Towards More Discriminative Feature Learning in SNNs With Temporal-Self-Erasing Supervision

Wei Liu[1], Li Yang[1], Mingxuan Zhao[1], Dengfeng Xue[1], Shuxun Wang[1], Boyu Cai[1,3], Jin Gao[1], Wenjuan Li[1], Bing Li[1], Weiming Hu[1,2,3] *

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]School of Information Science and Technology, ShanghaiTech University

Abstract

Spiking Neural Networks (SNNs) are biologically inspired models that process visual inputs over multiple time steps. However, they often struggle with limited feature discrimination along the temporal dimension due to inherent spatio-temporal invariance. This limitation arises from the redundant activation of certain regions and shared supervision for multiple time steps, constraining the network's ability to adapt and learn diverse features. To address this challenge, we propose a novel Temporal-Self-Erasing (TSE) supervision method that dynamically adapts the learning regions of interest for different time steps. The TSE method operates by identifying highly activated regions from predictions across multiple time steps and adaptively suppressing them during model training, thereby encouraging the network to focus on less activated yet potentially informative regions. This approach not only enhances the feature discrimination capability of SNNs but also facilitates more effective multi-time-step inference by exploiting more semantic information. Experimental results on benchmark datasets demonstrate that our TSE method significantly improves the classification accuracy and robustness of SNNs.

Introduction

Spiking Neural Networks (SNNs) have garnered considerable attention as a biologically plausible alternative to traditional artificial neural networks (ANNs), offering a unique temporal dimension for processing information (Roy, Jaiswal, and Panda 2019; Schuman et al. 2022; Li et al. 2023). Unlike conventional ANNs that rely on continuous and static representations, SNNs, inspired by the dynamics of biological neurons (Maass 1997), utilize spikes to encode and transmit information across multiple time steps. This temporal processing ability allows SNNs to mimic the sequential and dynamic nature of human perception and cognition. However, while SNNs are inherently designed for dynamic reasoning, they often face challenges related to limited feature discrimination along the temporal dimension. This limitation arises primarily from the conventional training approach used for SNNs, which produces identical back-propagation gradients and lacks diversity in feature representations across multiple time steps.

*Corresponding author: wmhu@nlpr.ia.ac.cn

In traditional SNN training paradigms, an input image is encoded and fed into the network across several time steps, generating a set of feature maps. These feature maps are then processed through the global average pooling and fully connected layers, ultimately resulting in a classification prediction for each time step. The predictions from all time steps are averaged to produce a final output, which is then supervised by a shared loss function. While this approach ensures temporal consistency, it inadvertently causes similar inference processes across time steps. Consequently, the SNN tends to focus on the same discriminative regions of the image repeatedly, failing to leverage the full potential of its temporal dynamics to explore other informative regions. This redundancy in feature representation limits the network's ability to capture a diverse set of features, thereby constraining its overall discriminative power.

To address this limitation, we propose a novel approach named Temporal-Self-Erasing (TSE) supervision, which aims to enhance the feature discriminability of SNNs by encouraging the network to explore different semantic regions across time steps. The core idea behind this approach is to dynamically modulate the feature maps at each time step, effectively "erasing" the regions that have already been well activated in previous time steps. By doing so, the network is forced to shift its focus to other, potentially less obvious but still informative, regions of the input image. This process not only enriches the diversity of the features learned by the network but also improves its ability to recognize objects based on a broader range of discriminative cues.

Our method begins by constructing erasing masks for each time step, which are used to modulate the feature maps before they are passed through the classification layers. Specifically, after the feature maps are generated at each time step, they are passed through a fully connected layer to produce classification prediction maps. For any given time step, we average the classification prediction maps from all previous time steps and apply a Softmax function along the category dimension. This results in a probability score map that highlights the regions of the image that have been most strongly associated with the true class label in earlier time steps.

To ensure that the network does not repeatedly focus on these regions, we construct an erasing mask by applying a threshold function to the probability score map. This thresh-

old is designed to suppress the regions with high predicted probability scores, thereby encouraging the network to focus on other, less activated regions. The threshold itself is a combination of a fixed component and a dynamic component, the latter of which is adaptively adjusted based on the distribution of probability scores. The resulting erasing mask is then applied to the feature map, effectively modulating the network's attention at the current time step and promoting the discovery of new discriminative regions.

To summarize, our contributions are threefold:

- We analyze the conventional supervision method of SNNs, highlighting issues of identical backpropagation gradients and similar feature representations across different time steps.
- We propose a novel Temporal-Self-Erasing supervision method that dynamically suppresses redundant regions across time steps, thereby enhancing the temporal feature discrimination ability of SNNs.
- Extensive experiments on multiple benchmarks demonstrate the efficacy of our method. Our TSE supervision brings significant improvements over existing training techniques for SNNs.

Related Work

The enhancement of feature representation in Spiking Neural Networks (SNNs) is crucial for improving their performance in various tasks. This enhancement can be achieved through several approaches, including the conversion from Artificial Neural Networks (ANNs) to SNNs, network structure optimization, and the incorporation of attention mechanisms.

Conversion From ANN to SNN

A promising approach to enhancing feature representation in SNNs is to convert pre-trained Artificial Neural Networks (ANNs) into SNNs (Cao, Chen, and Khosla 2015; Hunsberger and Eliasmith 2015; Rueckauer et al. 2017; Bu et al. 2023; Meng et al. 2022). This method leverages the robust feature extraction capabilities of ANNs, which are typically trained on large datasets using advanced optimization algorithms. The conversion process involves techniques such as weight normalization and threshold balancing, which translate the continuous activations of ANNs into spike-based representations compatible with SNNs. This allows SNNs to inherit the discriminative power of ANNs, thereby improving feature representations and making them more informative and capable of distinguishing complex patterns (Cao, Chen, and Khosla 2015; Deng et al. 2020). However, the extended time steps required during this conversion process result in increased computational overhead. Therefore, we adopted the direct training of SNNs to effectively reduce computational overhead, better align with the inherent characteristics of SNNs, and minimize delay and energy consumption during the inference stage.

Attention Mechanism

The attention mechanism has achieved great success in various vision tasks of ANNs, such as object detection(Dai et al. 2021; Yin et al. 2024; Yang et al. 2022a), object tracking(Gopal and Amer 2024), and visual grounding(Yang et al. 2022b, 2024). Integrating attention mechanisms into Spiking Neural Networks (SNNs) is a promising strategy for enhancing feature representation. Attention mechanisms enable the network to dynamically focus on the most relevant aspects of the input data, thereby improving the efficiency and accuracy of feature extraction. Within the context of SNNs, attention is employed as an auxiliary module to enhance the network's representational capacity (Hu, Shen, and Sun 2018; Woo et al. 2018; Yang et al. 2021; Li et al. 2022; Guo et al. 2022a), such as by adding attention modules across temporal, spatial, and channel dimensions to optimize membrane potential distribution (Yao et al. 2023; Guo, Huang, and Ma 2023). This approach enhances both the feature representation and the overall accuracy of the network. However, while multi-dimensional attention can reduce some unnecessary computations during specific time steps, it also introduces significant additional computational overhead. In this work, inspired by the principles of attention learning in SNNs, we employ the Temporal-Self-Erasing mechanism to expand the scope of SNNs' attention regions by suppressing feature learning in redundant areas, thereby optimizing the network's focus and improving its learning efficiency.

Erasing Strategy

"Erasing" is a well-established strategy in segmentation and classification tasks within Artificial Neural Networks (ANNs) (Yi and Wu 2019; Sun et al. 2020). This approach often involves masking specific regions of an image to help the model uncover additional semantic information, thereby enhancing its feature representation capabilities. For instance, in adversarial erasing networks (Wei et al. 2017), a Class Activation Mapping (CAM) (Zhou et al. 2016) branch identifies and erases discriminative regions within an image based on a set threshold. The modified image is then fed into another network to discover previously overlooked regions. Subsequent research has advanced this erasing strategy into an end-to-end training framework (Li et al. 2018; Zhang et al. 2018). The self-erasing network (Hou et al. 2018), for example, incorporates background priors to focus the network's attention on target areas, preventing the spread of attention to irrelevant regions. Drawing inspiration from adversarial and self-erasing techniques, we introduce a novel approach that decouples self-erasing SNNs along the temporal dimension. In this method, attention areas are progressively erased across different time steps, allowing the network to focus on distinct regional features at each time step. The resulting erased features are then leveraged for image classification, providing a more refined and temporally dynamic approach to feature representation.

Preliminary

The spiking neuron serves as the core element of Spiking Neural Networks (SNNs). It integrates incoming spikes, adjusts its membrane potential accordingly, and produces spike outputs over time. In this study, we utilize the Leaky

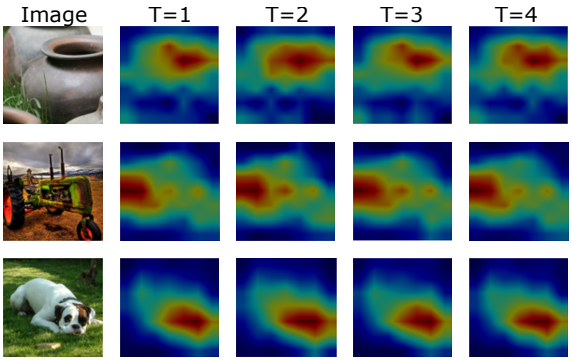

Figure 1: Class activation maps across different time steps in an SNN. The consistent distribution of high activation regions suggests that similar feature representations are shared across these time steps.

Integrate-and-Fire (LIF) neuron model, a well-established framework that governs the evolution of the neuronal membrane potential. The model is defined by the following equation:

$$u_i(t) = \lambda u_i'(t-1) + \sum_j w_{i,j}\, s_j(t). \tag{1}$$

In this equation, $u_i(t)$ denotes the membrane potential of neuron i at time step t, while $u_i'(t-1)$ represents the membrane potential after a spike has occurred at time step $t-1$, scaled by a constant leakage factor λ. The term $s_j(t)$ signifies the input spike from a presynaptic neuron j to neuron i, with $w_{i,j}$ representing the synaptic weight connecting these neurons. At each time step t, the neuron's membrane potential $u_i(t)$ is evaluated using a Heaviside step function $H(\cdot)$ to determine spike generation. When the membrane potential $u_i(t)$ exceeds the threshold v_{th}, the neuron emits a spike, resulting in an output of 1. If the threshold is not reached, no spike is generated, and the output remains 0. The spike output $s_i(t)$ is mathematically defined as follows:

$$s_i(t) = H(u_i(t) - v_{th}) = \begin{cases} 1, & \text{if } u_i(t) \geq v_{th} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

When a spike is fired, the membrane potential undergoes a reset. We apply a hard reset mechanism as proposed by (Ledinauskas et al. 2020), which resets the membrane potential to 0 upon spike generation, while leaving it unchanged if no spike occurs. This reset mechanism is described by:

$$u_i'(t) = u_i(t)(1 - s_i(t)). \tag{3}$$

This approach ensures proper management of the neuron's membrane potential, facilitating efficient temporal information processing within the SNN.

Method

In this section, we start by examining the traditional supervision method of SNNs and analyze its limitations in feature learning over multiple time steps. We then elaborate on the proposed temporal-self-erasing supervision method, aiming to enhance discriminavie feature learning in SNNs.

Analysis

For time step t in an SNN, we denote the feature map before the global average pooling (GAP) layer as $\mathbf{F}^t \in \mathbb{R}^{H\times W\times C}$. This feature map is processed through a GAP layer and a fully connected (FC) layer with weights $\mathbf{W}$ to produce the classification prediction:

$$\mathbf{p}^t = \mathrm{FC}(\mathrm{GAP}(\mathbf{F}^t)). \tag{4}$$

The predictions from all T time steps are averaged and normalized to produce the final classification probabilities: $\hat{\mathbf{p}} = \mathrm{Softmax}(\frac{1}{T}\sum_{t=1}^{T}\mathbf{p}^t)$. The cross-entropy loss is usually applied to $\hat{\mathbf{p}}$ to optimize the SNN:

$$\mathcal{L} = \mathcal{L}_{CE}(\hat{\mathbf{p}}, y) = -\log(\hat{\mathbf{p}}_y). \tag{5}$$

The backpropagation gradients for the feature map $\mathbf{F}^t$ at location (i,j) can be derived using the chain rule as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{F}_{i,j}^t} &= \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{p}}} \cdot \frac{\partial \hat{\mathbf{p}}}{\partial \mathbf{p}^t} \cdot \frac{\partial \mathbf{p}^t}{\partial \mathbf{F}_{i,j}^t} \\ &= \frac{1}{T} \cdot \frac{1}{HW} \cdot (\hat{\mathbf{p}} - \mathbf{e}_y) \cdot \mathbf{W}. \end{aligned} \tag{6}$$

Where $\mathbf{e}_y$ is the unit vector corresponding to the true class label y, and H and W denote the height and width of the feature map $\mathbf{F}^t$, respectively. We can find that the backpropagation gradients are always the same for feature maps at any time step t under this supervision, which does not favor SNNs learning richer and diverse features across different time steps. Figure 1 shows the class activation maps of different time steps for an SNN trained with the loss in Equation (5). Their high activation regions are in similar distributions, indicating that different time steps share similar feature representations. To achieve more accurate recognition, multi-time-step network paths of SNNs should fully exploit various semantic information for joint reasoning. This is challenging with the current supervision, prompting us to design new methods.

Temporal-Self-Erasing Supervision

Motivated by the above analysis, we argue that the supervision approach should be different for each time step, otherwise it may limit the SNNs from learning richer feature semantics. When observing an object, we humans can accurately recognize it in a dynamic process by seeking out the most distinct areas and then shifting attention to other semantic object parts. While SNNs, inspired by human brains, are designed to reason over multiple time steps, they lack such ability to uncover various discriminative regions of an object. To enhance feature learning from more discriminative regions over multiple time steps, we propose temporal-self-erasing supervision, which encourages SNNs to exploit more semantic regions during training (as shown in Figure 2).

Specifically, we first take the feature maps $\{\mathbf{F}^t\}_{t=1}^T$ before the global average pooling layer and apply the fully connected layer to the features at each location (i,j): $\mathbf{P}_{i,j}^t =$

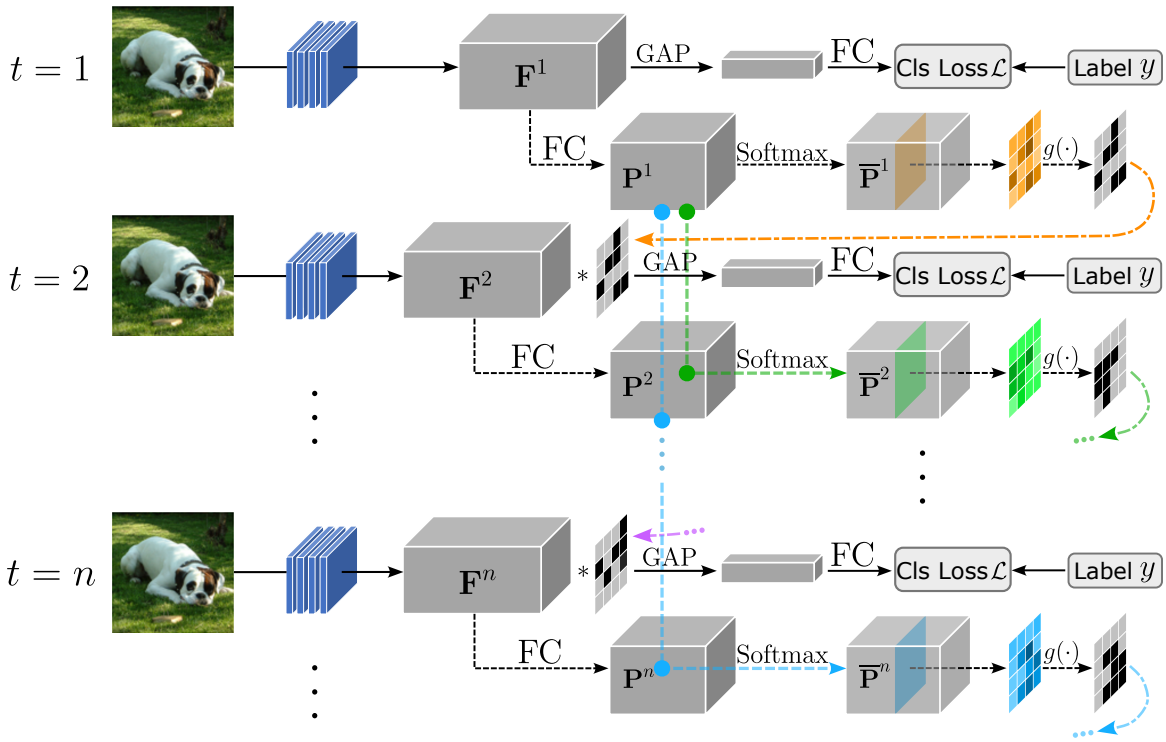

Figure 2: Temporal-Self-Erasing (TSE) supervision framework. The TSE encourages SNNs to explore diverse semantic regions over multiple time steps for enhanced feature discriminability. At each time step, an erasing mask is generated based on the predictions from previous time steps, suppressing regions with high predicted probability scores and redirecting attention to less activated areas. This process iteratively refines the feature maps, resulting in more robust and varied feature representations across time steps, ultimately improving classification accuracy.

$\mathrm{FC}(\mathbf{F}^t_{i,j})$. In this way, we obtain the classification prediction maps $\{\mathbf{P}^t\}_{t=1}^T$ for different time steps. For the current time step t ($t > 1$), we average the classification prediction maps from previous time steps $0 \sim (t-1)$ and apply the Softmax function along the category dimension:

$$\overline{\mathbf{P}}^{t-1} = \mathrm{Softmax}\left(\frac{1}{t-1}\sum_{k=1}^{t-1}(\mathbf{P}^k)\right). \tag{7}$$

The produced $\overline{\mathbf{P}}^{t-1}$ contains the probability scores for different categories at each location based on predictions before time step t. As the locations presenting high probability scores for the true class label y are already considered semantically discriminative in previous time steps, we seek to have the current time step focus on other regions to exploit more potentially informative regions. To this end, we take the predicted probability score map $\overline{\mathbf{P}}^{t-1}_y$ corresponding to the true label y to build the erasing mask for time step t by the function $g(\cdot)$:

$$\mathbf{M}^t = g\left(\overline{\mathbf{P}}^{t-1}\right). \tag{8}$$

This $g(\cdot)$ function basically applies a threshold to $\overline{\mathbf{P}}^{t-1}_y$ to suppress regions associated with higher predicted probability scores. While a fixed threshold can be used for selection, it is less flexible when predicted probability scores increase during training. Thus, we employ a combination of fixed and dynamic thresholds to generate the binary erasing mask. We represent the fixed threshold by a hyper-parameter τ_f. The dynamic threshold τ_d is calculated as the mean of the probability score map $\overline{\mathbf{P}}^{t-1}_{i,j}$ plus a hyper-parameter κ times the stand deviation of the scores:

$$\tau_d = \mathrm{mean}(\overline{\mathbf{P}}^{t-1}_{i,j}) + \kappa \cdot \mathrm{std}(\overline{\mathbf{P}}^{t-1}_{i,j}). \tag{9}$$

This dynamic threshold can be adaptively adjusted based on the distribution of probability scores. Thus, we determine the mask value at each location (i,j) of $\mathbf{M}^t$ by:

$$\mathbf{M}^t_{i,j} = \begin{cases} 0, & \text{if } \overline{\mathbf{P}}^{t-1}_{i,j} \geq \max(\tau_f, \tau_d) \\ 1, & \text{otherwise.} \end{cases} \tag{10}$$

After obtaining the erasing mask $\mathbf{M}^t$, we use it to modulate the feature map $\mathbf{F}^t$ at time step t, focusing on the regions less activated by previous time steps. The modulated feature maps is then fed through the global average pooling and fully connected layers to generate a classification prediction $\tilde{\mathbf{p}}^t$:

$$\tilde{\mathbf{p}}^t = \mathrm{FC}(\mathrm{GAP}(\mathbf{M}^t \cdot \mathbf{F}^t)). \tag{11}$$

This $\tilde{\mathbf{p}}^t$ essentially ignores predictions from erased regions' features, focusing on other potentially semantic areas.

Following the above approach, for each time step except the first one, we use classification predictions from previous time steps to build an erasing mask and apply it to the current feature map, producing a modulated classification prediction $\tilde{\mathbf{p}}^t$ $(t > 1)$. This allows the classification prediction at each time step t $(t > 1)$ to progressively focus on regions that are not well attended by previous time steps. Finally, we apply the cross-entropy loss on the classification prediction at the first time step and the modulated classification predictions at other time steps, respectively:

$$\mathcal{L} = \mathcal{L}_{CE}(\mathbf{p}^1, y) + \sum_{t=2}^{T} \mathcal{L}_{CE}(\tilde{\mathbf{p}}^t, y). \tag{12}$$

In this manner, our TSE method utilize the SNN's own predictions to modulate themselves at different time steps and apply supervision separately, helping to uncover and learn more discriminative features. The self-erasing process is only used to produce modulated predictions during model training, without introducing extra modules or computational cost. Experimental results demonstrate the efficacy of our method across various benchmarks.

Experiments

Datesets

In this section, we conduct extensive experiments to validate the efficacy of our proposed method. We employ Spiking ResNet-18/19 as our backbone and experiment on datasets such as CIFAR-100 (Krizhevsky, Nair, and Hinton 2010), ImageNet (Deng et al. 2009), and the neuromorphic DVS-CIFAR10 (Li et al. 2017). CIFAR-100 has 100 classes containing 600 images for each category. 500 images for training and 100 images for testing. ImageNet has over 1.2 million training images and 50 thousand valiting images. DVS-CIFAR10 is an event-based, dynamic vision sensor version of the CIFAR-10 dataset (Krizhevsky, Nair, and Hinton 2010). It captures pixel-level changes in brightness at high temporal resolution, resulting in a stream of events instead of conventional frames.

Training Setup

To be consistent with other research in direct SNN training (Che et al. 2022; Duan et al. 2022), we utilize LIF neurons for the direct training of our model. The threshold voltage and the membrane potential decay constant are set to 1 and 2, respectively. Our SNN model is trained using the Stochastic Gradient Descent (SGD) optimizer, with a momentum of 0.9, and an initial learning rate of 0.1, decreasing to 0 with cosine learning rate scheduler. The batch size and epoch numbers are set to 32 and 320, respectively. For the CIFAR-100 and DVS-CIFAR10 datasets, we conduct both training and inference processes utilizing a single NVIDIA A100 GPU. For the ImageNet dataset, we employ four NVIDIA A100 GPUs for training.

Comparisons With State-of-the-Art Methods

CIFAR-100 For CIFAR-100, as shown in Table 1, our method achieves an accuracy of 81.47% using a spiking ResNet-19 trained over 4 time steps. Moreover, it is 6.80% higher than that of SLTT, which also operates over 4 time steps. It is important to note that the recent SlipReLU, despite attaining a competitive accuracy of 78.55% using 128 time steps, is a conversion-based method that requires a considerable number of time steps to minimize information loss.

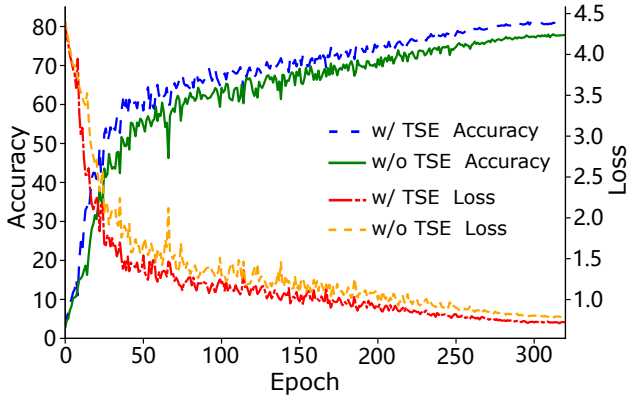

Figure 3: Comparison of accuracy and loss curves during training w/ and w/o TSE.

ImageNet For ImageNet, experiments are conducted utilizing ResNet-18, ResNet-34 and ResNet-50 as backbones, with the results detailed in Table 1. The proposed method is evaluated against existing direct SNN training methods. With the same network backbone (ResNet-34) and time step, our model is 0.50% higher than PSN (Fang et al. 2024). Using the same backbone architecture, our method outperforms MPBN (Guo et al. 2023b) on ResNet-34 with an accuracy of 71.04% versus 64.71%.

DVS-CIFAR10 We further evaluate the performance of our method using the DVS-CIFAR10 dataset. The comparative results, as delineated in Table 2, benchmark our approach against recent methods. Specifically, when implemented with a ResNet-19 architecture, our method utilizing 10 time steps attains an accuracy level of 83.60%. This performance is contrasted with that of the LSG and MPBN methods, which achieves accuracies of 77.90% and 74.40% respectively, under identical time step conditions.

Ablation Study

TSE Effectiveness To validate the effectiveness of the Temporal-Self-Erasing (TSE) mechanism proposed in the method section, we conducted ablation studies using ResNet-19 as the baseline model on the CIFAR-100 and DVS-CIFAR10 datasets. The experimental results are summarized in Table 3, evaluating the impact of TSE under time steps of $T = 4$ and $T = 10$.

On the CIFAR-100 dataset, the application of TSE led to a 3.46 percentage point improvement in accuracy compared to the baseline model without TSE. Additionally, TSE accelerated the network's convergence. Figure 3 compares the accuracy and loss curves during training between the baseline model and the model with TSE, demonstrating that TSE facilitates faster convergence. The TET method, optimized

Dataset	Method	Architecture	Time step	Train Method	Accuracy (%)
CIFAR-100	RTS (Deng and Gu 2021)	ResNet-18	64	ANN-SNN conversion	69.27
	QCFS (Bu et al. 2023)	ResNet-18	4	ANN-SNN conversion	75.67
	SlipReLU (Jiang et al. 2023)	ResNet-18	128	ANN-SNN conversion	78.55
	STBP-tdBN (Zheng et al. 2021)	ResNet-19	6	Surrogate Gradient	71.12
	Sew ResNet (Fang et al. 2021a)	ResNet-34	4	Surrogate Gradient	67.04
	TET (Deng et al. 2022)	ResNet-19	4	Surrogate Gradient	74.47
	SLTT (Meng et al. 2023)	ResNet-18	6	Surrogate Gradient	74.67
	TTS (Guo et al. 2023a)	ResNet-19	2	Surrogate Gradient	80.20
	GAC-SNN (Qiu et al. 2024)	ResNet-18	4	Surrogate Gradient	79.83
	Ours	**ResNet-19**	**4**	**Surrogate Gradient**	**81.47**
ImageNet	Sew ResNet (Fang et al. 2021a)	ResNet-18	4	Surrogate Gradient	63.18
		ResNet-34	4	Surrogate Gradient	67.04
	Real Spike (Guo et al. 2022c)	ResNet-18	4	Surrogate Gradient	63.68
	MPBN (Guo et al. 2023b)	ResNet-18	4	Surrogate Gradient	63.14
		ResNet-34	4	Surrogate Gradient	64.71
	PSN (Fang et al. 2024)	ResNet-18	4	Surrogate Gradient	67.63
		ResNet-34	4	Surrogate Gradient	70.54
	Ours	**ResNet-18**	**4**	**Surrogate Gradient**	**67.74**
		ResNet-34	**4**	**Surrogate Gradient**	**71.04**
		ResNet-50	**4**	**Surrogate Gradient**	**73.40**

Table 1: Comparison results with different methods on CIFAR100 and ImageNet.

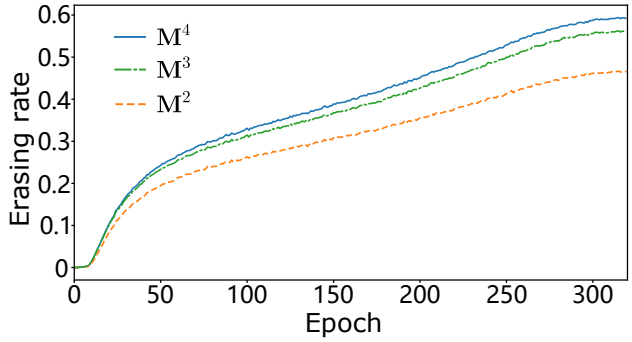

Figure 4: Feature map erasing rate curve. The curve illustrates the changes in the erasing rate of feature maps by the TSE mechanism at time steps $T = 2$, 3, and 4 during training on the CIFAR-100 dataset with a training time step of $T = 4$.

through temporal decoupling supervision, improved the accuracy on CIFAR-100 by 2.03 percentage points over the baseline. Furthermore, compared to the TET method, which also enhances SNNs through decoupled supervision, TSE provided an additional 1.43 percentage point improvement. Figure 4 illustrates the erasing rate of feature maps at different time steps under TSE (i.e., the proportion of pixels in the feature map set to zero by the TSE mechanism). The results indicate that the erasing rate increases with the time steps, suggesting that the network utilizes the self-erasing mechanism to uncover more semantic information.

On the DVS-CIFAR10 dataset, the impact of TSE was even more pronounced. Under the condition of $T = 10$, applying TSE resulted in a 4.4 percentage point increase in accuracy compared to the baseline model without TSE. However, the TET method may have suffered from overfitting on the DVS-CIFAR10 dataset, leading to a 1.6 percentage point decrease in accuracy compared to the baseline model.

These significant improvements indicate that TSE offers considerable advantages when applied to neuromorphic datasets and networks with extended time steps.

Analysis of Erasing Mask Construction Methods In Table 4, we present the performance of three different erasing mask construction methods on the CIFAR-100 dataset: Fixed Threshold (τ_f), Dynamic Threshold (τ_d), and Hybrid Threshold Method ($\max(\tau_f, \tau_d)$). The experimental results show that the fixed threshold method, which erases redundant activation regions based on a preset threshold, improves the accuracy of the baseline model (ResNet-19) by 2.79 percentage points. However, since it is challenging for the network to accurately identify discriminative regions during the early stages of training, dynamic threshold method applies a dynamic threshold calculated based on the network's prediction scores. This approach may lead to the erasure of critical information in the initial training phase, resulting in accuracy of dynamic threshold method being 0.82 percentage points lower than the baseline. The hybrid thresholding method, which combines fixed threshold with dynamic threshold, effectively guides the network by preserving essential region information while erasing redundant activation areas. This expands the learning scope of the feature regions. As a result, hybrid thresholding method outperforms fixed threshold and dynamic threshold, with accuracy improvements of 0.67 percentage points and 4.28 percentage points, respectively.

The Hyper-Parameters We conducted an ablation study on the two hyperparameters, the fixed threshold τ_f and the κ, within the erasing mask construction, to identify the op-

Dataset	Method	Architecture	Time step	Train Method	Accuracy (%)
DVS-CIFAR10	Dspike (Li et al. 2021)	ResNet-18	10	Surrogate Gradient	75.40
	SLTT (Meng et al. 2023)	VGG-11	10	Surrogate Gradient	77.30
	STBP-tdBN (Zheng et al. 2021)	ResNet-19	10	Surrogate Gradient	67.80
	RecDis-SNN (Guo et al. 2022b)	ResNet-19	10	Surrogate Gradient	72.42
	Real Spike (Guo et al. 2022c)	ResNet-19	10	Surrogate Gradient	72.85
	PLIF (Fang et al. 2021b)	7-layer CNN	20	Surrogate Gradient	74.80
	MPBN (Guo et al. 2023b)	ResNet-19	10	Surrogate Gradient	74.40
	MLF (Feng et al. 2022)	ResNet-19	10	Surrogate Gradient	70.36
	IM-Loss (Guo et al. 2022b)	ResNet-19	10	Surrogate Gradient	72.60
	LSG (Lian et al. 2023)	ResNet-19	10	Surrogate Gradient	77.90
	LM-H (Hao et al. 2023)	ResNet-19	10	Surrogate Gradient	79.10
	Ours	**ResNet-19**	**10**	**Surrogate Gradient**	**83.60**

Table 2: Comparison results with different methods on DVS-CIFAR10.

Dataset	Model	T	Accuracy(%)
CIFAR-100	baseline	4	78.01
	TET (Deng et al. 2022)	4	80.04
	TSE	4	**81.47**
DVS-CIFAR10	baseline	10	79.20
	TET (Deng et al. 2022)	10	77.60
	TSE	10	**83.60**

Table 3: Comparison of TSE with other supervision methods on the CIFAR-100 and DVS-CIFAR10 datasets with time steps $T = 4$ and $T = 10$, respectively.

Model	Accuracy(%)
baseline	78.01
Fixed Threshold	80.80
Dynamic Threshold	77.19
Hybrid Threshold	**81.47**

Table 4: Performance comparison of erasing mask construction methods on CIFAR-100 with time steps $T = 4$.

timal combination that enhances feature learning across different time steps while optimizing training efficiency. Table 5 presents the accuracy of the model under various τ_f and κ configurations. The experimental results indicate that the configuration of $\tau_f = 0.6$ and $\kappa = 0$ effectively suppresses redundant activation regions, thereby guiding the model to discover more semantic information. Based on this evidence, we determined that $\tau_f = 0.6$ and $\kappa = 0$ represent the optimal settings for the Hybrid Threshold method, ensuring more efficient model training. Additionally, using ResNet-19 as the backbone and setting the time step $T = 4$, we visualized the activation maps after masking in the last three time steps within the TSE method. We resize these masks to match the input image sizes and overlay them semi-transparently on the input images, as shown in Figure 5. The results demonstrate that as the time steps progress, TSE gradually expands the suppressed regions within the image, thereby facilitating the model in extracting more discriminative features and improving recognition accuracy.

κ	τ_f	Accuracy(%)
0.1	0.5	80.43
	0.6	80.72
	0.7	81.01
0	0.5	80.84
	0.6	**81.47**
	0.7	81.33
-0.1	0.5	80.83
	0.6	80.91
	0.7	80.95

Table 5: Performance comparison of mask construction methods on the CIFAR-100 dataset with time steps $T = 4$.

Figure 5: Visualization of the erased regions. The first column shows the input image, while the second to fourth columns depict the regions erased by the TSE mechanism at time steps $T = 2$, 3, and 4, respectively.

Conclusion

In conclusion, this paper presents a novel Temporal-Self-Erasing (TSE) supervision method that effectively addresses the limitations of SNNs in feature discrimination across the temporal dimension. By dynamically suppressing redundant activations and promoting focus on less activated regions, TSE significantly enhances the discriminative capability and robustness of SNNs. Experimental results on benchmark datasets, including CIFAR-100, ImageNet, and DVS-CIFAR10, demonstrate that TSE outperforms state-of-the-art methods, paving the way for more efficient and adaptive spatio-temporal processing in neuromorphic computing systems.

Acknowledgments

This work was supported by the National Science and Technology Major Project (2020AAA0105802, 2020AAA0105801), the National Natural Science Foundation of China (No. 62202469, 62403462, 62036011, 62192782, U2441241), Beijing Natural Science Foundation (L223003), the Project of Beijing Science and Technology Committee (No. Z231100005923046). The Key Research and Development Program of Xinjiang Uyghur Autonomous Region No. 2023B03024.

References

Bu, T.; Fang, W.; Ding, J.; Dai, P.; Yu, Z.; and Huang, T. 2023. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. *arXiv preprint arXiv:2303.04347*.

Cao, Y.; Chen, Y.; and Khosla, D. 2015. Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 113: 54–66.

Che, K.; Leng, L.; Zhang, K.; Zhang, J.; Meng, Q.; Cheng, J.; Guo, Q.; and Liao, J. 2022. Differentiable hierarchical and surrogate gradient search for spiking neural networks. *Advances in Neural Information Processing Systems*, 35: 24975–24990.

Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; and Zhang, L. 2021. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2988–2997.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, L.; Wu, Y.; Hu, X.; Liang, L.; Ding, Y.; Li, G.; Zhao, G.; Li, P.; and Xie, Y. 2020. Rethinking the performance comparison between SNNS and ANNS. *Neural networks*, 121: 294–307.

Deng, S.; and Gu, S. 2021. Optimal conversion of conventional artificial neural networks to spiking neural networks. *arXiv preprint arXiv:2103.00476*.

Deng, S.; Li, Y.; Zhang, S.; and Gu, S. 2022. Temporal efficient training of spiking neural network via gradient reweighting. *arXiv preprint arXiv:2202.11946*.

Duan, C.; Ding, J.; Chen, S.; Yu, Z.; and Huang, T. 2022. Temporal effective batch normalization in spiking neural networks. *Advances in Neural Information Processing Systems*, 35: 34377–34390.

Fang, W.; Yu, Z.; Chen, Y.; Huang, T.; Masquelier, T.; and Tian, Y. 2021a. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34: 21056–21069.

Fang, W.; Yu, Z.; Chen, Y.; Masquelier, T.; Huang, T.; and Tian, Y. 2021b. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2661–2671.

Fang, W.; Yu, Z.; Zhou, Z.; Chen, D.; Chen, Y.; Ma, Z.; Masquelier, T.; and Tian, Y. 2024. Parallel spiking neurons with high efficiency and ability to learn long-term dependencies. *Advances in Neural Information Processing Systems*, 36.

Feng, L.; Liu, Q.; Tang, H.; Ma, D.; and Pan, G. 2022. Multi-level firing with spiking ds-resnet: Enabling better and deeper directly-trained spiking neural networks. *arXiv preprint arXiv:2210.06386*.

Gopal, G. Y.; and Amer, M. A. 2024. Separable self and mixed attention transformers for efficient object tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6708–6717.

Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R. R.; Cheng, M.-M.; and Hu, S.-M. 2022a. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3): 331–368.

Guo, Y.; Chen, Y.; Liu, X.; Peng, W.; Zhang, Y.; Huang, X.; and Ma, Z. 2023a. Ternary Spike: Learning Ternary Spikes for Spiking Neural Networks. arXiv:2312.06372.

Guo, Y.; Huang, X.; and Ma, Z. 2023. Direct learning-based deep spiking neural networks: a review. *Frontiers in Neuroscience*, 17: 1209795.

Guo, Y.; Tong, X.; Chen, Y.; Zhang, L.; Liu, X.; Ma, Z.; and Huang, X. 2022b. Recdis-snn: Rectifying membrane potential distribution for directly training spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 326–335.

Guo, Y.; Zhang, L.; Chen, Y.; Tong, X.; Liu, X.; Wang, Y.; Huang, X.; and Ma, Z. 2022c. Real spike: Learning real-valued spikes for spiking neural networks. In *European Conference on Computer Vision*, 52–68. Springer.

Guo, Y.; Zhang, Y.; Chen, Y.; Peng, W.; Liu, X.; Zhang, L.; Huang, X.; and Ma, Z. 2023b. Membrane potential batch normalization for spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19420–19430.

Hao, Z.; Shi, X.; Huang, Z.; Bu, T.; Yu, Z.; and Huang, T. 2023. A progressive training framework for spiking neural networks with learnable multi-hierarchical model. In *The Twelfth International Conference on Learning Representations*.

Hou, Q.; Jiang, P.; Wei, Y.; and Cheng, M.-M. 2018. Self-erasing network for integral object attention. *Advances in neural information processing systems*, 31.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hunsberger, E.; and Eliasmith, C. 2015. Spiking deep networks with LIF neurons. *arXiv preprint arXiv:1510.08829*.

Jiang, H.; Anumasa, S.; De Masi, G.; Xiong, H.; and Gu, B. 2023. A Unified Optimization Framework of ANN-SNN Conversion: Towards Optimal Mapping from Activation Values to Firing Rates.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2010. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4): 1.

Ledinauskas, E.; Ruseckas, J.; Juršėnas, A.; and Buračas, G. 2020. Training deep spiking neural networks. *arXiv preprint arXiv:2006.04436*.

Li, G.; Deng, L.; Tang, H.; Pan, G.; Tian, Y.; Roy, K.; and Maass, W. 2023. Brain inspired computing: A systematic survey and future trends.

Li, G.; Fang, Q.; Zha, L.; Gao, X.; and Zheng, N. 2022. HAM: Hybrid attention module in deep convolutional neural networks for image classification. *Pattern Recognition*, 129: 108785.

Li, H.; Liu, H.; Ji, X.; Li, G.; and Shi, L. 2017. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11: 309.

Li, K.; Wu, Z.; Peng, K.-C.; Ernst, J.; and Fu, Y. 2018. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9215–9223.

Li, Y.; Guo, Y.; Zhang, S.; Deng, S.; Hai, Y.; and Gu, S. 2021. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. *Advances in Neural Information Processing Systems*, 34: 23426–23439.

Lian, S.; Shen, J.; Liu, Q.; Wang, Z.; Yan, R.; and Tang, H. 2023. Learnable Surrogate Gradient for Direct Training Spiking Neural Networks. In *IJCAI*, 3002–3010.

Maass, W. 1997. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9): 1659–1671.

Meng, Q.; Xiao, M.; Yan, S.; Wang, Y.; Lin, Z.; and Luo, Z.-Q. 2022. Training high-performance low-latency spiking neural networks by differentiation on spike representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12444–12453.

Meng, Q.; Xiao, M.; Yan, S.; Wang, Y.; Lin, Z.; and Luo, Z.-Q. 2023. Towards Memory-and Time-Efficient Backpropagation for Training Spiking Neural Networks. *arXiv preprint arXiv:2302.14311*.

Qiu, X.; Zhu, R.-J.; Chou, Y.; Wang, Z.; Deng, L.-j.; and Li, G. 2024. Gated attention coding for training high-performance and efficient spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 601–610.

Roy, K.; Jaiswal, A.; and Panda, P. 2019. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784): 607–617.

Rueckauer, B.; Lungu, I.-A.; Hu, Y.; Pfeiffer, M.; and Liu, S.-C. 2017. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11: 682.

Schuman, C. D.; Kulkarni, S. R.; Parsa, M.; Mitchell, J. P.; Kay, B.; et al. 2022. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2(1): 10–19.

Sun, G.; Cholakkal, H.; Khan, S.; Khan, F.; and Shao, L. 2020. Fine-grained recognition: Accounting for subtle differences between similar classes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12047–12054.

Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1568–1576.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Yang, L.; Xu, Y.; Wang, S.; Yuan, C.; Zhang, Z.; Li, B.; and Hu, W. 2022a. PDNet: Toward better one-stage object detection with prediction decoupling. *IEEE Transactions on Image Processing*, 31: 5121–5133.

Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022b. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9499–9508.

Yang, L.; Zhang, R.-Y.; Li, L.; and Xie, X. 2021. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning*, 11863–11874. PMLR.

Yang, L.; Zhang, Z.; Qi, Z.; Xu, Y.; Liu, W.; Shan, Y.; Li, B.; Yang, W.; Li, P.; Wang, Y.; et al. 2024. Exploiting contextual objects and relations for 3d visual grounding. *Advances in Neural Information Processing Systems*, 36.

Yao, M.; Zhang, H.; Zhao, G.; Zhang, X.; Wang, D.; Cao, G.; and Li, G. 2023. Sparser spiking activity can be better: Feature Refine-and-Mask spiking neural network for event-based visual recognition. *Neural Networks*, 166: 410–423.

Yi, K.; and Wu, J. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7017–7025.

Yin, B.; Zhang, X.; Fan, D.-P.; Jiao, S.; Cheng, M.-M.; Van Gool, L.; and Hou, Q. 2024. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, X.; Wei, Y.; Feng, J.; Yang, Y.; and Huang, T. S. 2018. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1325–1334.

Zheng, H.; Wu, Y.; Deng, L.; Hu, Y.; and Li, G. 2021. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11062–11070.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.