Sparsely Timing the Change: A Spiking Temporal Framework for Remote Sensing Interpretation

Supplementary Material

Algorithm 1 Geo-Spike Interpolation (GSI-P)

Require: Bi-temporal images $I_1, I_2 \in \mathbb{R}^{B\times C\times H\times W}$, temporal step length T.

Ensure: Sparse spike sequence $E' \in \{0,1\}^{B\times T\times 2C\times H\times W}$, geographical response map G, temporal delay τ.

1: Compute the radiometric discrepancy M between I_1, I_2 and perform normalization;
2: Estimate the $G(x, y)$ from I_1, I_2;
3: Apply geographical modulation using the $G(x, y)$ and learnable temporal correction θ on the M to obtain $\tilde{M}'$;
4: Map the $\tilde{M}'$ to $\tau(x, y, c)$.
5: **for** $t = 1$ **to** T **do**
6: Combin $\tau(x, y, c)$, $|I_1-I_2|$, and $\tilde{M}'$ perform polarity modeling $E_t^{\text{on}}, E_t^{\text{off}}$;
7: **end for**
8: Stack $E = \{concat(E_t^{\text{on}}, E_t^{\text{off}})\}_{t=1}^{T}$;
9: Apply temporal smoothing and re-binarization to obtain E'.

7. Supplementary Material

S.1. Computation of spectral and structural responses

In remote sensing scenarios, land-cover changes involve variations in both spectral reflectance and geometric structure or texture. Therefore, we construct a geographical response map $G(x, y)$ from two complementary perspectives to characterize the dynamic properties of land surfaces over time.

Spectral Index Difference $S(x, y)$. Used to measure the trend of spectral variation in land-cover materials. For imagery containing a red band R and a pseudo–near-infrared band $N = 0.5(R + G)$, it is defined as follows:

$$\text{NDVI} = \frac{N - R}{N + R + \epsilon}, \quad S(x, y) = \text{NDVI}_2(x, y) - \text{NDVI}_1(x, y), \tag{22}$$

Gradient Direction Similarity $C(x, y)$. This metric characterizes the consistency of structural and textural patterns. Given grayscale images I_1, I_2 he gradients are computed using Sobel kernels K_x, K_y as follows:

$$\nabla_x I = I * K_x, \quad \nabla_y I = I * K_y, \tag{23}$$

The directional similarity is defined as follows:

$$C(x, y) = \frac{\nabla_x I_1 \nabla_x I_2 + \nabla_y I_1 \nabla_y I_2}{\sqrt{(\nabla_x I_1^2 + \nabla_y I_1^2 + \epsilon)(\nabla_x I_2^2 + \nabla_y I_2^2 + \epsilon)}}. \tag{24}$$

Finally, the two components are fused using a hyperbolic tangent function:

$$G(x, y) = \tanh(\lambda_s S(x, y) + \lambda_g C(x, y)), \tag{25}$$

Where λ_s and λ_g denote the weights of the spectral and structural components, respectively, typically set as $\lambda_s = 0.6, \lambda_g = 0.4$.

S.2. Learnable Temporal correction Network f_θ, g_θ

To compensate for local reflectance and texture variations, GSI-P introduces two lightweight learnable temporal correction subnets,$f_\theta and g_\theta$, structured as follows:

$$[f_\theta, g_\theta] : \mathbb{R}^{2C\times H\times W} \to \mathbb{R}^{2\times H\times W}. \tag{26}$$

The implementation is formulated as follows:

```
nn.Sequential(
Conv2d(), ReLU(),
Conv2d(), ReLU(),
Conv2d()
)
```

Where f_θ corresponds to the local scaling branch, constrained to $[0, 1]$ by a Sigmoid activation, And g_θ represents the offset branch, constrained to $[-1, 1]$ through a Tanh activation. Together, they jointly refine local temporal responses, enabling the model to adapt to regional reflectance variations while preserving geographical consistency.

S.3. Definitions and Physical Interpretations of the Polarities

In the polarity modeling stage, GSI-P employs three types of logical masks:

- Temporal matching mask $\mathbf{1}\{t = \tau(x, y, c)\}$, indicating whether a pixel fires at time t;
- Polarity mask $\mathbf{1}\{\Delta I > 0\}$ or $\mathbf{1}\{\Delta I < 0\}$, distinguishing between brightening and darkening pixel;
- Existence mask $\mathbf{1}_{\text{pres}} = \mathbf{1}\{\tilde{M}'(x, y, c) \geq \eta\}$, activated only when the change intensity exceeds the threshold η.

The three masks are combined using a logical AND operation, triggering an event only when the temporal match, polarity consistency, and significant change conditions are simultaneously satisfied. This mechanism ensures that spike

activations correspond to genuine land-cover changes rather than noise or minor perturbations.

S.4. Motivation and Analysis of the Triangular Kernel Selection

Since each pixel fires only once within the sparse temporal sequence, the sequence becomes highly sparse along the time dimension. To enhance temporal continuity and noise robustness, we apply one dimensional triangular kernel smoothing:

$$E'_t = \sum_{\delta=-k}^{k} w_\delta E_{t+\delta}, \quad w_\delta \propto (k-|\delta|+1), \quad \sum_{\delta=-k}^{k} w_\delta = 1. \tag{27}$$

The kernel width k is adaptively adjusted according to the geographical response map G:

$$k = \max\Big(1, \lfloor k_0(1+0.5\langle|G|\rangle)\rfloor\Big), \tag{28}$$

Where $\langle|G|\rangle$ denotes the spatial mean of $|G|$, reflecting the overall intensity of global scene variations. In instances where the scene displays significant variation (for example, in the context of urban expansion areas), the smoothing kernel is expanded to facilitate the bridging of discontinuities between adjacent time steps. In instances where the scene remains stable (for example, in the case of water bodies or forest regions), the kernel narrows to prevent oversmoothing. Compared to mean or Gaussian kernels, the triangular kernel better aligns with the linear decay property of membrane potential integration, effectively smoothing noise while preserving the shape of the primary peak.

S.5. SNN Encoder

To extract temporal structural information from the pseudo-event sequences generated by the GSI-P module, we design a lightweight Spiking Neural Network Encoder (SNN Encoder). The encoder adopts a multi-layer Conv–BN–LIF (Convolution–Batch Normalization–Leaky Integrate-and-Fire) architecture, enabling efficient encoding of dynamic features while maintaining temporal precision. The SNN encoder takes as input a tensor $x \in \mathbb{R}^{B\times T\times C\times H\times W}$, where B is the batch size, T the temporal length, and C, H, and W denote the channel and spatial dimensions, respectively. Each layer consists of the following three stages: (1) Spatial Convolution, extracts local features and reduces spatial resolution; (2) Batch Normalization, stabilizes the membrane potential distribution; and (3) LIF Neuron, performs temporal integration and spike firing. This architecture progressively extracts spike features at multiple scales, yielding three hierarchical levels of temporal feature representations.

```
nn.Sequential(
Conv2d(),BN(), LIF(),
Conv2d(),BN(), LIF(),
)
```

Table 8. Performance comparison under different temporal step lengths T.

T	P(%)	R(%)	F1(%)	IoU(%)	memory
4	93.29	92.6	92.95	86.82	12865
8	93.45	92.71	93.08	87.05	15330
12	93.44	92.4	92.92	86.77	17786
16	93.04	92.79	92.91	86.77	20147
24	93.33	92.3	92.81	86.58	22576

S.6. Effect of the Number of Temporal Frames T

The temporal step length, denoted T, is a crucial parameter that determines the length of the sparse interpolated sequence generated by the GSI-P module between bi-temporal images. This parameter directly affects the temporal resolution and dynamic continuity perceived by the SNN encoder. When T is small, the temporal discretisation becomes coarse, causing spike events to cluster within a few steps and limiting the representation of intermediate transitions in gradual changes. Conversely, when T is too large, temporal continuity is enhanced, but the variation between pseudo-frames is reduced, leading to a significant decrease in spike density and a reduction in temporal saliency. Furthermore, excessively lengthy temporal sequences can introduce additional computational overhead and may result in gradient sparsity during temporal propagation.

Therefore, the temporal step length T plays a crucial role in balancing temporal resolution and sparsity within the STSpikeFuse framework. To evaluate its impact, we conducted systematic ablation experiments under different settings of $T \in 4, 8, 12, 16, 24$ and reported the trends in accuracy and computational efficiency, show in Table 8. The results demonstrate that performance peaks at $T = 8$, and that further increases in temporal resolution provide no significant gain. This confirms the effectiveness of medium-length pseudo-temporal sequences for dynamic modeling in the network.

S.7. Training curves of different delay mapping form

The training curves in Fig. 5 further confirm its stability, whereas the learnable variant shows clear fluctuations. Overall, the linear mapping offers the most stable and consistent temporal behaviour.

S.8. Analysis of Polarity modeling

In sparse temporal encoding, the directionality of brightness change is of paramount importance. It is important to

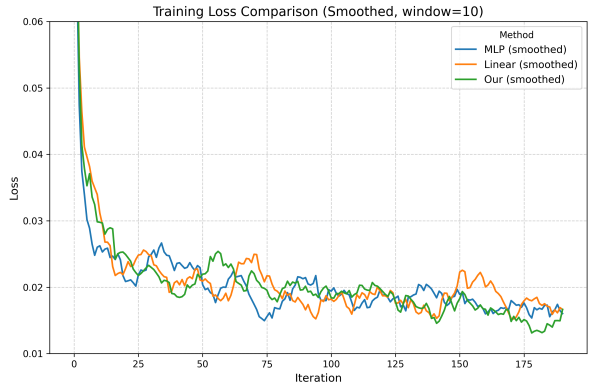

Figure 5. The convergence trends of training loss under different mapping forms.

note that disregarding the distinction between reflection enhancement (ON) and reflection attenuation (OFF) hinders the model's capacity to discern between brightening and darkening dynamics. This results in temporal signals becoming intertwined within the semantic space. To address this, we introduce polarity modelinging in the GSI-P module to separate ON and OFF channels while preserving the sign of changes:

$$E_t^{\text{off}}(x, y, c) = \mathbf{1}\{t = \tau(x, y, c)\}\, \mathbf{1}\{\Delta I < 0\}\, \mathbf{1}_{\text{pres}}, \quad (29)$$

Where $\mathbf{1}\{\Delta I < 0\}$ denotes the mask for the OFF channel, $\tau(x, y, c)$ is the firing delay, and $\mathbf{1}_{\text{pres}}$ indicates whether the pixel participates in encoding at the current time. The ON channel is computed symmetrically to record reflection-enhancement. We compare three configurations: (1) Ignore direction: encode only the radiometric discrepancy of brightness change; (2) Ignore saliency, encode direction without considering radiometric discrepancy weighting; and (3) Bipolar modeling (ours). explicitly separate ON and OFF channels while retaining radiometric discrepancy information. As shown in Table 9, single polarity models that ignore direction mix positive and negative changes into a single response during training, causing label ambiguity and gradient conflicts, which lead to unstable or even divergent optimization. It is evident that disregarding radiometric discrepancy saliency engenders substandard performance, concomitant with the emergence of convergence issues. In contrast, the explicitly separated bipolar modeling achieves stable convergence and improves F1 and IoU by 0.38% and 1.69%, respectively.

S.9. Analysis of the Time-to-First-Spike

The Time-to-First-Spike (TfS) reflects the temporal saliency of changes and guides the model to focus on early changing regions. However, the manner in which TfS is integrated exerts a direct influence on model stability and temporal alignment. Should TfS be disregarded, temporal information will degenerate into static statistics. Conversely, overly deep injection has the potential to distort the backbone feature distribution. Therefore, we evaluate four integration strategies: (1) No TfS, temporal cues are entirely removed; (2) Bias-only, TfS is used solely as an attention bias to modulate attention logits; (3) Q-only, TfS is injected into the query feature Q to explicitly modulate spatial saliency; and (4) Both bias and Q, TfS is simultaneously used as a bias term and query modulation.

Table 9. Effect of Polarity modeling on Performance.

modeling	P(%)	R(%)	F1(%)	IoU(%)
Ignore direction	Fail to converge			
Ignore saliency	93.53	92.34	92.93	86.79
Bipolar	93.45	92.71	93.08	98.07

As shown in Table 10, completely removing TfS causes the model fail to converge, indicating that temporal saliency is a prerequisite for stable SNN–ANN collaboration. Using only bias or query modulation partially improves performance but introduces overfitting or unstable attention behavior. When both are combined, TfS provides early temporal guidance at the attention level and enhances spatial alignment at the feature level, resulting in smoother temporal responses and more consistent detection.

Table 10. Ablation Results on TfS Utilization Strategies.

Strategies	P(%)	R(%)	F1(%)	IoU(%)
No TfS	Fail to converge			
Bias-only	93.21	92.57	92.89	86.72
Q-only	93.45	92.15	92.8	86.57
Both	93.45	92.71	93.08	98.07