

A Appendix

In the Supplementary Material, we detail the **implementation details of Hybrid Dimension Attention (HDA)**, while **providing more visual results and ablation analysis**.

A.1 Hybrid Dimension Attention

We adhere the Temporal-Channel Joint Attention (TCJA) module [1] to achieve hybrid feature modulation, which has been demonstrated to deliver efficient and effective performance in spiking neural networks (SNNs), thus facilitating the joint optimization of membrane potentials. The overall flow of TCJA is illustrated in Fig. 1. For simplicity, the notation of batch size B is omitted in the following description.

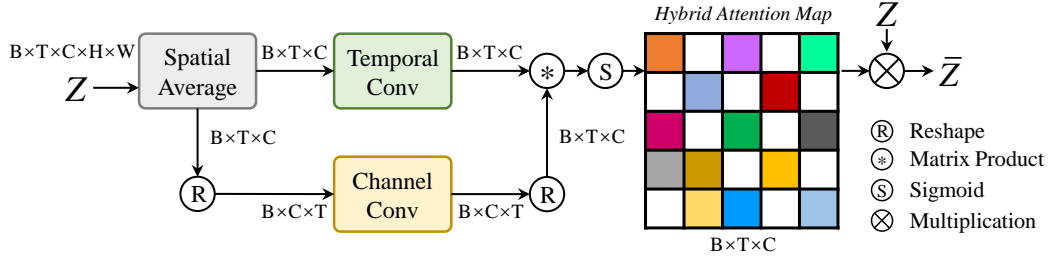


Figure 1: The illustration of the adopted TJCA.

Spatial Average. In particular, to better focus on the temporal and channel-wise correlations, the input spiking sequences $Z \in \mathbb{R}^{T \times C \times H \times W}$ is spatially averaged, which can be implemented by calculated an average matrix $\mathcal{M} \in \mathbb{R}^{C \times T}$:

$$\mathcal{M}_{(c,t)} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W Z_{i,j}^{(c,t)}, \quad (1)$$

where $\mathcal{M}_{(c,t)}$ denotes each element of the averaging matrix \mathcal{M} , and $Z^{(c,t)}$ represents the input feature at the c -th channel and time step t .

Temporal Conv. Next, a 1-D convolution is applied along the channel dimension to each row of the averaging matrix \mathcal{M} . The resulting feature maps from different rows are then aggregated. This temporal convolution process can be formulated as:

$$\mathcal{T}_{i,j} = \sum_{n=1}^C \sum_{m=0}^{K_T-1} W_{(n,i)}^m \mathcal{M}_{(n,j+m)}, \quad (2)$$

where K_T denotes the size of the convolution kernel, specifying the number of time steps involved in the convolution operation. $W_{(n,i)}^m$ represents the m -th learnable parameter for the i -th channel when performing 1-D convolution on the n -th row of the input matrix \mathcal{M} . The resulting attention score matrix after temporal convolution is denoted as $\mathcal{T} \in \mathbb{R}^{C \times T}$.

Channel Conv. To model the correlations between different features and their neighboring channels, we perform a 1-D convolution along the temporal axis (T) for each column of the matrix \mathcal{M} . The outputs from all columns are then summed to produce the final representation. This process can be formulated as:

$$\mathcal{C}_{i,j} = \sum_{n=1}^T \sum_{m=0}^{K_C-1} E_{(n,j)}^m \mathcal{M}_{(i+m,n)}, \quad (3)$$

where K_C denotes the size of the convolution kernel. $E_{(n,j)}^m$ is a learnable parameters that represents the m -th parameter of the j -th channel when performing a 1-D convolution on the n -th row of the input tensor \mathcal{M} . $\mathcal{C} \in \mathbb{R}^{C \times T}$ is the attention score matrix after channel convolution operation.

Finally, to obtain a hybrid attention map from the temporal and channel matrices, i.e., \mathcal{T} and \mathcal{C} , we integrate them through element-wise multiplication:

$$\mathcal{F} = \sigma(\mathcal{T} \cdot \mathcal{C}), \quad (4)$$

where σ denotes the sigmoid activation function. The output of the TJCA module is obtained by modulating Z with the joint attention score \mathcal{F} , enabling the exploration of potential correlations across the joint temporal and channel dimensions.

A.2 Additional Visual Results

We further present visual comparisons across various remote sensing image scenes to demonstrate the superiority and generalizability of SpikeSR. Fig. 2, Fig. 3, and Fig. 4 show the results on the AID, DOTA, and DIOR datasets, respectively, where our method consistently produces the sharpest details and maintains the highest visual fidelity.

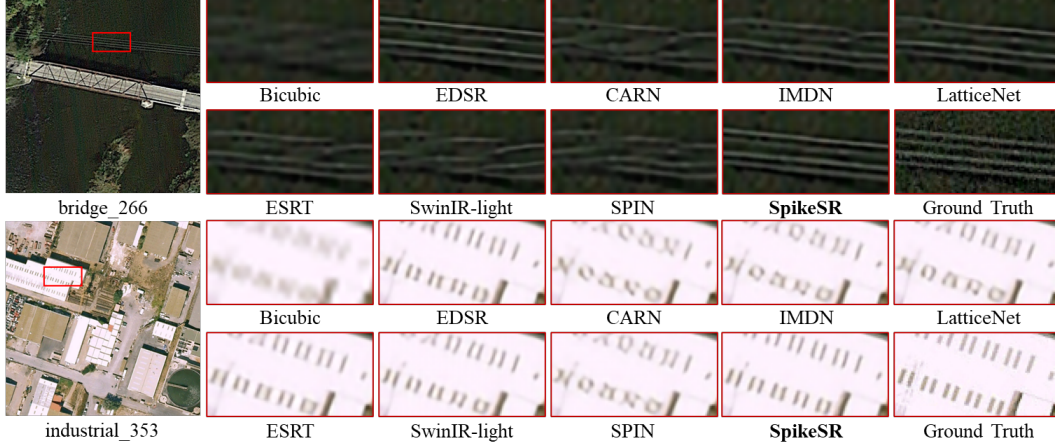


Figure 2: Qualitative comparison of state-of-the-art efficient models for $\times 4$ SR task on AID test set.



Figure 3: Qualitative comparison of state-of-the-art efficient models for $\times 4$ SR task on DOTA test set.

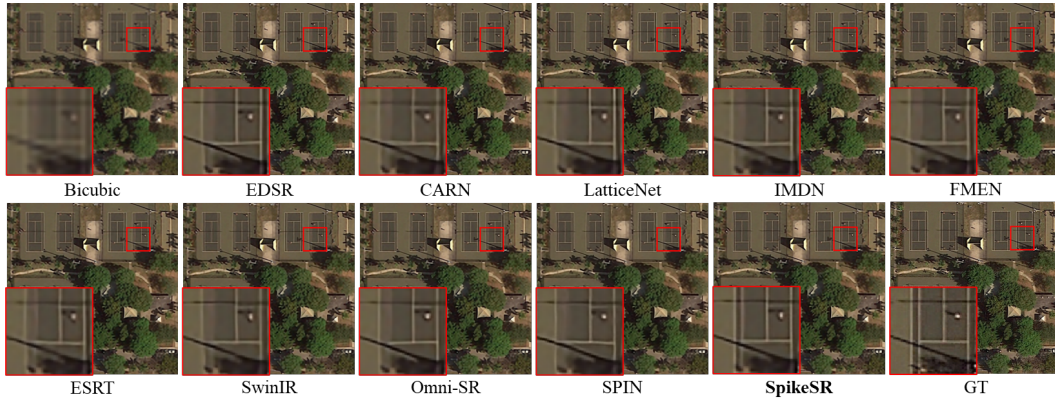


Figure 4: Qualitative comparison of state-of-the-art efficient models for $\times 4$ SR task on DIOR test set.

39 A.3 Additional Ablation Analysis

40 We provide the best-matched patches at the first level of the feature pyramid in our Deformable
41 Similarity Attention (DSA), as shown in Fig. 5. It can be seen that the matched patches share some
42 visually similar patterns, which could potentially facilitate recovering more details.

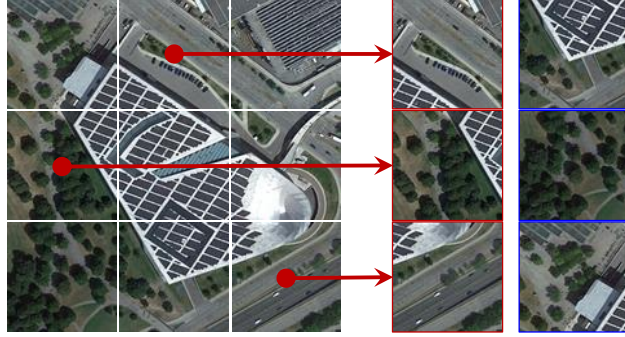


Figure 5: Visualization of patch pairs that best match.

43 References

- 44 [1] Rui-Jie Zhu, Malu Zhang, Qihang Zhao, Haoyu Deng, Yule Duan, and Liang-Jian Deng. Tcja-snn: Temporal-
45 channel joint attention for spiking neural networks. *IEEE Transactions on Neural Networks and Learning*
46 *Systems*, pages 1–14, 2024.