.70	0.7895	30.74	0.7909	30.78	0.7921	30.75	0.7911	**30.79**	**0.7930**	**30.79**	**0.7930**
Park	30.54	0.8130	30.63	0.8136	30.65	0.8141	30.72	0.8169	30.76	0.8177	30.73	0.8162	30.81	0.8198	**30.82**	**0.8203**
Parking	27.25	0.8317	27.08	0.8245	27.23	0.8270	27.47	0.8352	27.42	0.8354	27.50	0.8363	27.70	**0.8435**	**27.72**	0.8424
Playground	35.37	0.8943	35.27	0.892	35.42	0.8929	35.47	0.8942	35.49	0.8946	35.45	0.8946	35.59	0.8968	**35.70**	**0.8976**
Pond	32.11	0.8542	32.10	0.8532	32.11	0.8534	32.17	0.8546	32.22	0.8553	32.15	0.8545	32.23	0.8561	**32.25**	**0.8563**
Port	28.50	0.8596	28.61	0.8593	28.67	0.8597	28.75	0.8620	28.81	0.8631	28.79	0.8624	28.85	0.8651	**28.94**	**0.8658**
Railway Station	28.72	0.7738	28.68	0.7699	28.77	0.7718	28.84	0.7744	28.92	0.7777	28.89	0.7759	28.97	0.7802	**29.02**	**0.7816**
Resort	28.52	0.7799	28.59	0.7791	28.62	0.7795	28.68	0.7819	28.74	0.7837	28.66	0.7801	28.78	0.7864	**28.82**	**0.7869**
River	31.55	0.7891	31.55	0.7882	31.57	0.7885	31.60	0.7900	31.64	0.7905	31.61	0.7904	31.66	0.7918	**31.68**	**0.7922**
School	29.36	0.8044	29.41	0.8033	29.45	0.8041	29.51	0.8067	29.59	0.8091	29.50	0.8048	29.67	0.8123	**29.68**	**0.8124**
S-Residential	27.71	0.6728	27.79	0.6723	27.80	0.6725	27.85	0.6752	27.88	0.6758	27.80	0.6728	27.91	**0.6782**	**27.92**	0.6775
Square	30.84	0.8200	30.83	0.8181	30.87	0.8183	30.97	0.8218	31.06	0.8236	30.98	0.821	31.11	0.8256	**31.15**	**0.8266**
Stadium	29.63	0.8387	29.51	0.834	29.62	0.8358	29.74	0.8388	29.82	0.8413	29.76	0.8394	29.80	0.8420	**29.93**	**0.8439**
Storage Tanks	27.44	0.7664	27.50	0.7649	27.52	0.7648	27.58	0.7671	27.63	0.7692	27.58	0.767	27.66	0.7720	**27.68**	**0.7718**
Viaduct	28.99	0.7757	28.92	0.7711	28.96	0.7722	29.06	0.7759	29.14	0.7784	29.05	0.7753	29.16	0.7811	**29.25**	**0.7831**
Average	30.65	0.8086	30.66	0.8068	30.71	0.8076	30.77	0.8102	30.83	0.8114	30.78	0.8098	30.87	0.8138	**30.91**	**0.8142**

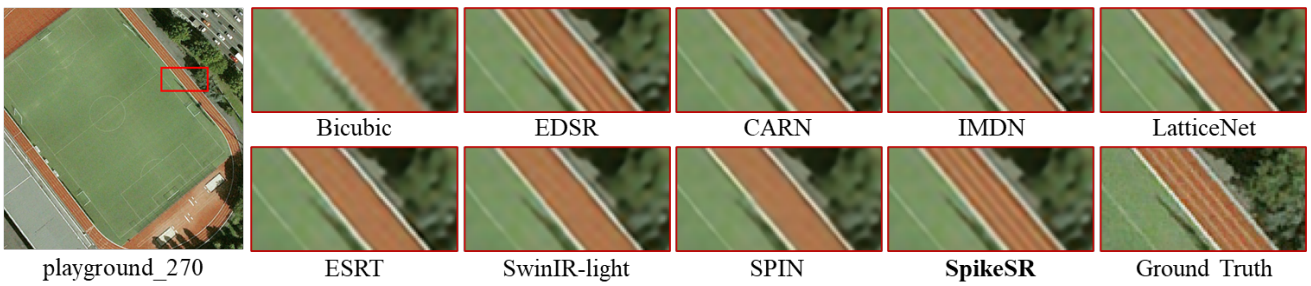

Fig. 5. Qualitative comparison of state-of-the-art efficient models for ×4 SR task on AID test set.

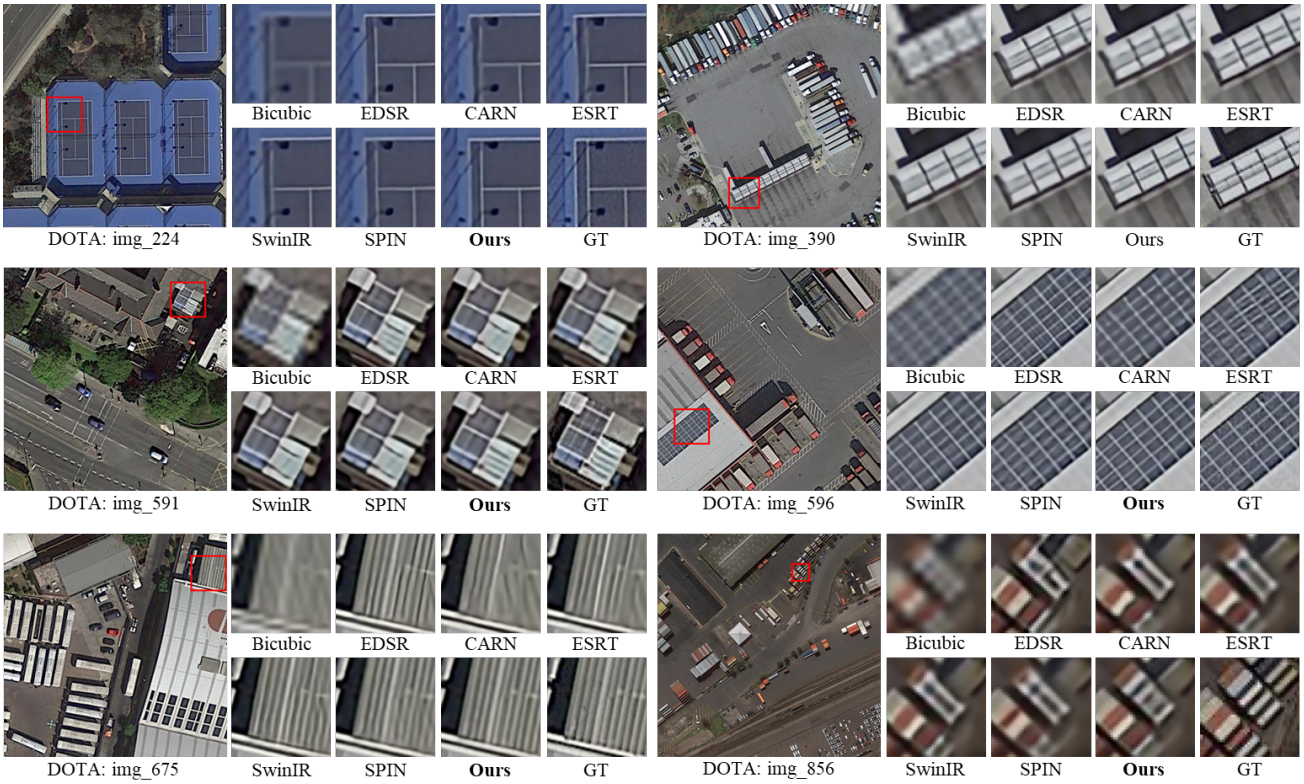

Fig. 6. Qualitative comparison of state-of-the-art efficient SR models for ×4 SR task on DOTA dataset.

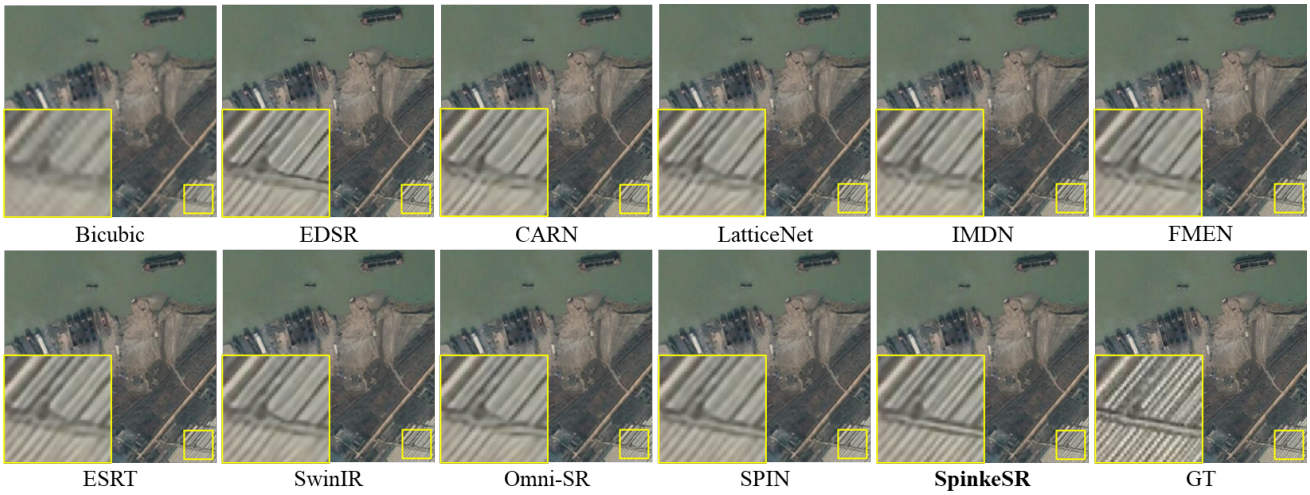

Fig. 7. Qualitative comparison of state-of-the-art efficient SR models for ×4 SR task on DIOR dataset.

IV. Experiments

Datasets. We use the AID dataset [55] as the training set, a large-scale remote sensing benchmark for scene classification, consisting of 30 different scene categories. The AID dataset includes 10,000 HR images, where we randomly select 3,000 for training and 900 for validation. The LR samples are generated by bicubic downsampling. Following TTST [65], we also evaluate our method on the DOTA [56] and DIOR [57] datasets, which contain 900 and 1,000 images, respectively.

Implementation Details. During model training, the learning rate is fixed to 10^{-4}, and the training procedure stops after 1000 epochs with a batch size of 4. Adam optimizer is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Data augmentation includes random rotations of 90°, 180°, 270°, and horizontal flips on 64×64 patches. The channel number, embedding dimension of cross-attention, and MLP rate of small, medium, and final SpikeSR are set to $\{40, 24, 72\}$, $\{56, 24, 72\}$, and $\{64, 32, 100\}$, respectively. The number of SAGs is 4, with 2 SABs in each SAG. Time step is set to $\mathrm{T} = 4$.

Metrics. We use the widely used peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to evaluate SR performance. The results are measured on the Y channel after

TABLE III. Ablation on different variants of our SpikeSR.

Methods	TA	CA	SA	HAD	DSA	#Param.	PSNR
Baseline	✓	✓	✓			1120K	27.80
Varian-A			✓	✓		1062K	27.92
Varian-B	✓	✓			✓	1009K	28.11
SpikeSR				✓	✓	1042K	**28.19**

converting RGB to YCbCr space. For a fair comparison, comparative methods are retained from scratch on the AID dataset, adhering to their official implementation settings.

A. Comparison with Efficient Models

We compare our SpikeSR with state-of-the-art (SOTA) efficient SR methods, including CNN-based models of CARN [28], LatticeNet [59], RLFN [61], etc, and transformer-based approaches of SwinIR [27], SPIN [31], HiT-SR [32], etc. We also report results for the small and medium versions of our SpikeSR, denoted as SpikeSR-S and Spike-M, respectively.

Quantitative comparisons. The quantitative results of various methods are reported in Table I. We can observe that our SpikeSR achieves the best performance across three benchmarks, outperforming SOTA CNN- and transformer-based SR models. For example, on AID, DOTA, and DIOR datasets, SpikeSR improves PSNR by 0.08 dB, 0.04 dB, and 0.1 dB, respectively, compared to the impressive SwinIR. Moreover, the small version of SpikeSR requires fewer parameters (472K vs. 897K) and FLOPs (15.21G vs. 23.56G) than SwinIR, yet achieves superior average performance.

Qualitative comparisons. Fig. 5, Fig. 6, and Fig. 7 present visual comparisons on AID, DOTA, and DIOR datasets, respectively. As shown in Fig. 5, SpikeSR effectively restores severely damaged textures, *e.g.*, the runway line in the playground. By contrast, other SR models fail to recover such weak high-frequency details. In Fig. 6, the reconstruction of "img_591" highlights that SpikeSR produces results closest to the GT, while other methods like SPIN recovers unrealistic results. Moreover, Fig. 7 further demonstrates that SpikeSR consistently delivers superior visual quality, restoring more textural information compared to large-capacity ANN-based model like Omni-SR.

V. Ablation Study

We conduct ablation studies to assess the impact of key components in SpikeSR. The experimental results in Table III are measured on the AID-tiny dataset [65]. In particular, the Baseline model is constructed by removing the HDA and DSA and replacing them with standard temporal attention (TA), channel attention (CA), and spatial attention (SA) mechanisms. For a fair comparison, we increase the number of m, n, and channel dimensions to 8, 4, and 256, respectively, which ensures a similar number of parameters with our SpikeSR. Similarly, those settings of Variant-A are modified to 10, 5, and 128, respectively.

Effectiveness of HDA and DSA. Table III indicates how the SR performance is influenced by the HDA and DSA. Comparing the PSNR values of the Baseline and Variant-A reveals that HDA contributes a 0.12 dB improvement. This suggests that enhancing the correlation between the temporal and channel dimensions delivers better recalibration of the membrane potentials. More intuitively, we visualize the feature maps to highlight the impact of HDA, as shown in Fig. 8. The results illustrate that HDA effectively refines the feature representation by emphasizing salient details and suppressing irrelevant information.

By introducing the proposed DSA to the Baseline model, the PSNR results can be improved by a large margin of 0.21 dB. When employing HDA and DSA simultaneously, the resulting SpikeSR achieves an additional 0.08 dB improvement compared to Variant-B. To better demonstrate its effectiveness in capturing global self-similarity priors, we provide the LAM visualization and diffusion index in Fig. 9. As observed, DSA generates more pronounced LAMs and significantly increases the DI, indicating that the model activates more valuable pixels for accurate SR.

Feature Pyramid and DConv. Due to the scale diversity of objects in RSIs, we constructed a feature pyramid to grasp the self-similarity in multiple levels. As listed in Table IV, the use of feature pyramid improves the performance by 0.05 dB without introducing additional parameters. Furthermore, to demonstrate the effectiveness of deformable convolution, we remove this component, which leads to a severe performance

TABLE IV. Ablation on feature pyramid and deformable convolution.

Methods	w/o Pyramid	w/o DConv	DSA (Full)
#Param.	1042K	918K	1042K
PSNR (dB)	28.14	28.07	**28.19**

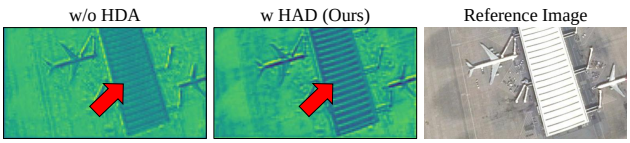

Fig. 8. Feature visualizations. The feature obtained by HDA is sharper and preserves more details, indicating high-quality feature representation.

TABLE V. Performance and complexity analysis of SpikeSR with different numbers of blocks.

Number of m	2	3	4	5	6
#Param.	558K	800K	1042K	1284K	1526K
FLOPs	17.47G	25.26G	33.00G	40.84G	48.60G
PSNR (dB)	28.11	28.17	**28.19**	28.16	**28.20**

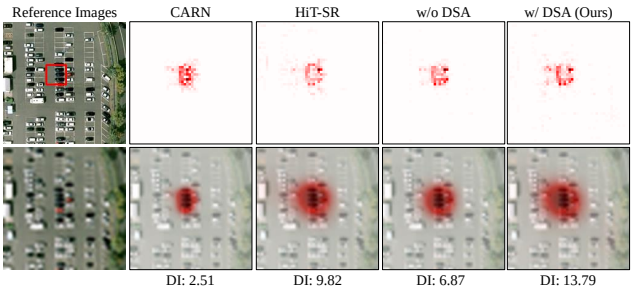

Fig. 9. Analysis of local attribution maps (LAMs) [66] and diffusion index (DI). The proposed DAS helps SpikeSR exploit more useful information against CARN and HiT-SR.

drop by 0.12 dB. This illustrates that the self-similar patches contain massive irrelevant and misaligned contents, and direct fusion may introduce interference, thus resulting in suboptimal performance.

Network Depth. We evaluate the impact of the network depth by changing the number of SAGs of our SpikeSR from 2 to 6 blocks. As reported in Table V, SpikeSR achieves the highest SR performance when $m = 3$. While increasing the number of m may further improve the reconstruction, it also brings larger model size. Therefore, we set $m = 4$ in our final model, considering the trade-off between performance and computational complexity.

VI. Conclusion

In this paper, we investigate the application of SNNs for efficient SR of remote sensing images. Motivated by the observation that LIF neurons exhibit a higher spike rate in degraded images, we integrate SNNs with convolutions for improved feature representation. Besides, a hybrid dimension attention is employed to modulate the spike response, further refining salient information. To incorporate valuable prior knowledge for more accurate SR, we propose a deformable similarity attention module, capturing global self-similarity across multiple feature levels. Extensive experiments on various remote sensing datasets demonstrate the efficacy and effectiveness of the proposed SpikeSR model.

References

[1] M. T. Razzak, G. Mateo-García, G. Lecuyer, L. Gómez-Chova, Y. Gal, and F. Kalaitzis, "Multi-spectral multi-image super-resolution of sentinel-2 with radiometric consistency losses and its effect on building delineation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 1–13, 2023.

[2] R. Dong, S. Yuan, B. Luo, M. Chen, J. Zhang, L. Zhang, W. Li, J. Zheng, and H. Fu, "Building bridges across spatial and temporal resolutions: Reference-based super-resolution via change priors and conditional diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 684–27 694.

[3] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 979–25 991, 2022.

[4] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[5] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *European Conference on Computer Vision*. Springer, 2020, pp. 208–224.

[6] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4544–4556, 2012.

[7] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5344–5353.

[8] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[9] L. Sun, J. Dong, J. Tang, and J. Pan, "Spatially-adaptive feature modulation for efficient image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 190–13 199.

[10] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3147–3155.

[11] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 367–22 377.

[12] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 312–12 321.

[13] M. Zhou, K. Yan, J. Pan, W. Ren, Q. Xie, and X. Cao, "Memory-augmented deep unfolding network for guided image super-resolution," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 215–242, 2023.

[14] Y. Lan, Y. Zhang, X. Ma, Y. Qu, and Y. Fu, "Efficient converted spiking neural network for 3d and 2d classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9211–9220.

[15] Z. Yang, Y. Wu, G. Wang, Y. Yang, G. Li, L. Deng, J. Zhu, and L. Shi, "Dashnet: A hybrid artificial and spiking neural network for high-speed object tracking," *arXiv preprint arXiv:1909.12942*, 2019.

[16] F. Paredes-Vallés, K. Y. Scheper, and G. C. De Croon, "Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2051–2064, 2019.

[17] M. Yao, H. Gao, G. Zhao, D. Wang, Y. Lin, Z. Yang, and G. Li, "Temporal-wise attention spiking neural networks for event streams classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 221–10 230.

[18] W. Fang, Z. Yu, Y. Chen, T. Huang, T. Masquelier, and Y. Tian, "Deep residual learning in spiking neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 056–21 069, 2021.

[19] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[20] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.

[22] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 286–301.

[23] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *European Conference of Computer Vision*, 2020, pp. 191–207.

[24] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[25] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3517–3526.

[26] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[27] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[28] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 252–268.

[29] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proceedings of the ACM International Conference on Multimedia*, 2019, pp. 2024–2032.

[30] Z. Du, D. Liu, J. Liu, J. Tang, G. Wu, and L. Fu, "Fast and memory-efficient network towards efficient image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 853–862.

[31] A. Zhang, W. Ren, Y. Liu, and X. Cao, "Lightweight image super-resolution with superpixel token interaction," in *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, 2023, pp. 12 728–12 737.

[32] X. Zhang, Y. Zhang, and F. Yu, "Hit-sr: Hierarchical transformer for efficient image super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2024, pp. 483–500.

[33] X. Shi, Z. Hao, and Z. Yu, "Spikingresformer: Bridging resnet and vision transformer in spiking neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 5610–5619.

[34] S. Kim, S. Park, B. Na, and S. Yoon, "Spiking-yolo: spiking neural network for energy-efficient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11 270–11 277.

[35] C. Lee, A. K. Kosta, A. Z. Zhu, K. Chaney, K. Daniilidis, and K. Roy, "Spike-flownet: event-based optical flow estimation with energy-efficient hybrid neural networks," in *European Conference on Computer Vision*, 2020, pp. 366–382.

[36] P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *International Joint Conference on Neural Networks*, 2015, pp. 1–8.

[37] Y. Li, S. Deng, X. Dong, and S. Gu, "Error-aware conversion from ann to snn via post-training parameter calibration," *International Journal of Computer Vision*, vol. 132, p. 3586–3609, 2024.

[38] J. Ding, Z. Yu, Y. Tian, and T. Huang, "Optimal ann-snn conversion for fast and accurate inference in deep spiking neural networks," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021, pp. 2328–2336.

[39] C. Stöckl and W. Maass, "Optimized spiking neurons can classify images with high accuracy through temporal coding with two spikes," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 230–238, 2021.

[40] J. H. Lee, T. Delbruck, and M. Pfeiffer, "Training deep spiking neural networks using backpropagation," *Frontiers in neuroscience*, vol. 10, p. 508, 2016.

[41] Z. Wang, Y. Fang, J. Cao, Q. Zhang, Z. Wang, and R. Xu, "Masked spiking transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1761–1771.

[42] H. Zheng, Y. Wu, L. Deng, Y. Hu, and G. Li, "Going deeper with directly-trained larger spiking neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 062–11 070.

[43] Y. Hu, L. Deng, Y. Wu, M. Yao, and G. Li, "Advancing spiking neural networks toward deep residual learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 2353–2367, 2025.

[44] M. Yao, G. Zhao, H. Zhang, Y. Hu, L. Deng, Y. Tian, B. Xu, and G. Li, "Attention spiking neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9393–9410, 2023.

[45] X. Qiu, R.-J. Zhu, Y. Chou, Z. Wang, L.-j. Deng, and G. Li, "Gated attention coding for training high-performance and efficient spiking neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 601–610.

[46] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[47] W. Fang, Y. Chen, J. Ding, Z. Yu, T. Masquelier, D. Chen, L. Huang, H. Zhou, G. Li, and Y. Tian, "Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence," *Science Advances*, vol. 9, no. 40, p. eadi1480, 2023.

[48] R.-J. Zhu, M. Zhang, Q. Zhao, H. Deng, Y. Duan, and L.-J. Deng, "Tcja-snn: Temporal-channel joint attention for spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024.

[49] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[50] S. Kundu, G. Datta, M. Pedram, and P. A. Beerel, "Spike-thrift: Towards energy-efficient deep spiking neural networks by limiting spiking activity via attention-guided compression," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3953–3962.

[51] T. Song, G. Jin, P. Li, K. Jiang, X. Chen, and J. Jin, "Learning a spiking neural network for efficient image deraining," *arXiv preprint arXiv:2405.06277*, 2024.

[52] J.-N. Su, G. Fan, M. Gan, G.-Y. Chen, W. Guo, and C. L. P. Chen, "Revealing the dark side of non-local attention in single image super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 11 476–11 490, 2024.

[53] J. Liu, C. Chen, J. Tang, and G. Wu, "From coarse to fine: Hierarchical pixel integration for lightweight image super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1666–1674.

[54] L. Wang, X. Dong, Y. Wang, X. Ying, Z. Lin, W. An, and Y. Guo, "Exploring sparsity in image super-resolution for efficient inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4917–4926.

[55] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[56] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[57] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[58] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 41–55.

[59] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu, "Latticenet: Towards lightweight image super-resolution with lattice block," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 272–289.

[60] J. Fang, H. Lin, X. Chen, and K. Zeng, "A hybrid network of cnn and transformer for lightweight image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 1103–1112.

[61] F. Kong, M. Li, S. Liu, D. Liu, J. He, Y. Bai, F. Chen, and L. Fu, "Residual local feature network for efficient super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 766–776.

[62] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022, pp. 457–466.

[63] H. Wang, X. Chen, B. Ni, Y. Liu, and J. Liu, "Omni aggregation networks for lightweight image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 378–22 387.

[64] H. Choi, J. Lee, and J. Yang, "N-gram in swin transformers for efficient lightweight image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2071–2081.

[65] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang, "Ttst: A top-k token selective transformer for remote sensing image super-resolution," *IEEE Transactions on Image Processing*, vol. 33, pp. 738–752, 2024.

[66] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9199–9208.