SDPGO: Efficient Self-Distillation Training Meets Proximal Gradient Optimization

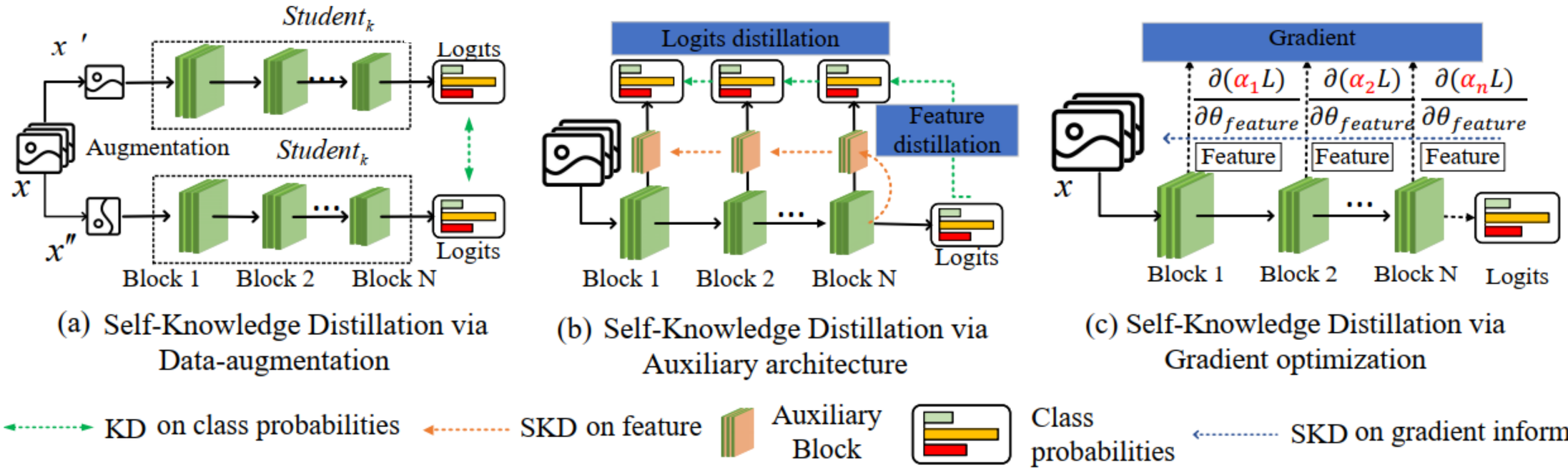
Tongtong Su[1], Liao Yun2*, Fengbo Zheng1*

[1]School of Computer and Information Engineering, Tianjin Normal University
[2]College of Artificial Intelligence, Tianjin University of Science and Technology

天津师范大学
TIANJIN NORMAL UNIVERSITY

Background

Existing SKD approaches often rely on fixed or heuristically defined weights to prioritize features during knowledge transfer.



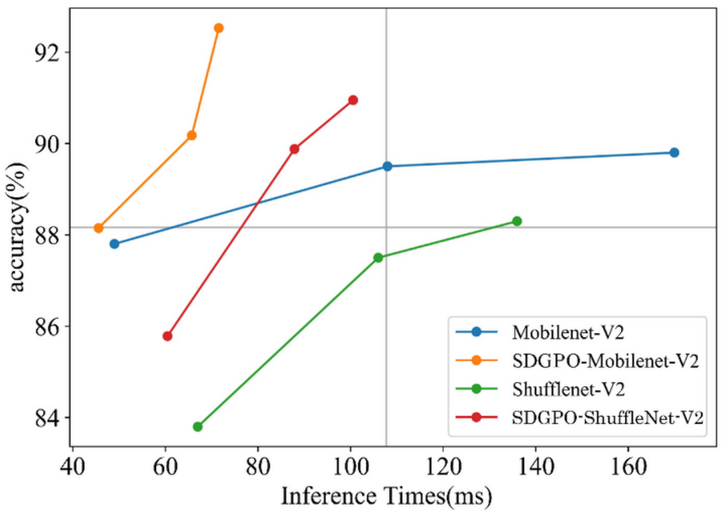
remain static weights

real-time gradient analysis

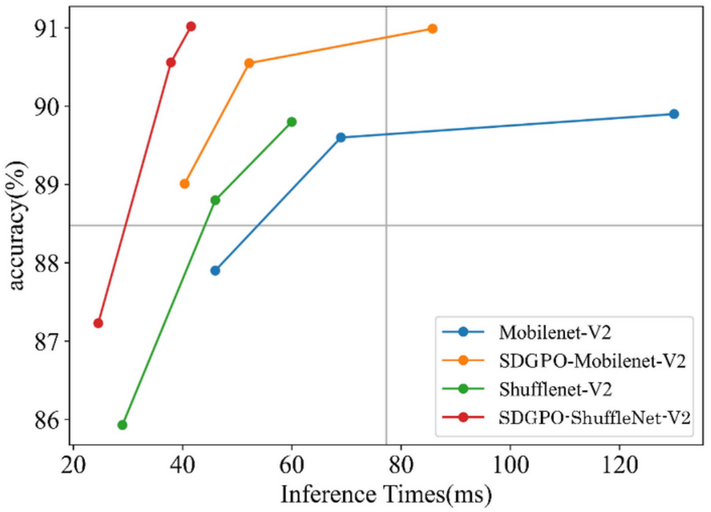
Exploratory Experiments

Efficiency and Memory Consumption :

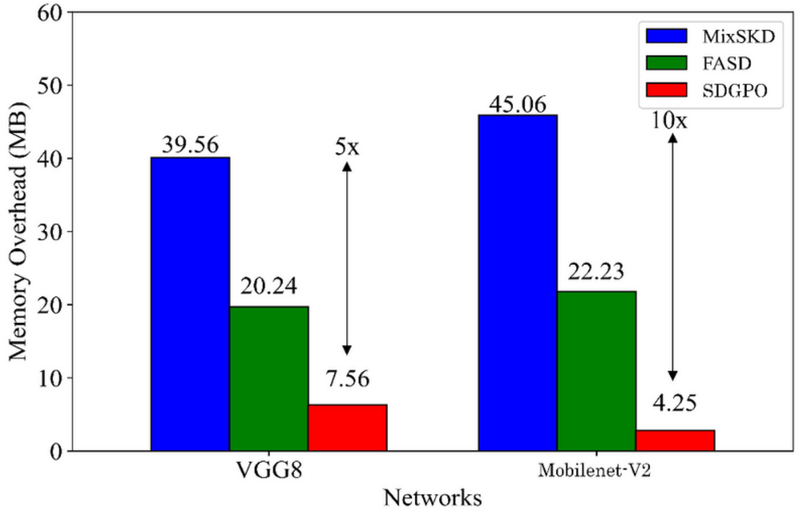
- SDPGO achieves an average inference acceleration optimization of 23.46% on Shuffle-V2 and Mobile-V2 backbone.
- SDPGO delivers significant memory compression, achieving approximately 3 × reduction compared to the state-of-the-art FASD[1] method and 5 × reduction compared to MixSKD[2] on VGG-8.



(a) Raspberty Pi 4B (CPU：BCM2711)



(b) Raspberty Pi 5 (CPU：BCM2711)



(c) Memory consumption on Raspberty Pi 5

[1] Kai Xu, Lichun Wang, Shuang Li, Jianjia Xin, and Baocai Yin. Self-distillation with augmentation in feature space. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[2] Chuanguang Yang, Zhulin An, Helong Zhou, and Qian Zhang. Mixskd: Self-knowledge distillation from mixup for image recognition. In *European Conference on Computer Vision*, pages 534–551. Springer, 2022. 0

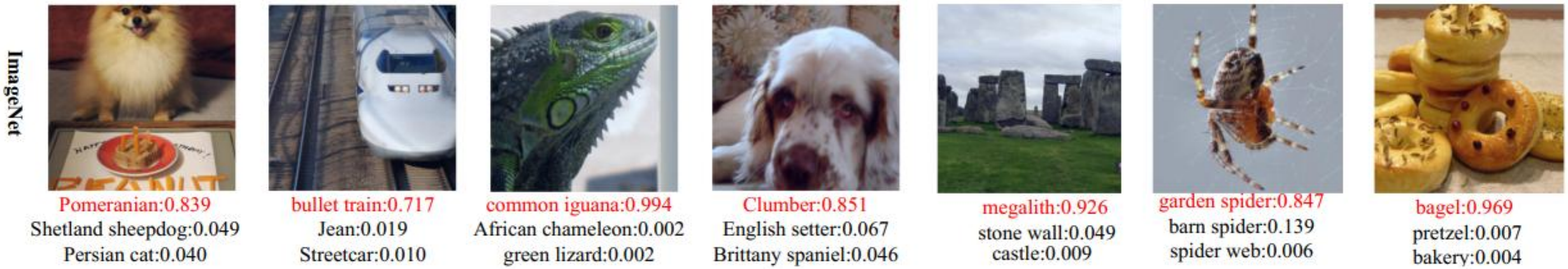
Why can our SDPGO method win?

Visualization：

- the target and top-2 non-target class values of our customized soft labels

Observation：

- SDPGO, generates customized soft labels for each image during training, including the target class and the top two non-target classes.
- By using the gradient-based dynamic weight assignment , higher weights are assigned to classes similar to the target, enhancing the overall learning process.



Why can our SDPGO method win?

Sequential Iterative Learning Module

The optimization goal of DNN:

- we partition each mini-batch into two sequential segments: half aligned with the prior iteration and half with the next iteration.

- Our work uses historical information from the last batch to efficiently generate soft targets as more instant smoothed labels for regularization.

$$\mathcal{B}_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n} \rightarrow \mathbf{p}_i^{\tau,t}, \quad \mathbf{p}_i^{\tau,t-1}$$

$$\mathcal{L}_{SIL} = -\tau^2 \sum_{i=1}^{n} \frac{1}{n} \cdot \underbrace{\sum_{i=1}^{n} \mathbf{p}_i^{\tau,t-1} \log \frac{\mathbf{p}_i^{\tau,t}}{\mathbf{p}_i^{\tau,t-1}}}_{D_{KL}}.$$

Difference:

- the teacher model dynamically evolves during training, with the *t*-th iteration predictions used as the teacher's knowledge without incurringany loss.

Why can our SDPGO method win?

Proximally Weight Assignment Module

- Feature importance is evaluated via gradient magnitudes
- Refine the raw gradient-derived weights via a proximal operator to enforce sparsity and stability

$$w_k^l = \frac{1}{W}\sum_{i=1}^{W}\sum_{j=1}^{H}\left|\frac{\partial L_{CE}}{\partial F_{i,j,k}^l}\right|$$

$$\hat{w}_k^l = \mathrm{Prox}_\lambda\left(w_k^l\right) = \begin{cases} w_k^l - \lambda & \text{if } w_k^l > \lambda \\ 0 & \text{otherwise} \end{cases}$$

- Use an adaptive threshold controlling feature sparsity using only moving averages
- Compute the mean absolute value of the gradient for each weight parameter. Then, we apply z-score normalization to these values

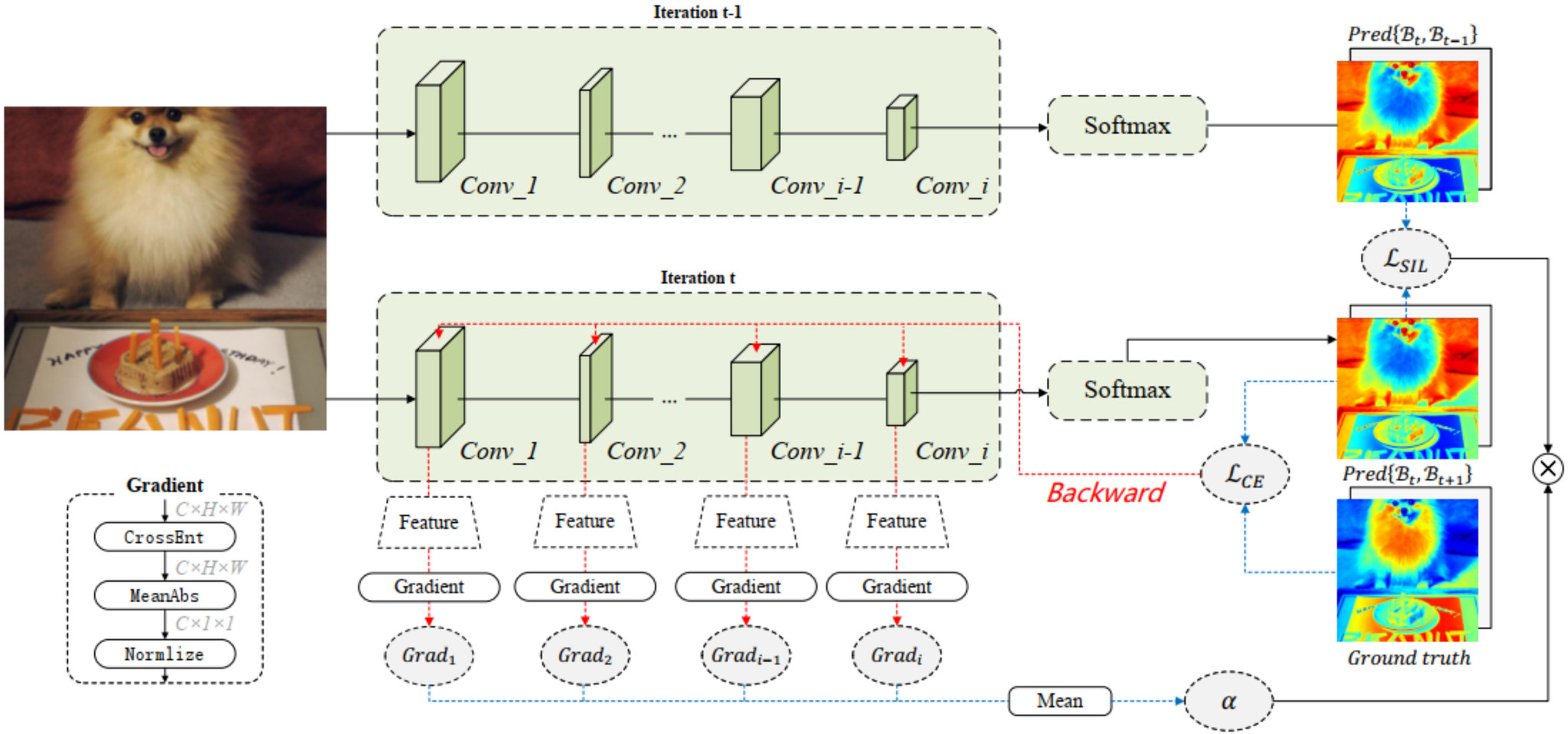
$$\lambda_t = \beta\lambda_{t-1} + (1-\beta)\cdot\frac{|w_h|_1}{N}$$

$$M^l = \text{Z-score}\left(\left|\sum_{k=1}^{M}\hat{w}_k^l\right|\right)$$

- incorporates a dynamic weight α that balances the task and distillation losses.

$$\alpha_t = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}M^l[i,j]\right)$$

Core Idea of SDPGO



SDPGO uses sequential iterative learning and gradient-based feature weighting to generate soft labels.

Experiments

CIFAR Classification Result

Table 1: Top-1 accuracy (%) of various SKD methods across widely used networks on CIFAR-10 (C10) and CIFAR-100 (C100). The best results are highlighted in bold, while the second-best results are underlined. We use Δ to show its performance gain.

Dataset	Methods	Vgg-16	R-32	R-110	WRN-20-8	DNet-40-12	SN-V2	MN-V2
C10	Baseline	93.97	93.46	94.79	94.53	92.91	92.70	93.31
	BYOT [57]	94.03	93.57	94.86	94.14	93.01	92.99	93.73
	EFWSNet [61]	93.85	93.97	94.92	94.68	93.39	93.23	93.88
	PS-KD [20]	94.10	94.04	94.91	95.01	93.23	93.45	94.02
	FRSKD [18]	94.38	94.78	95.23	95.27	94.21	94.17	94.76
	DLB [35]	<u>94.62</u>	94.15	95.15	<u>95.54</u>	<u>93.43</u>	95.10	94.46
	MSKD [50]	93.82	<u>95.59</u>	<u>95.93</u>	93.93	94.25	95.29	<u>94.91</u>
	FASD [47]	94.21	95.45	95.66	94.52	94.39	<u>95.34</u>	94.86
	SDPGO	**95.90**	**96.44**	**95.98**	**95.70**	**95.60**	**95.73**	**95.43**
	Δ	1.93	2.98	1.19	1.17	2.69	3.03	2.12
C100	Baseline	73.63	71.74	76.36	77.58	71.69	71.82	68.08
	BYOT [57]	73.79	72.39	77.75	77.68	77.04	72.97	68.72
	EFWSNet [61]	73.92	73.54	75.81	78.02	76.95	72.87	69.45
	PS-KD [20]	74.05	72.51	77.15	78.74	72.52	74.55	69.74
	FRSKD [18]	<u>76.72</u>	75.34	<u>79.15</u>	78.95	77.12	75.23	70.25
	DLB [35]	76.12	74.07	78.18	<u>79.21</u>	72.52	75.51	69.47
	MSKD [50]	76.57	75.12	78.86	78.07	76.85	76.52	71.66
	FASD [47]	75.52	<u>75.42</u>	78.52	78.62	<u>77.24</u>	<u>76.76</u>	<u>71.75</u>
	SDPGO	**76.85**	**75.57**	**79.31**	**79.36**	**78.04**	**77.29**	**72.25**
	Δ	3.22	3.83	1.95	1.78	6.35	5.47	4.17

Experiments

ImageNet Classification Result & Fine-grained Classification Result

Table 2: Top-1 and Top-5 accuracy (%) of our SDPGO method. ResNet-18 is used as classifier network on the ImageNet dataset.

Methods	ResNet-18		Gain(↑)
	Top-1	Top-5	
ResNet-18	69.75	89.07	-
FitNet [1]	71.61	90.51	1.86
Review [4]	70.81	89.98	1.06
CAT-KD [11]	71.22	90.26	1.47
CRD [41]	71.17	90.13	1.42
SSKD [46]	71.62	90.67	1.87
DCCD [30]	71.95	90.88	2.2
BYOT [57]	69.84	89.62	0.09
EFWSNet [61]	72.36	91.74	2.61
PS-KD [20]	71.59	90.85	1.84
FRSKD [18]	70.17	90.52	0.42
DLB [35]	70.12	90.27	0.37
MSKD [50]	71.67	91.20	1.92
FASD [47]	71.70	90.91	1.95
SDPGO (Ours)	**72.47**	**92.56**	2.72

Table 3: Top-1 accuracy (%) of various self-knowledge distillation methods across widely used networks on CUB200 and Cars196 dataset.

Method	CUB200		Cars196	
	Top-1	Top-5	Top-1	Top-5
ResNet-18	69.66	91.66	71.82	91.04
+ BYOT [57]	73.38	91.18	79.35	94.70
+ DLB [35]	76.10	93.37	78.28	93.13
+ MSKD [50]	71.11	91.37	82.94	95.83
+ SDPGO	**78.06**	**94.77**	**84.17**	**96.44**
ResNet-50	74.36	92.52	76.44	92.26
+ BYOT [57]	77.76	94.22	80.17	94.78
+ DLB [35]	80.69	95.62	82.94	95.83
+ MSKD [50]	75.96	93.03	77.90	93.55
+ SDPGO	**81.69**	**95.88**	**89.20**	**98.16**
MobNet-V2	73.09	91.82	72.15	93.01
+ BYOT [57]	74.25	91.92	82.51	95.03
+ DLB [35]	78.08	94.32	81.72	92.93
+ MSKD [50]	73.65	91.23	82.69	94.41
+ SDPGO	**78.67**	**94.75**	**85.28**	**96.69**

Experiments

ImageNet Classification Result & Fine-grained Classification Result

Table 4: The downstream tasks results with SDPGO on the COCO-2017 dataset.

Methods	ResNet-18		ResNet-50	
	bbox	segm	bbox	segm
baseline	33.4	30.2	36.9	33.4
MixSKD	33.9	31.05	37.0	33.8
FASD	34.1	30.9	37.3	34.4
SDPGO	**35.08**	**32.10**	**38.08**	**36.69**

Table 5: Results of training more models, including ViT-liked models with SDPGO on ImageNet.

Model	Baseline	SDPGO (Ours)
ResNet-50	73.56	74.03 (+0.47)
DeiT-Tiny	74.42	75.01 (+0.59)
DeiT-small	80.55	80.89 (+0.34)
Swin-Tiny	81.18	81.95 (+0.77)
Swin-small	84.36	86.24 (+1.88)

Table 6: Overall Performance Comparison on Semantic Segmentation Tasks.

Model	Method	ADE20K	Cityscapes
ResNet-50	Baseline	39.72	74.85
	MixSKD	42.37	74.96
	FASD	40.78	72.89
	SDPGO	42.75	75.01

Table 7: Performance comparison on Pascal VOC segmentation task.

Model	Method	mIOU	Model	Method	mIOU
EfficientDet-d0	Baseline	79.07	EfficientDet-d1	Baseline	81.95
	MixSKD	79.52		MixSKD	82.51
	FASD	80.54		FASD	83.43
	SDPGO	80.67		SDPGO	83.97

Efficiency and expenditure

Robustness Analysis & Training time

Table 8: Top-1 Acc (%) and Training time for each batch of data of competitive KD methods.

Method	Baseline	ReviewKD	CATKD	BYoT	PS-KD	DLB	FASD	Ours
Training Time	12	25	17	19	15	17	40	12
Top-1 Acc (%)	69.66	76.58	72.6	76.10	75.43	78.06	75.43	78.06

Table 9: Performance Comparison when a fraction of training data is noisy.

η	BYOT	PS-KD	DLB	FASD	Our
0	72.39	72.51	74.07	75.42	75.57
10	65.41	62.75	67.56	64.25	71.56
20	58.05	57.56	65.17	60.51	68.53
30	53.25	51.71	54.45	57.27	61.58
40	42.22	41.26	51.25	55.39	59.81

Table 10: Performance Comparison between SOTA SKD and SDPGO when a fraction of data present for training purpose.

F	BYOT	PS-KD	DLB	FASD	Our
25	49.57	48.75	51.28	60.34	65.29
50	58.25	56.23	59.56	68.62	70.07
75	63.43	60.58	68.12	70.47	71.58
100	72.39	72.51	74.07	75.42	75.57

Summary

- Proposed a self-distillation framework based on gradient-driven feature importance evaluation.
- No teacher model or auxiliary architecture required.
- Achieves consistent improvements across CIFAR, ImageNet, and fine-grained benchmarks.
- Faster inference and lower memory consumption, suitable for deployment.