

Cross Knowledge Distillation between Artificial and Spiking Neural Networks

Shuhan Ye^{2†}, Yuanbin Qian^{2†}, Chong Wang^{1,2*}, Sunqi Lin², Jiazen Xu², Jiangbo Qian^{1,2}, and Yuqi Li²

¹Merchants' Guild Economics and Cultural Intelligent Computing Laboratory, Ningbo University, Ningbo, China

²Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China

Abstract—Recently, Spiking Neural Networks (SNNs) have demonstrated rich potential in computer vision domain due to their high biological plausibility, event-driven characteristic and energy-saving efficiency. Still, limited annotated event-based datasets and immature SNN architectures result in their performance inferior to that of Artificial Neural Networks (ANNs). To enhance the performance of SNNs on their optimal data format, DVS data, we explore using RGB data and well-performing ANNs to implement knowledge distillation. In this case, solving cross-modality and cross-architecture challenges is necessary. In this paper, we propose cross knowledge distillation (CKD), which not only leverages semantic similarity and sliding replacement to mitigate the cross-modality challenge, but also uses an indirect phased knowledge distillation to mitigate the cross-architecture challenge. We validated our method on main-stream neuromorphic datasets, including N-Caltech101 and CEP-DVS. The experimental results show that our method outperforms current State-of-the-Art methods. The code will be available at <https://github.com/ShawnYE618/CKD>

Index Terms—Spiking neural networks, knowledge distillation, cross-architecture, cross-modality, transfer learning

I. INTRODUCTION

Spiking Neural Networks (SNNs), regarded as the third generation of neural networks [1], have gained significant attention due to their high biological plausibility, event-driven characteristics [2] and energy-saving efficiency [3]. The ambition of SNNs aligns with the algorithm-hardware co-design paradigm of neuromorphic computing, aiming to serve as a low-power alternative to traditional machine intelligence based on Artificial Neural Networks (ANNs) [2]. Unlike ANNs, which use continuous signals and high-power multiply-accumulation operations, SNNs utilize binary spike transmit mechanism, thus leading to substantial improvements in energy efficiency when deployed on neuromorphic hardware [4].

However, many aspects of SNN remain underdeveloped. Binary spike activation maps in SNNs have limited information capacity compared to full-precision activation maps in ANNs [5]. As a result, SNNs struggle to retain sufficient information from membrane potentials during the quantization process, leading to information loss and a subsequent accuracy drop. Moreover, the data modalities most suitable for ANNs and SNNs differ significantly; ANNs are well suited for dense and synchronous (RGB) data, whereas SNNs excel with sparse and asynchronous (DVS) data. Capturing DVS data is both time-consuming and costly [6], so available datasets are not

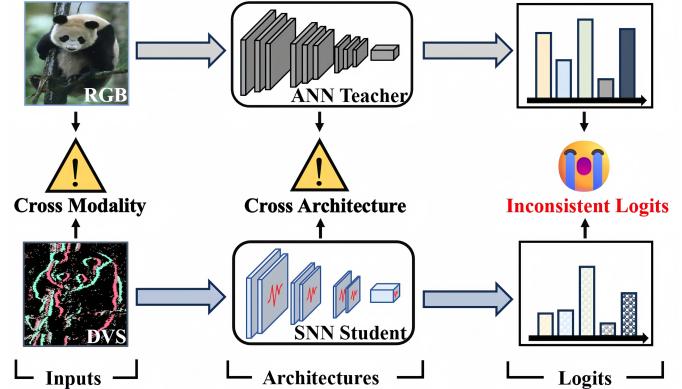


Fig. 1: The challenge of cross-modality and cross-architecture in knowledge distillation between ANNs and SNNs.

only more challenging to collect than RGB frame datasets, but are also typically smaller in scale, which hinders the generalizability of SNN for event-based vision tasks. These challenges put current SNN-based models at a disadvantage compared to their ANN counterparts. Therefore, it is logical to investigate how more effective ANNs, along with RGB datasets that are better suited and more readily available, can be leveraged to enhance the performance of SNNs.

Knowledge distillation (KD) [7] has proved its effectiveness in similar contexts. However, there is limited work for SNN-based knowledge distillation, the few existing cross-architecture approaches failed to consider the correspondence between modality and architecture [8]. In these studies, both models either use RGB data or DVS data. Using ANNs to extract sparse DVS data for distilling SNNs does not fully leverage the potential of ANNs while training SNNs with dense RGB data presents certain risks [9].

As a result, ensuring modality-architecture correspondence is crucial. We aim to leverage non-event data and ANNs to mitigate the limitations imposed by the small dataset scale and the limited performance of SNNs. Therefore, employing cross-architecture knowledge distillation is both necessary and challenging. Due to the inherent differences between RGB and DVS data, the distillation process also presents a cross-modality challenge, as shown in Fig. 1.

Recently, He et al. [10] proposed a method which transfers knowledge from RGB data to DVS data in a shared SNN model. Inspired by their work, we present cross knowledge distillation (CKD), an indirect phased distillation method to

† Equal Contribution

* Corresponding author: Chong Wang

improve the efficiency of RGB data. CKD utilizes a well-trained ANN model to extract knowledge from RGB data and guide the semantically similar cross-modal data stream within the weight-shared SNN model. Subsequently, superior knowledge is transferred to another dynamic stream via cross-modality knowledge transfer. Through our method, which actually employs a RGB-DVS hybrid stream as an intermediary, the gaps of cross-modality and cross-architecture are effectively bridged. Our main contributions are as follows:

- 1) We are the first work to accomplish knowledge distillation that bridges both architecture and modality, simultaneously addressing both RGB data and DVS data, between ANN and SNN.
- 2) We leverage the semantic similarity between RGB and DVS data from the same category to achieve modality and architecture correspondence. By employing an RGB-DVS hybrid stream, we enable an indirect phased distillation, allowing ANNs to process the RGB data, while guiding SNNs to achieve maximum performance improvement on the DVS data best suited for them. This also lays the foundation for SNNs to tackle other temporally demanding tasks in the future.
- 3) Our sufficient experiments on event-based datasets prove the effectiveness of CKD. Remarkably on N-Caltech101 [11], we report a new state-of-the-art top-1 accuracy (97.13%), which is 3.68% higher than our baseline [10] (93.45%), closely approaching the SOTA result of its RGB version Caltech101 [12] on ANN [13] (98.02%). Our results demonstrate that with our CKD method, SNNs could perform almost as well as ANNs.

II. RELATED WORK

Spiking Neural Networks. SNNs draw inspiration from human brain, using discrete spikes for information processing. This method achieves effects comparable to continuous activation functions by accumulating spikes over an additional temporal dimension, making it highly suitable for processing temporal data. Concretely, SNNs replace the traditional activation function by using a spiking neuron model, such as the integrate-and-fire (IF) neuron model and the widely-used Leaky Integrate-and-Fire (LIF) neuron model [14]. The LIF neuron model integrates incoming spikes over time, with its membrane potential and spiking behavior governed by the following equations:

$$\mathbf{u}^{t+1,l} = \tau \mathbf{u}^{t,l} + \mathbf{W}^l \mathbf{s}^{t,l-1} \quad (1)$$

$$\mathbf{s}^{t,l} = H(\mathbf{u}^{t,l} - V_{\text{th}}) \quad (2)$$

$$\mathbf{u}^{t+1,l} = \tau \mathbf{u}^{t,l} \cdot (1 - \mathbf{s}^{t,l}) + \mathbf{W}^l \mathbf{s}^{t+1,l-1} \quad (3)$$

where $\mathbf{u}^{t,l}$ denotes the membrane potential of neurons in layer l at time step t , \mathbf{W}^l represents the weight matrix of layer l , and $\mathbf{s}^{t,l}$ corresponds to the binary spikes emitted by neurons. The Heaviside step function H determines whether a spike is emitted, based on the comparison between $\mathbf{u}^{t,l}$ and the

threshold V_{th} . The leaky factor τ controls the temporal decay of the membrane potential.

Multimodal Machine Learning. Multimodal machine learning refers to the integration and processing of information from multiple data sources or modalities, such as images, text, and audio. In machine learning, multimodal models are designed to understand and fuse different types of data to improve performance on complex tasks by leveraging the complementary strengths of each modality. In this work, we use traditional RGB images and DVS (event) data, which is captured by an event camera, also known as a Dynamic Vision Sensor (DVS). DVS excels in high temporal resolution (microsecond-level), low latency, and high dynamic range (>120 dB). These advantages make them highly compatible with SNNs. Specifically, they asynchronously output a positive or negative event at a pixel location where the brightness change exceeds a certain threshold, which are then integrated into sparse event streams. The output data format is represented as a four-dimensional array \mathbf{e}_i :

$$\mathbf{e}_i = (t_i, x_i, y_i, p_i) \quad (4)$$

where t_i is the microsecond-level timestamp, x_i and y_i are the 2D spatial coordinates, p_i represents the polarity (1/0 or 1/-1) and i denotes the index of the i -th event in the event stream. Although some recent works are attempting to convert the event stream into other forms of event representation [15]–[17] to fully leverage the characteristics of DVS data, the best-performing methods so far still encapsulate the event stream into event frames within fixed time intervals.

Knowledge Distillation. Knowledge distillation is a technique in which a well-performing pre-trained model serves as the teacher, transferring its superior knowledge to a smaller student model for accuracy improvement. Recently, there has been increasing research on knowledge distillation for ANNs, including both intra-architecture and cross-architecture distillation [18]–[20], but this research for SNNs remains relatively limited. This is mainly due to the inherent limitations of the performance of SNNs, where distillation within the same architecture is not very effective [21]. Distilling knowledge from ANNs to SNNs faces tricky cross-architecture challenges. Moreover, ANNs are well-suited for traditional dense RGB data, while SNNs excel in processing sparse DVS data. The mismatch between modality and architecture will lead to inferior performance or limited task applicability [9]. However, previous works have overlooked this important consideration [8]. In this work, we exploit the semantic similarity of cross-modal data within the same category to bridge the cross-modality gap, and use an indirect phased distillation method to bridge the cross-architecture gap.

III. METHODOLOGY

In this section, we introduce our cross knowledge distillation (CKD) method, which contains a cross-modality knowledge transfer module and a cross-architecture knowledge distillation module. CKD achieves effective knowledge distillation from

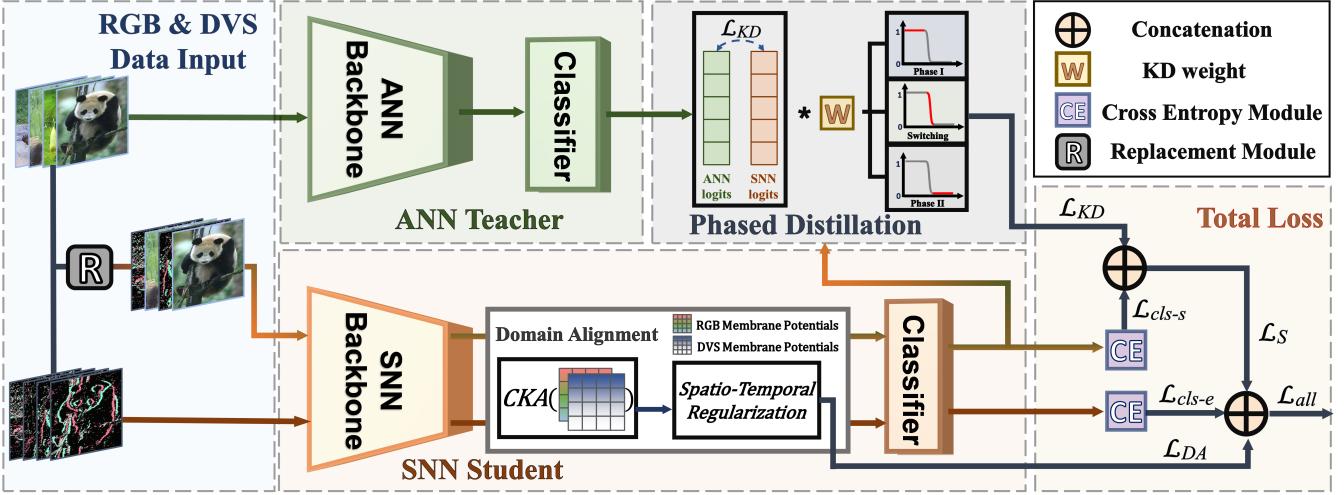


Fig. 2: Overview of our Cross Knowledge Distillation framework between Artificial and Spiking neural networks.

ANNs to SNNs under the modality and architecture correspondence. We utilize both RGB images and DVS data during training, but testing is conducted exclusively with DVS data.

A. Cross-Modality Knowledge Transfer

The cross-modality knowledge transfer module is the basis of our framework, which contains SNN model, domain-alignment module and replacement module.

1) *Domain-alignment Module*: Feeding DVS data and its RGB format counterpart into an identical network results in misaligned feature distributions [10], so domain-alignment processing is necessary to safeguard the effectiveness of knowledge transfer. In this work, we employ a category index to ensure that each pair of inputs corresponds to the same category. Following this, we convert RGB images to HSV (Hue, Saturation, Value) color space and replicate the V channel to match the dimensions of dynamic data in order to feed them in a shared SNN model.

As shown in Fig. 2, domain-alignment loss is employed to constrain the feature distribution differences of the two streams in SNN. Here, we use CKA [22], a proven effective index of network representation similarity, to measure that between the two output feature. The closer the value of CKA is to 1, the more correlated the two vectors are. For this reason, we subtract the CKA from 1 as the domain-alignment loss function \mathcal{L}_{DA} . Considering category-alignment and the unique spatial-temporal domain characteristics of SNNs, in particular, the time steps, the domain alignment loss is expressed as:

$$\mathcal{L}_{DA} = 1 - \frac{1}{T} \sum_{t=1}^T \sum_{y_i=y_j, y \in \mathcal{Y}} CKA(\mathbf{F}_s^{i,t}, \mathbf{F}_d^{j,t}) \quad (5)$$

where T denotes the timesteps of input data, t represents current moment, emphasizing that for dynamic data containing rich temporal information, CKA is calculated at each timestep and then averaged across all of them. The y represents the category of input data, included in the overall set of categories \mathcal{Y} . The formula $y_i = y_j$ indicates that for paired static and

dynamic samples, features $\mathbf{F}_s^{i,t}$ and $\mathbf{F}_d^{j,t}$ correspond to inputs from the same category. Moreover, at the temporal level, we add learnable coefficients to assign weight for each timestep. To prevent the model from overfitting at a certain time step, we adopt DVS data classification loss \mathcal{L}_{cls-e} as regularization term. In this case, the domain alignment loss is defined as:

$$\begin{aligned} \mathcal{L}_{DA} = & \frac{1}{T} \sum_{t=1}^T \sigma(\theta_t) (1 - CKA(\mathbf{F}_s^{i,t}, \mathbf{F}_d^{j,t})) \\ & + \frac{1}{T} \sum_{t=1}^T (1 - \sigma(\theta_t)) \mathcal{L}_{cls-e} \end{aligned} \quad (6)$$

where σ denotes sigmoid function, θ_t represents the coefficient at time step t . For classification loss, we employ TET loss which can compensate the momentum loss of surrogate gradient to improve SNN's generalizability [23]. The overall loss for two streams of static data \mathcal{L}_S can be expressed as:

$$\mathcal{L}_S = \alpha \mathcal{L}_{cls-s} + \beta \mathcal{L}_{DA} \quad (7)$$

where α and β are the coefficients of classification and domain-alignment loss. \mathcal{L}_{cls-s} indicates the classification loss of static data.

2) *Semantically Similar Replacement Module*: Leveraging the semantic similarity between RGB frame-based data and its event-based (DVS) counterpart, this module replaces static data with DVS data by a non-linear probability function $P_{replace}$, which gradually increases from 0 to 1.

$$P_{replace} = \left(\frac{b_i + e_c * b_l}{N_b} \right)^3 \quad (8)$$

where b_i is the index within current training batch, e_c is the number of current epoch while b_l is the length of training batches. N_b represents the sum of batches during the entire training process.

This approach enables DVS data to benefit from the rich knowledge of static data during the early training stage, while progressively shifting the focus to DVS data in later stages,

ensuring a smooth and stable transition throughout the process. Moreover, this module builds a more gradual hybrid modal data stream to alleviate the limitation of SNNs in handling large amounts of pure RGB dense data input. This advantage will become more evident in the future, especially in action recognition with more temporal information and video-based detection tasks.

B. Cross-Architecture Knowledge Distillation

In this work, we use a well-performing ANN teacher model to extract superior knowledge from best-matched modal data, static data, guiding the hybrid data stream within the SNN student model. Subsequently, superior knowledge is transferred from the hybrid stream to the dynamic stream via the cross-modality knowledge transfer module. The distillation between multimodal but same category data effectively bridges the gap of cross-modality and simplifies the problem we face into a more familiar and manageable cross-architecture issue.

1) Knowledge Distillation Loss: As shown in Fig. 2, we choose logits distillation rather than feature distillation. On the one hand, our knowledge transfer module already operates in the feature domain, introducing an extra knowledge distillation module is too direct and influential, in contrast, our CKD correct the features indirectly and effectively as shown in Fig.3. On the other hand, unlike previous cross-architecture feature-domain distillation [24] with strict correspondence between architectural layers, our framework appropriately relaxes the constraints on architecture, thereby reducing the costs of computation and training while expanding the possibilities for cross-architecture distillation.

In this work, we use vanilla KD loss function to align the logits from the hybrid stream of SNN with those of ANN, distilling valuable information to SNN in logits domain and laying the groundwork for subsequent feature-domain knowledge transfer. The vanilla knowledge distillation loss function \mathcal{L}_{KD} , which can be expressed as:

$$\mathcal{L}_{KD} = \sum_{i=1}^N KL(\mathcal{SM}(\mathcal{Z}_t^{st}/\tau)_i \| \mathcal{SM}(\mathcal{Z}_s^{st}/\tau)_i) \quad (9)$$

where \mathcal{Z}_t^{st} and \mathcal{Z}_s^{st} represent the logits of the two static data streams from teacher model and student model respectively. The function \mathcal{SM} denotes the softmax operation, which transforms logits into probability distributions. KL denotes the Kullback-Leibler divergence, a measure of the similarity between two probability distributions, which is computed for each class i here. N denotes the total number of classes while τ represents the temperature of the distillation.

This indirect knowledge distillation effectively bridges the significant representation gap and mitigates the cross-architecture issue, since the knowledge extracted from the ANN is further corrected by the cross-modality knowledge transfer module, which ensures the mitigation of architectural differences.

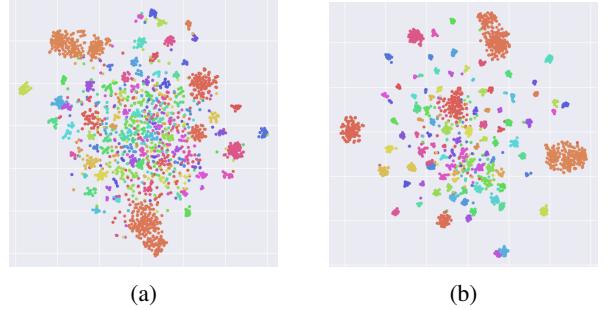


Fig. 3: t-SNE visualizations of (a) the baseline and (b) our method on dataset N-Caltech101.

2) Phased Distillation Strategy: As previously outlined, to ensure training stable and smooth, we adopt a semantically similar replacement module. Consequently, as training deepens, the hybrid stream which initially designated for static data is gradually incorporated and eventually become dominated by dynamic data. This phenomenon contradicts our initial intent of using the static stream within the SNN as an intermediary, to ideally mitigate cross-modality and cross-architecture influences. Therefore, it is crucial to stop the distillation process at a certain point. The probability of replacement in (8) is rising slowly from 0, even remaining below 0.125 at midpoint. This level of mixing is acceptable, so we aim to switch the weight of KD from 1 to 0 when KD brings negative impacts. We use a adjustable function $\gamma(e)$ to enable our switching process manageable:

$$\gamma(e) = 1 - \frac{1}{1 + \exp(-k \cdot (e - e_{th}))} \quad (10)$$

where e_{th} is the threshold of switching epoch, used to control the position of switching, while k is set to control the slope of the switching process. In this case, the overall loss for two streams of static data \mathcal{L}_S in (7) could be redefined as:

$$\mathcal{L}_S = \alpha \mathcal{L}_{cls-s} + \beta \mathcal{L}_{DA} + \gamma(e) \mathcal{L}_{KD} \quad (11)$$

where the \mathcal{L}_{KD} is denied as shown in (9). The total training loss can be expressed as $\mathcal{L}_{all} = \mathcal{L}_S + \mathcal{L}_{cls-e}$.

IV. EXPERIMENTS

We conduct experiments on mainstream event-based datasets N-Caltech101 [11] with its static version Caltech101 [12], and the image-event paired dataset CEP-DVS [24], to evaluate the effectiveness of the proposed method.

A. Implement Details

We integrate all DVS data into frames and resize them to 48x48 for both the N-Caltech101 and CEP-DVS datasets. For a fair comparison, we use the VGGSNN model [10] trained on N-Caltech101 for 300 epochs with 10 time steps, and the Spiking-ResNet18 model [10] trained on CEP-DVS for 200 epochs with 6 time steps. We select the Wide_ResNet101_2 (WRN101_2) [29] as ANN teacher model, which is finetuned 20 epochs from ImageNet in HSV-converted Caltech101. Additionally, we set both the α and β in (11) to 1. All experiments are conducted using the BrainCog framework [30].

TABLE I: Experimental results compared with other works. T is the set time steps

Dataset	Category	Methods	SNN Model	ANN Model	Timesteps	Accuracy(%)
N-Caltech101	Data augmentation	NDA [25]	VGGSN	-	10	78.2
		EventMixer [26]	ResNet-18	-	10	79.2
	Efficient training	TET [23]	VGGSN	-	10	79.27
		TKS [27]	VGGSN	-	10	84.1
		ETC [28]	VGGSN	-	10	85.53
	Domain adaptation	Knowledge-Transfer [10]	VGGSN	-	10	93.18 \pm 0.38 (93.33*)
CEP-DVS	Knowledge distillation	CKD (Ours)	VGGSN	WRN101_2	10	96.71 \pm 0.30 (97.13)
	Efficient training	TET [23]	ResNet-18	-	10	25.05
	Domain adaptation	Knowledge-Transfer [10]	ResNet-18	-	10	30.05 \pm 0.50 (35.40*)
	Knowledge distillation	CKD (Ours)	ResNet-18	WRN101_2	10	38.80 \pm 1.23 (40.20)

^a The results are mean and standard deviation after taking three different seeds. The symbol (*) denotes our implementation of other methods. The best accuracy is shown in parentheses. The accuracy of WRN101_2 is 97.48% on Caltech101 and 61.45% on static part of CEP-DVS.

TABLE II: Ablation studies of different knowledge distillation loss functions on CKD framework.

Network	Methods	Acc(%)	Network	Methods	Acc(%)	
VGGSN	N-Caltech101	baseline	93.18	CEP-DVS	baseline	30.50
		DKD	96.48		DKD	37.63
		LumiNet	95.82		LumiNet	38.43
	DIST	95.59	ResNet-18	DIST	37.75	
	KD	96.71		KD	38.80	

^a The results are mean after taking three different seeds.

B. Comparison with the State-of-the-Art

We first evaluate our CKD on N-Caltech101 and compare it with TET [23], TKS [27], ETC [28], and Knowledge-Transfer [10]. The results presented in TABLE I demonstrate that our method outperforms all the compared methods. We report a new state-of-the-art accuracy of 97.13% on N-Caltech101, which is close to that of our ANN teacher model [29] (97.48%) and the current state-of-the-art for Caltech101 [13] (98.02%).

The results show that with our CKD, SNNs can perform almost as well as ANNs on image classification tasks. At the same time, our CKD achieves cross-modal and cross-architecture knowledge distillation. It also ensures modality-architecture correspondence, thereby fully leveraging the advantages of SNNs, such as energy efficiency, biological plausibility, and the asynchronous nature and high dynamic range of DVS data. This lays a solid foundation for future deployment on brain-inspired chips. As shown in the t-SNE visualization in Fig. 3, compared to our baseline [10], the data points corresponding to different classes are distinctly more separated, with tighter clusters and less overlap between classes. This indicates that knowledge extracted from the ANN has been efficiently transferred to the SNN, leading to more discriminative features and higher accuracy.

C. Ablation Studies

To verify the effect of CKD, we conduct ablation studies on KD loss function and the phase switching function. As shown in TABLE II, on the N-Caltech101, we achieve a 97.13% accuracy using vanilla KD. In contrast, DKD [19], LumiNet [31] and DIST [32] yield relatively lower accuracies than vanilla KD, since these methods exert relatively excessive influence

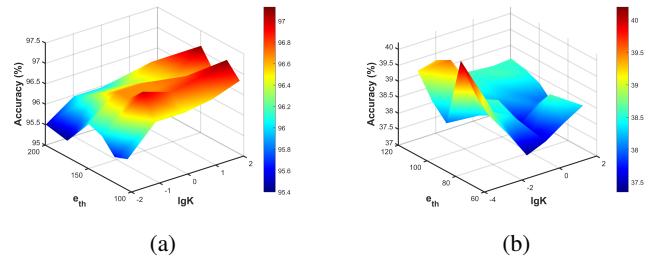


Fig. 4: Visualization of phase switching function parameters ablation study on (a) N-Caltech101 and (b) CEP-DVS.

on the static stream within the SNN, adversely affecting our domain-alignment module and leading to a slightly decline in accuracy. However, they still surpass our baseline by over 2%, demonstrating the superiority of our CKD framework. On the CEP-DVS, the results are similar, further validating our conclusion and the effectiveness of CKD.

For the other ablation study, we perform visualizations using various sets of parameters e_{th} and k in phase switching function. As shown in Eq. 10, the k is set to control the slope of our switching function, when k becomes greater, the switching becomes swifter. The e_{th} controls the beginning epoch of the switching. As shown in Fig. 4, we apply base-10 logarithm to the k values to improve their readability.

On the N-Caltech101, peaks (97.13%) occur at $e_{th} = 149.5, k = 100$ and $e_{th} = 119.5, k = 100$ while declining when k turns small or e_{th} deviates excessively from midpoint. The lowest point (95.40%) occurs at $e_{th} = 174.5, k = 0.01$. These results indicate that on this dataset the performance of this SNN model improves the most when the switching is swift and occurs around the midpoint, whereas the effect is relatively modest during a delayed and smooth switching process. On the CEP-DVS, peak (40.2%) are observed at small values of k , with $e_{th} = 89.5, k = 0.001$, representing that slower and steadier phase switching allows SNNs to learn more and better from ANNs. The lowest point (37.35%) occurs at $e_{th} = 69.5, k = 0.1$. These visualizations illustrated the effects of different switching strategies. Nevertheless, all results from this ablation study surpass our baseline, validating the effectiveness of our CKD method.

V. CONCLUSION

In this paper, we proposed cross knowledge distillation (CKD), a novel method to bridge both architectural and modal differences between ANNs and SNNs as well as RGB and DVS data. Through extensive experiments on main-stream neuromorphic datasets, we proved the effectiveness of CKD. We achieve a new SOTA top-1 accuracy of 97.13% on N-Caltech101, a significant improvement over previous works, as well as a competitive accuracy of 40.20% on CEP-DVS. Our results show that SNNs, with the help of CKD, can perform on par with ANNs, making them viable for real-world vision tasks. Moreover, the proposed method lays a solid foundation for future research, enabling SNNs to tackle more complex and temporally demanding tasks, potentially paving the way for their deployment in neuromorphic computing systems.

VI. ACKNOWLEDGMENTS

This work was supported by the Ningbo Municipal Natural Science Foundation of China (No. 2022J114), National Natural Science Foundation of China (No. 62271274), Ningbo S&T Project (No.2024Z004) and Ningbo Major Research and Development Plan Project (No.2023Z225)

REFERENCES

- [1] Wolfgang Maass, “Networks of spiking neurons: the third generation of neural network models,” *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.
- [2] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda, “Towards spike-based machine intelligence with neuromorphic computing,” *Nature*, vol. 575, no. 7784, pp. 607–617, 2019.
- [3] Philipp Akopyan, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam, Yutaka Nakamura, Pallab Datta, Gi-Joon Nam, et al., “Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip,” *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 34, no. 10, pp. 1537–1557, 2015.
- [4] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al., “Loihi: A neuromorphic manycore processor with on-chip learning,” *Ieee Micro*, vol. 38, no. 1, pp. 82–99, 2018.
- [5] Yufei Guo, Yuanpei Chen, Xiaode Liu, Weihang Peng, Yuhang Zhang, Xuhui Huang, and Zhe Ma, “Ternary spike: Learning ternary spikes for spiking neural networks,” *arXiv preprint arXiv:2312.06372*, 2023.
- [6] Yuanbin Qian, Shuhan Ye, Chong Wang, Xiaojie Cai, Jiangbo Qian, and Jiafei Wu, “Ucf-crime-dvs: A novel event-based dataset for video anomaly detection with spiking neural networks,” *arXiv preprint arXiv:2503.12905*, 2025.
- [7] Geoffrey Hinton, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [8] Di Hong, Jiangrong Shen, Yu Qi, and Yueming Wang, “Lasnn: Layer-wise ann-to-snn distillation for effective and efficient training in deep spiking neural networks,” *arXiv preprint arXiv:2304.09101*, 2023.
- [9] Shiting Xiao, Yuhang Li, Youngeun Kim, Donghyun Lee, and Priyadarshini Panda, “Respike: Residual frames-based hybrid spiking neural networks for efficient action recognition,” *arXiv preprint arXiv:2409.01564*, 2024.
- [10] Xiang He, Dongcheng Zhao, Yang Li, Guobin Shen, Qingqun Kong, and Yi Zeng, “An efficient knowledge transfer strategy for spiking neural networks from static to event domain,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 512–520.
- [11] Orchard Garrick, Jayawant Ajinkya, Gregory K. Cohen, and Thakor Nitish, “Converting static image datasets to spiking neuromorphic datasets using saccades,” *Frontiers in Neuroscience*, vol. 9, no. 178, 2015.
- [12] Li Fei-Fei, Rob Fergus, and Pietro Perona, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” *Computer Vision and Pattern Recognition Workshop*, 2004.
- [13] H M Dipu Kabir, “Reduction of class activation uncertainty with background information,” 2024.
- [14] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition*, Cambridge University Press, 2014.
- [15] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis, “Unsupervised event-based learning of optical flow, depth, and egomotion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 989–997.
- [16] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman, “Hots: a hierarchy of event-based time-surfaces for pattern recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1346–1359, 2016.
- [17] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman, “Hats: Histograms of averaged time surfaces for robust event-based object classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1731–1740.
- [18] Yujie Zheng, Chong Wang, Chenchen Tao, Sunqi Lin, Jiangbo Qian, and Jiafei Wu, “Restructuring the teacher and student in self-distillation,” *IEEE Transactions on Image Processing*, 2024.
- [19] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang, “Decoupled knowledge distillation,” 2022.
- [20] Sunqi Lin, Chong Wang, Yujie Zheng, Chenchen Tao, Xinmiao Dai, and Yuqi Li, “Distill vision transformers to cnns via teacher collaboration,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 5925–5929.
- [21] Ravi Kumar Kushawaha, Saurabh Kumar, Biplab Banerjee, and Rajbabu Velmurugan, “Distilling spikes: Knowledge distillation in spiking neural networks,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4536–4543.
- [22] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton, “Similarity of neural network representations revisited,” in *International conference on machine learning*. PMLR, 2019, pp. 3519–3529.
- [23] Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu, “Temporal efficient training of spiking neural network via gradient re-weighting,” in *International Conference on Learning Representations*, 2022.
- [24] Yongjian Deng, Hao Chen, Huiying Chen, and Youfu Li, “Learning from images: A distillation learning framework for event cameras,” *IEEE Transactions on Image Processing*, vol. 30, pp. 4919–4931, 2021.
- [25] Yuhang Li, Youngeun Kim, Hyoungseob Park, Tamar Geller, and Priyadarshini Panda, “Neuromorphic data augmentation for training spiking neural networks,” 2022.
- [26] Guobin Shen, Dongcheng Zhao, and Yi Zeng, “Eventmix: An efficient data augmentation strategy for event-based learning,” *Information Sciences*, vol. 644, pp. 119170, 2023.
- [27] Yiting Dong, Dongcheng Zhao, Yang Li, and Yi Zeng, “An unsupervised stdp-based spiking neural network inspired by biologically plausible learning rules and connections,” *Neural Networks*, vol. 165, pp. 799–808, 2023.
- [28] Dongcheng Zhao, Guobin Shen, Yiting Dong, Yang Li, and Yi Zeng, “Improving stability and performance of spiking neural networks through enhancing temporal consistency,” *Pattern Recognition*, vol. 159, pp. 111094, 2025.
- [29] H M Dipu Kabir, Moloud Abdar, Abbas Khosravi, Seyed Mohammad Jafar Jalali, Amir F. Atiya, Saeid Nahavandi, and Dipti Srinivasan, “Spinalnet: Deep neural network with gradual input,” *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 5, pp. 1165–1177, Oct. 2023.
- [30] Yi Zeng, Dongcheng Zhao, Feifei Zhao, Guobin Shen, Yiting Dong, Enmeng Lu, Qian Zhang, Yingqian Sun, Qian Liang, Yuxuan Zhao, et al., “Braincog: A spiking neural network based, brain-inspired cognitive intelligence engine for brain-inspired ai and brain simulation,” *Patterns*, vol. 4, no. 8, 2023.
- [31] Md Ismail Hossain, MM Elahi, Sameera Ramasinghe, Ali Cheraghian, Fuad Rahman, Nabeel Mohammed, and Shafin Rahman, “Luminet: The bright side of perceptual knowledge distillation,” *arXiv preprint arXiv:2310.03669*, 2023.
- [32] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu, “Knowledge distillation from a stronger teacher,” *arXiv preprint arXiv:2205.10536*, 2022.