Efficient Logit-based Knowledge Distillation of Deep Spiking Neural Networks for Full-Range Timestep Deployment

Chengting Yu†, Xiaochen Zhao†, Lei Liu, Shu Yang, Gaoang Wang, Erping Li, and Aili Wang*

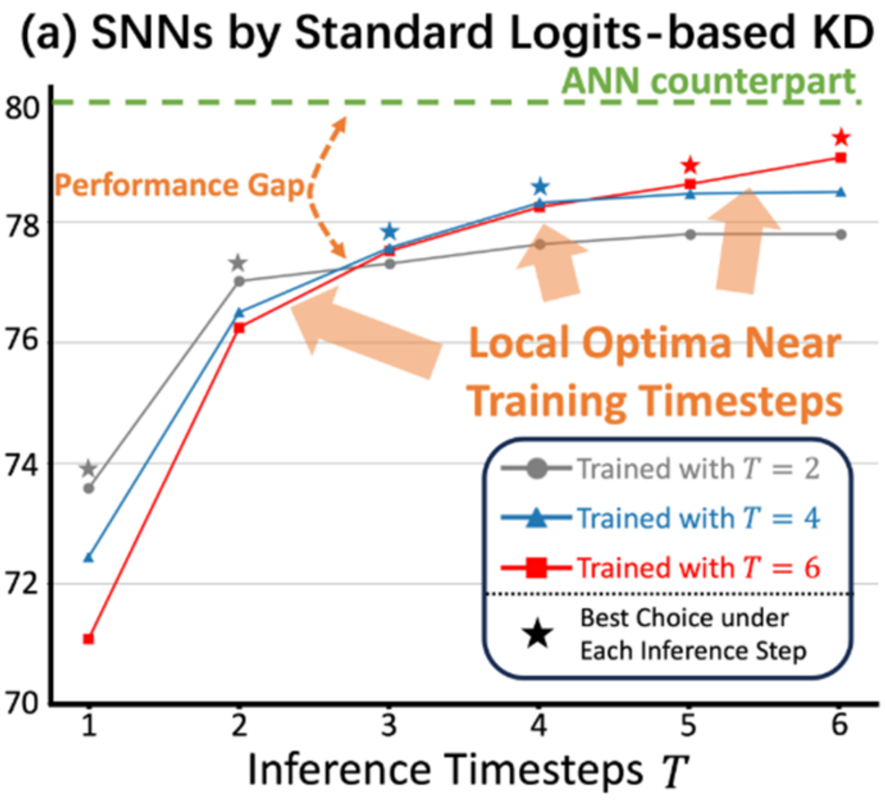
ZJUI Institute, Zhejiang University
Haining, Zhejiang, China

Code link: https://github.com/Intelli-Chip-Lab/snn_temporal_decoupling_distillation

Motivation

- SNNs are brain-inspired models
  - Offer a potential **energy efficiency** advantage on neuromorphic hardware
  - An alternative to traditional ANNs

- Major limitations of SNNs:
  - **Lower accuracy** compared to ANNs
  - The fixed inference timesteps **restrict adaptability**
  - Changing inference timesteps requires **retraining**

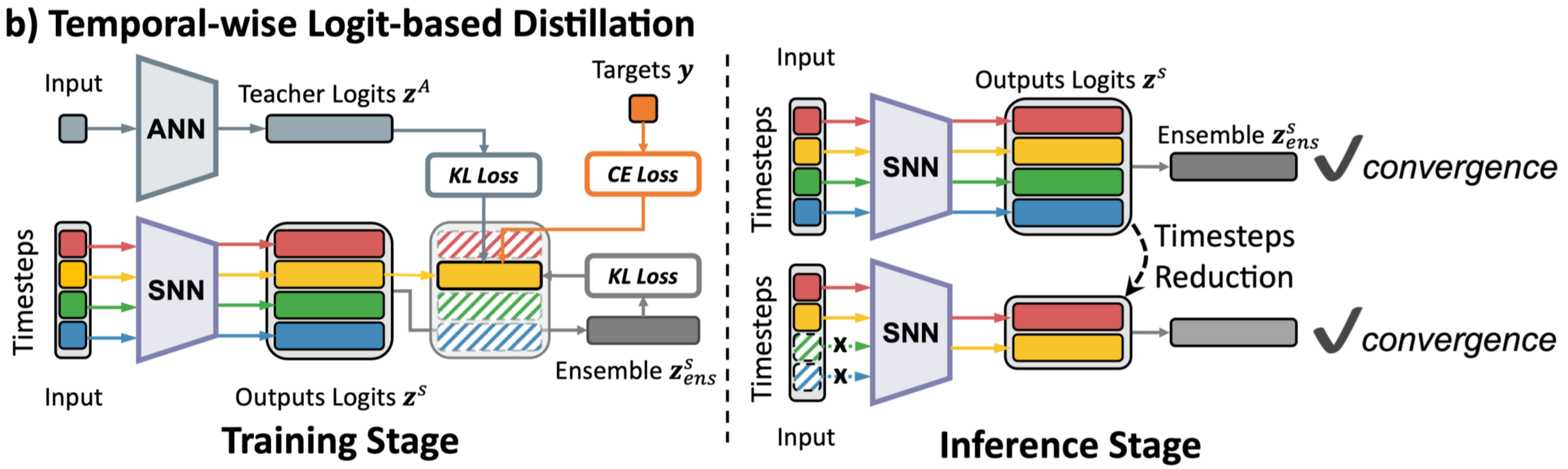


Key Innovation

- **Approach**: leverages the **spatiotemporal properties** of SNNs
- **Proposed Solution**: a **novel distillation framework** for deep SNNs
  - Works across a **full range of timesteps**
  - **No retraining needed** when inference timesteps change.
- **Theoretical Contribution**:
  - Proof that training leads to **convergence for all time-based models**.
- **Empirical Results**:
  - Tested on CIFAR-10, CIFAR-100, CIFAR10-DVS, and ImageNet
  - Achieves **state-of-the-art performance**

Method Overview

- Transforms traditional logits-based distillation into **temporal-wise distillation**
- Integrates ensemble learning-based **self-distillation**



Temporal-wise Distillation for Deep SNNs

- Unique to SNNs: SNNs **generate logits at multiple timesteps**
- Insight:
  - Viewing SNN outputs over time as an **ensemble**
  - Accuracy improves when **each timestep's output** becomes better
- Proposed Method
  - Redefine distillation targets to **include logits from all timesteps** $z^S_{ens} = \frac{1}{T}\sum_t z^S(t)$
  - Temporal-wise cross-entropy (TWCE) for **hard targets**

$$\mathcal{L}_{TWCE} = \frac{1}{T}\sum_t \mathcal{L}_{CE}\left(S\left(z^S(t)\right), y\right)$$

  - Temporal-wise KL divergence for **soft labels**

$$\mathcal{L}_{TWKL} = \frac{1}{T}\sum_t \mathcal{L}_{KL}\left(S\left(z^S(t)/\tau\right), S\left(z^A/\tau\right)\right)$$

  - The **overall objectives** for temporal-wise distillations $\mathcal{L}_{TWKD} = \mathcal{L}_{TWCE} + \alpha\mathcal{L}_{TWKL}$

Ensemble Learning-based Self-Distillation

- **Key Observation**:
  - **Voting logits (averaged over time)** are more effective
  - Consistent with results from **student-ensemble learning research**
- **Proposed Method**:
  - Adding **final voting logits** as an **additional soft label** for self-distillation

$$\mathcal{L}_{TWSD} = \frac{1}{T}\sum_{t} \mathcal{L}_{KL}\left(S\left(z^{S}(t)/\tau\right), S\left(z^{S}_{ens}/\tau\right)\right)$$

- **Effectiveness**:
  - **Enhances the model's learning** without increasing computational cost
  - Integrates seamlessly with the temporal-wise framework **for better performance**
- **Overall Training Objective**: $\mathcal{L}_{TWKD} = \mathcal{L}_{TWCE} + \alpha\mathcal{L}_{TWKL} + \beta\mathcal{L}_{TWSD}$

Convergence of Temporal-wise Distillation

- Problem Identified by Deng et al., 2022:
  - SNNs may **struggle with convergence** in classification tasks due to **high second-order moments**
- Solution:
  - Optimize **outputs at each timestep** helps avoid convergence issues
- Theoretical Support:
  - **$\mathcal{L}_{TWCE}$ forms the upper bound of $\mathcal{L}_{SCE}$**

$$\mathcal{L}_{SCE} = -\sum_i y_i log\, S_i\left(z^S_{ens}(t), y\right) \leq -\frac{1}{T}\sum_t \sum_i y_i log\, S_i\left(z^S(t), y\right) = \mathcal{L}_{TWCE}$$

  - Similarly, **soft-label objectives** can also be **temporally decoupled**
  - Thus, we have $\mathcal{L}_{SKD} \leq \mathcal{L}_{TWKD}$

Results--Performance Comp. on Benchmarks

Results on CIFAR10 and CIFAR100 Datasets

Table 1. Performance comparison of top-1 accuracy (%) on CIFAR-10 and CIFAR-100 datasets, averaged over three experimental runs.

	Method	Model	Timestep	Top-1 Acc. (%)	
				CIFAR-10	CIFAR-100
Direct-training	STBP-tdBN (Zheng et al., 2021)	ResNet-19	6 4 2	93.16 92.92 92.34	- - -
	Dspike (Li et al., 2021b)	ResNet-18	6 4 2	94.25 93.66 93.13	74.24 73.35 71.68
	TET (Deng et al., 2022)	ResNet-19	6 4 2	94.50 94.44 94.16	74.72 74.47 72.87
	RecDis (Guo et al., 2022b)	ResNet-19	6 4 2	95.55 95.53 93.64	- 74.10 -
	DSR (Meng et al., 2022)	ResNet-18	20	95.10	78.50
	SSF (Wang et al., 2023a)	ResNet-18	20	94.90	75.48
	SLTT (Meng et al., 2023)	ResNet-18	6	94.4	74.38
	OS (Zhu et al., 2023)	ResNet-19	4	95.20	77.86
	RateBP (Yu et al., 2024)	ResNet-18	6 4 2	95.90 95.61 94.75	79.02 78.26 75.97
		ResNet-19	6 4 2	96.36 96.26 96.23	80.83 80.71 79.87
w/ distillation	KDSNN (Xu et al., 2023b)	ResNet-18	4	93.41	-
	Joint A-SNN (Guo et al., 2023b)	ResNet-18	4 2	95.45 94.01	77.39 75.79
		ResNet-34	4 2	96.07 95.13	79.76 77.11
	SM (Deng et al., 2023)	ResNet-18	4	94.07	79.49
		ResNet-19	4	96.82	81.70
	SAKD (Qiu et al., 2024a)	ResNet-19	4	96.06	80.10
	BKDSNN (Xu et al., 2024)	ResNet-19	4	94.64	74.95
	TSSD (Zuo et al., 2024)	ResNet-18	2	93.37	73.40
	TKS (Dong et al., 2024)	ResNet-19	4	96.35	79.89
	EnOF (Guo et al.)	ResNet-19	2	96.19	82.43
	SuperSNN (Zhang et al.)	ResNet-19	6 2	95.61 95.08	77.45 76.49
	Our	ResNet-18	6 4 2	**95.96** **95.57** **95.11**	**79.80** **79.10** **77.32**
		ResNet-19	6 4 2	**97.00** **96.97** **96.65**	**82.56** **82.47** **81.47**

Results on ImageNet and CIFAR10-DVS Datasets

Table 2. Performance comparison of top-1 accuracy (%) on ImageNet with single crop.

Method	Model	Timestep	Acc. (%)
STBP-tdBN (Zheng et al., 2021)	ResNet-34 ResNet-50	6 6	63.72 64.88
Dspike (Li et al., 2021b)	ResNet-34	6	68.19
RecDis (Guo et al., 2022b)	ResNet-34	6	67.33
TET (Deng et al., 2022)	ResNet-34	4	68.00
OS (Zhu et al., 2023)	ResNet-34	4	67.54
RateBP (Yu et al., 2024)	ResNet-34	4	70.01
KDSNN (Xu et al., 2023b)	ResNet-34	4	67.18
LaSNN (Hong et al., 2023)	ResNet-34	4	66.94
SM (Deng et al., 2023)	ResNet-34	6 4	69.35 68.25
SAKD (Qiu et al., 2024a)	ResNet-34	4	70.04
TKS (Dong et al., 2024)	ResNet-34	4	69.60
EnOF (Guo et al.)	ResNet-34	4	67.40
Our	ResNet-34	4	**71.04**

Table 3. Performance comparison of top-1 accuracy (%) on CIFAR10-DVS, averaged over three experimental runs.

Method	Model	Timestep	Acc. (%)
STBP-tdBN (Zheng et al., 2021)	ResNet-19	10	67.80
Dspike (Li et al., 2021b)	ResNet-18	10	75.40
RecDis (Guo et al., 2022b)	ResNet-19	10	72.42
TET (Deng et al., 2022)	VGGSNN	10	83.17
SM (Deng et al., 2023)	ResNet-18	10	83.19
SSF (Wang et al., 2023a)	VGG-11	20	78.00
SLTT (Meng et al., 2023)	VGG-11	10	77.17
SAKD (Qiu et al., 2024a)	VGG-11 ResNet-19	4 4	81.50 80.30
Our	ResNet-18	4 10	**83.50** **86.40**

- **Performance**:
  - Achieves **comparable or superior accuracy**
  - **Effectively reduces the accuracy gap** between SNNs and ANNs.
- **ANN-Guided Distillation Cost**: **running the ANN teacher** model to generate soft labels.

Results--Ablation Study

Ablation Study of Training Objectives

Table 5. Performance comparison on objectives combinations using ResNet-18 on the CIFAR100 dataset.

T	$\mathcal{L}_{\text{TWCE}}$	w/ $\mathcal{L}_{\text{TWSD}}$	w/ $\mathcal{L}_{\text{TWKL}}$	w/ $\mathcal{L}_{\text{TWKL}}$&$\mathcal{L}_{\text{TWSD}}$
4	78.58	78.94	79.05	**79.10**
6	79.26	79.63	79.56	**79.80**

- With $\mathcal{L}_{TWKL}$ > $\mathcal{L}_{TWCE}$ only
- $\mathcal{L}_{TWSD}$ **improves further**
- All three components ($\mathcal{L}_{TWCE}$, $\mathcal{L}_{TWKL}$, $\mathcal{L}_{TWSD}$) are **mutually compatible**
- Work together to improve the model's **accuracy and stability**.

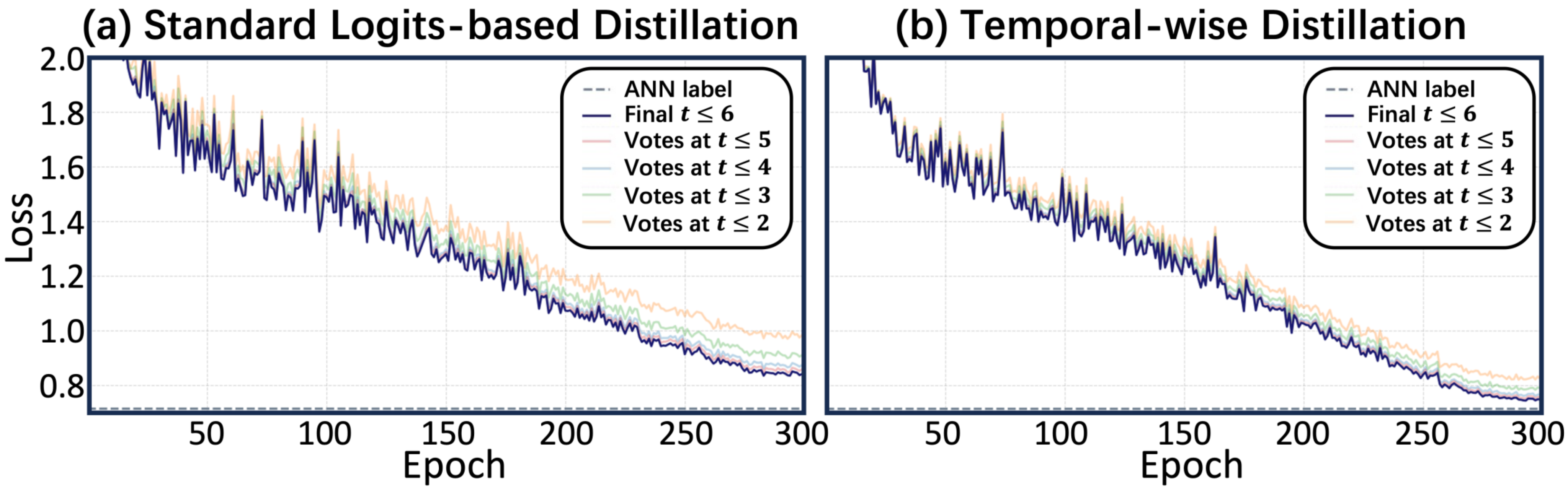
Comparison Study on Temporal Decoupling

Table 6. Performance comparison of temporal decoupling on hard targets and soft labels using ResNet-18 on the CIFAR100 dataset.

T	$\mathcal{L}_{\text{SCE}}$	$\mathcal{L}_{\text{TWCE}}$	$\mathcal{L}_{\text{SKL}}$	$\mathcal{L}_{\text{TWKL}}$	Accuracy (%)
4	✓		✓		78.32
	✓			✓	78.60
		✓	✓		78.74
		✓		✓	**79.05**
6	✓		✓		79.07
	✓			✓	79.15
		✓	✓		79.32
		✓		✓	**79.56**

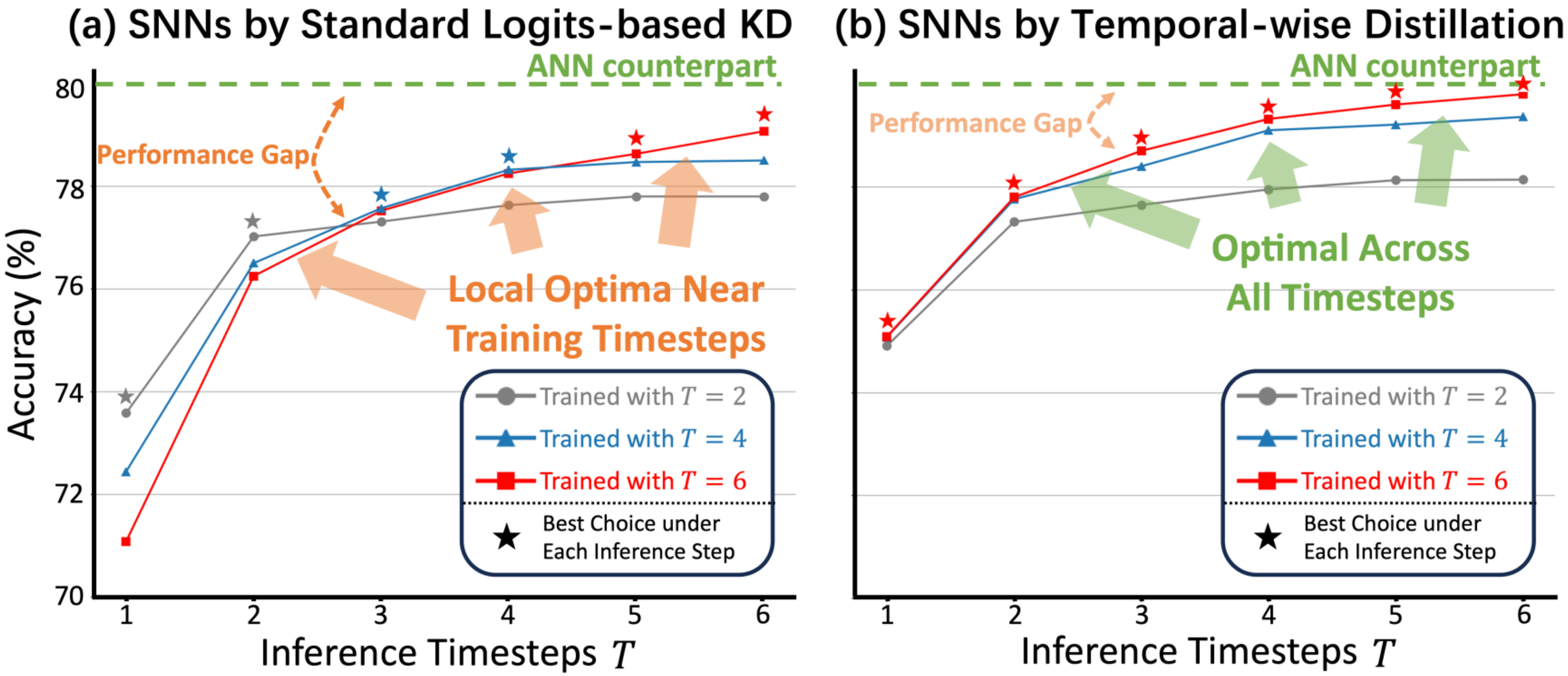
- Decoupling either $\mathcal{L}_{SCE}$ **or** $\mathcal{L}_{SKL}$ **individually** improves performance
- **Combining both decoupled losses** leads to the **best overall performance**.

Results--Loss Convergence



- Temporal decoupling:
  - **Enhances convergence** of loss across different timesteps
  - Loss trajectories become **tighter and more uniform**, indicating stable learning
  - **Matches theoretical expectations**

Results--Analysis of Full-Range Performance



- In standard logits-based distillation
  - Each model performs best only in **a narrow timestep range**
- Proposed temporal-wise logits-based distillation
  - A single model trained at **T = 6 performs well across all inference timesteps (1 to 6)**.
  - **Reducing the need to retrain** for different deployment scenarios

Conclusion

- **Problem Addressed:** Inflexibility and performance issues in SNNs
- **Proposed Method**: A novel knowledge distillation framework for deep SNNs
  - Introduces **temporal decoupling** into the **logits-based** distillation framework for SNNs
  - Integrates ensemble learning-based self-distillation
  - Provides both theoretical analysis and empirical experiments
- **Experimental Results**:
  - One of the **most efficient ANN-guided training strategies** for SNNs in terms of performance and computational cost
  - Enables **robust training and generalization** across a full range of inference timesteps
  - Aims to support **broader adoption and development** of SNN-based technologies

Acknowledgement & Resources

- **Funding Grants**
  - NSFC with Grant No. 62304203
  - NSF of Zhejiang Province, China with Grant No. LQ22F010011
  - The ZJU-YST joint research center for fundamental science.

- **Resources**

Thank you!

Contact us: {chengting.21, xiaochen.24, ailiwang}@intl.zju.edu.cn