Model Compression

Yuxuan Zhang

Table of contents

01. Summary

Challenges of Knowledge Distillation

- Performance gap between the teacher model and the student model
- The teacher model is not strong enough
- It is difficult to transfer knowledge without information loss

How EfficientKD Solves the Problems?

Self-Supervised Learning

Make the teacher model strong enough

Reused Teacher Classifier

Minimize the information loss during the knowledge transfer

Abstract

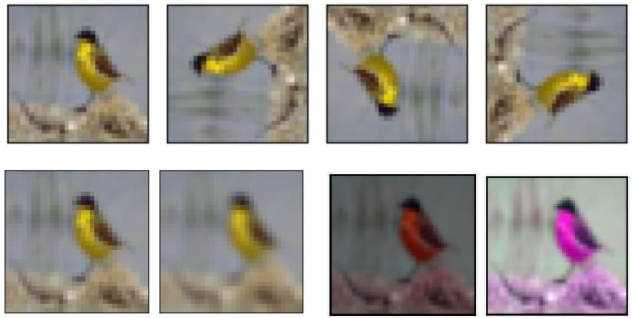
Knowledge distillation aims to transfer knowledge from a powerful teacher model to a lightweight student model. To enhance the student model's performance, multitudinous methods have been introduced with intricate distillation loss functions. However, these approaches suffer from transmission loss during knowledge transfer and suboptimal teacher model performance, thus resulting in insufficient performance. To address these two problems, we propose EfficientKD, a novel and efficient knowledge distillation approach. EfficientKD employs a self-supervised learning strategy, augmented with eight auxiliary decoder branches, to bolster the teacher model. Furthermore, by reusing the teacher's decoder head and training the student model's encoder via feature alignment with an optional feature projector, EfficientKD significantly reduces knowledge loss during distillation. Notably, our method is also adaptable for semantic segmentation tasks. Comprehensive experiments substantiate the superiority of our proposed method. For instance, using a ResNet32x4 and ResNet8x4 teacher-student pairing on CIFAR-100, EfficientKD achieves a classification accuracy of **79.91%**, attaining a **1.83%** improvement over leading methods like SimKD (CVPR 2022), DKD (CVPR 2022), and DIST (NeurIPS 2022).

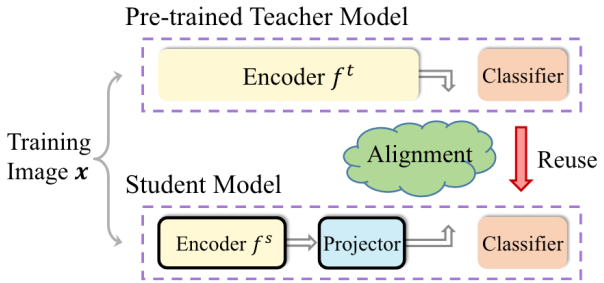
Contributions

- Image classification
  - Combine self-supervised learning with reused classifier
  - Achieve **1.83%** better classification accuracy than SOTA methods (These include SimKD (CVPR 2022), DKD (CVPR 2022), and DIST (NeurIPS 2022))
- Semantic Segmentation
  - Propose a method using the teacher model's segmentation head
  - Achieve about **0.5%** mIoU increase compared with training from scratch

EfficientKD Framework

- Add eight auxiliary classification branches
- Employ reused teacher classifier
- Achieve **2%** better classification accuracy than SOTA methods





Knowledge Distillation Benchmark

Rank	Model	Top-1 Accuracy (%)	Paper	Code	Result	Year	Tags
1	**resnet8x4** (T: resnet32x4 S: resnet8x4 [modified])	78.08	Knowledge Distillation with the Reused Teacher Classifier			2022	
2	**ReviewKD++** (T:resnet-32x4, S:shufflenet-v2)	77.93	Improving Knowledge Distillation via Regularizing Feature Norm and Direction			2023	
3	**ReviewKD++** (T:resnet-32x4, S:shufflenet-v1)	77.68	Improving Knowledge Distillation via Regularizing Feature Norm and Direction			2023	
4	**resnet8x4** (T: resnet32x4 S: resnet8x4)	76.68	Information Theoretic Representation Distillation			2021	
5	**resnet8x4** (T: resnet32x4 S: resnet8x4)	76.31	Knowledge Distillation from A Stronger Teacher			2022	
6	**DKD++** (T:resnet-32x4, S:resnet-8x4)	76.28	Improving Knowledge Distillation via Regularizing Feature Norm and Direction			2023	
7	**resnet8x4** (T: resnet32x4 S: resnet8x4)	76.15	Wasserstein Contrastive Representation Distillation			2020	

Ours: **79.91** (T: resnet32x4 S: resnet8x4)

Model Compression Pipeline

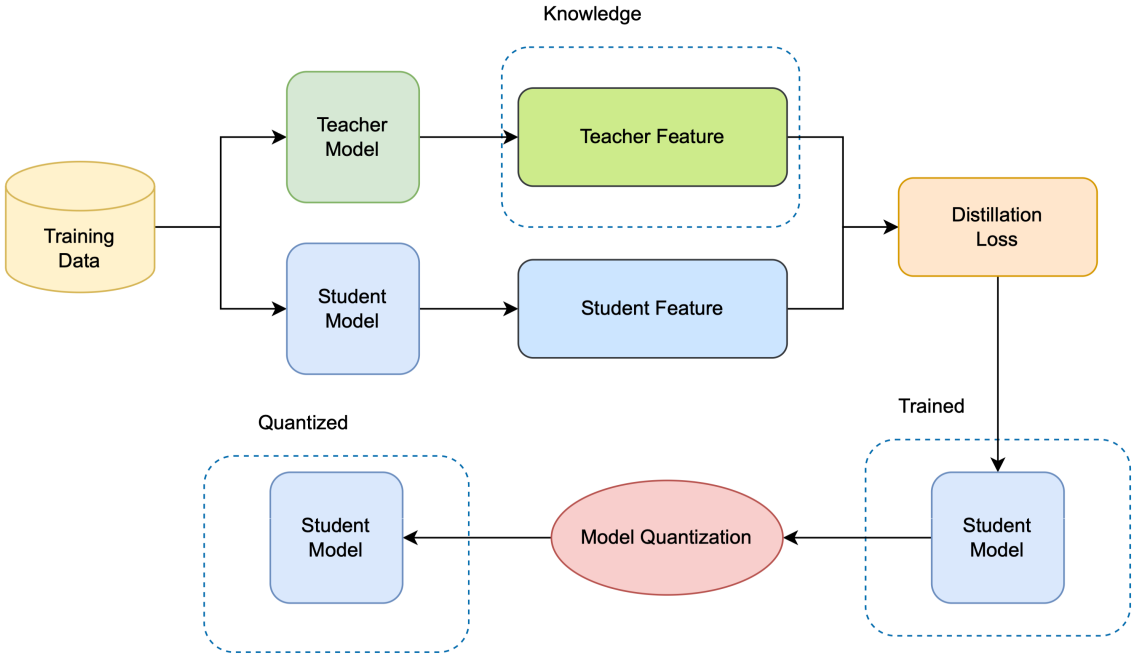
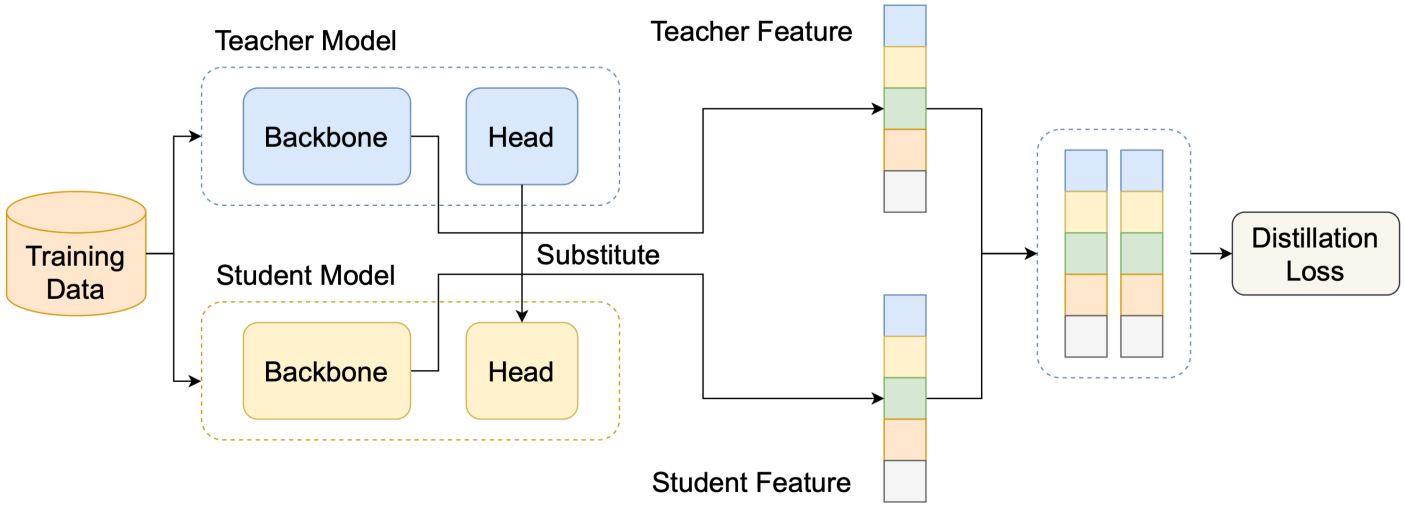


Illustration of Feature Alignment Knowledge Distillation



Summary of Model Compression

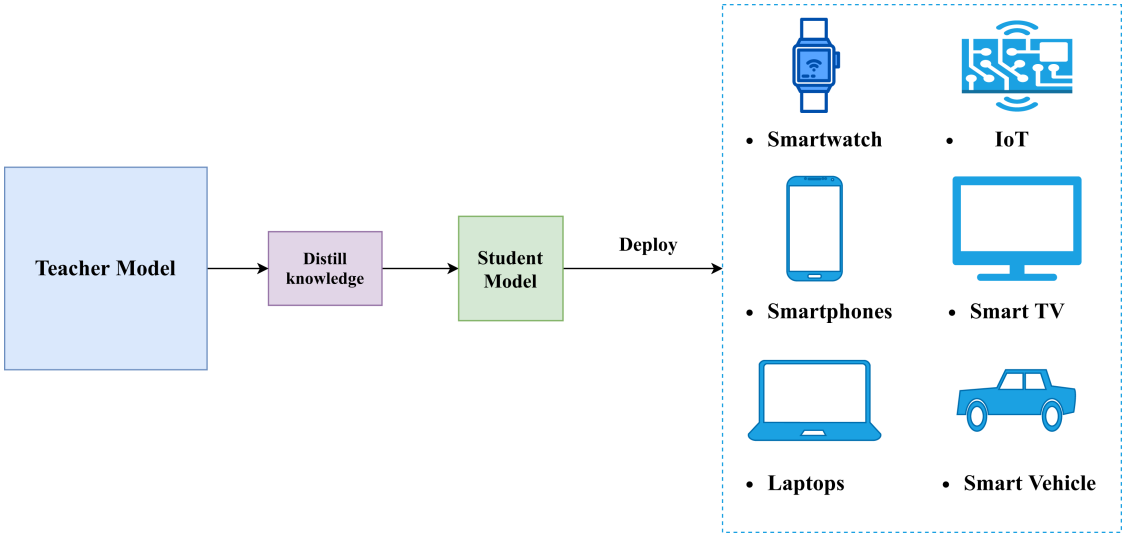
- Implement EfficientViT with model compression techniques for faster semantic segmentation with better performance

	Previous Model (SegFormer)	Current Model (EfficientViT with Model Compression and Inference Optimization)	Δ
mIoU	46.51	50.48	**+ 8.5%**
Latency (s) (per frame)	0.0927	0.0310	**- 66.56%**
FPS	10.79	32.26	**+ 198.98%**

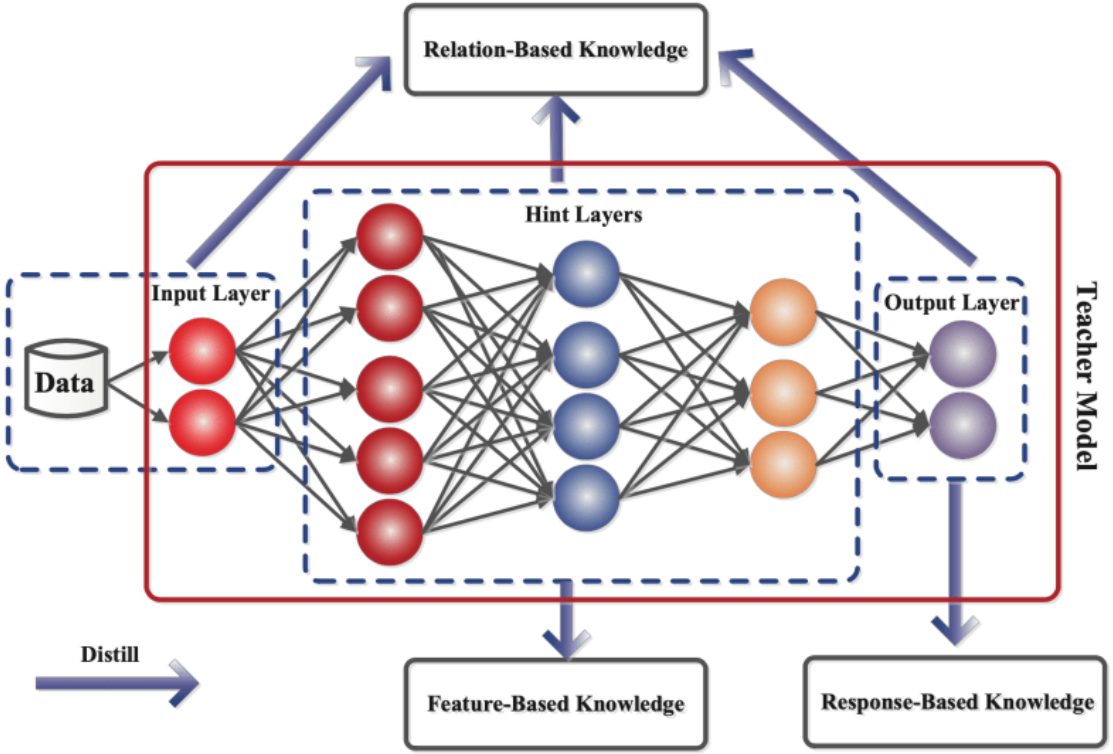
02. Introduction

Knowledge Distillation Objective

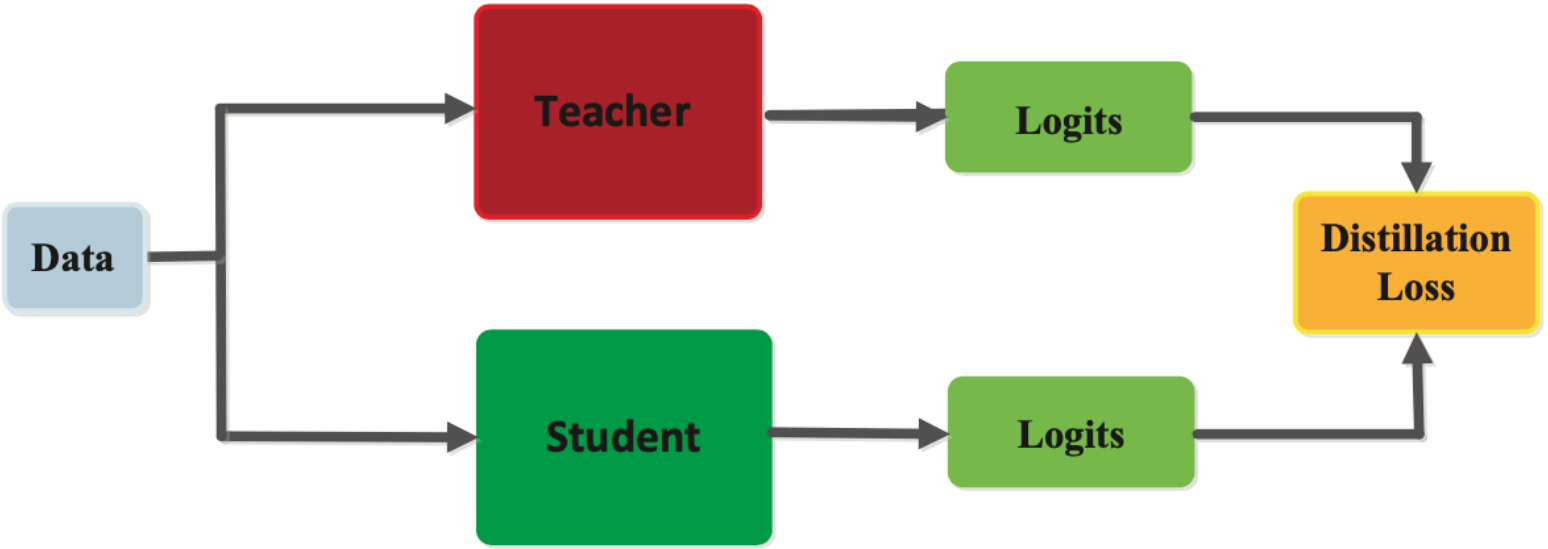
- Compress a powerful yet cumbersome teacher model into a lightweight student model without much sacrifice of performance
- Narrow down the gap between the teacher and the student model



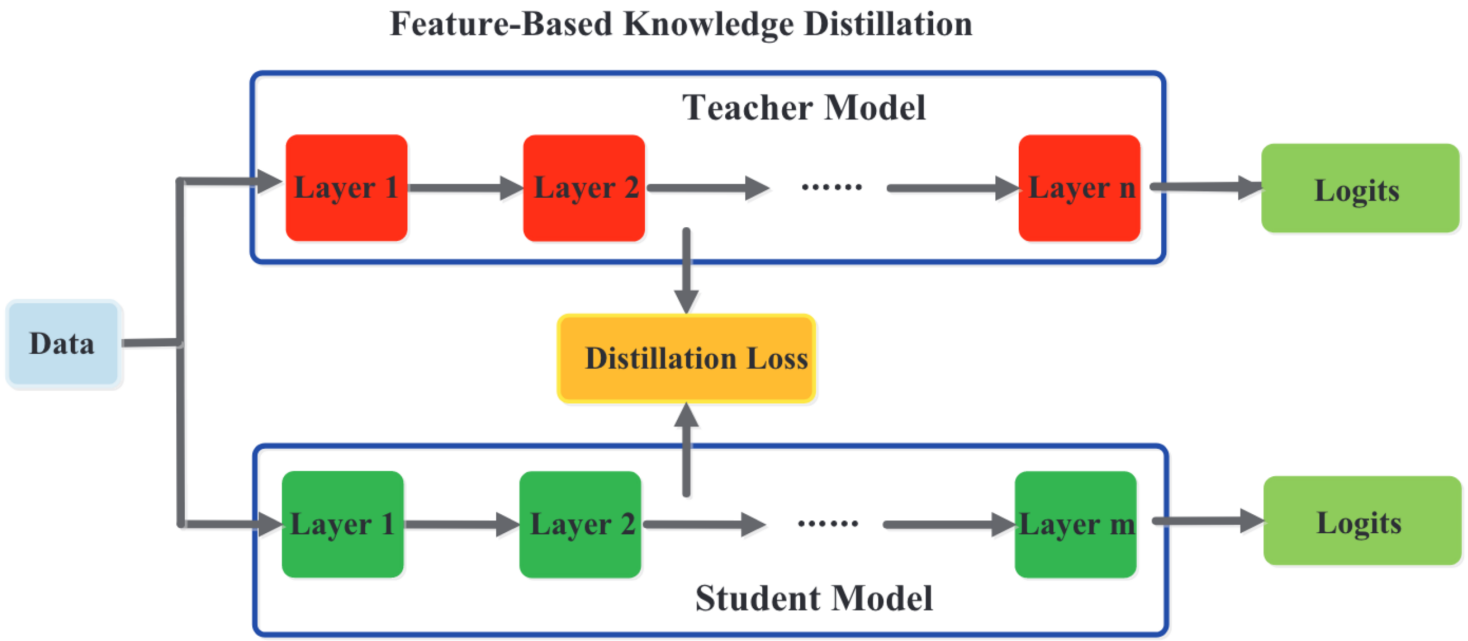
Different Kinds of Knowledge



Logit-based Knowledge Distillation



Feature-based Knowledge Distillation



KD Loss Function

$$D_{\mathrm{KL}}\left(y_{\mathrm{pred}}, y_{\mathrm{true}}\right) = y_{\mathrm{true}} \cdot \log \frac{y_{\mathrm{true}}}{y_{\mathrm{pred}}}$$

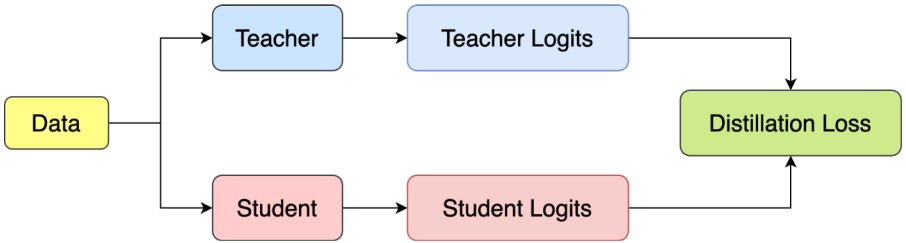
$$l_n = -w_{y_n} \log \frac{\exp\left(x_{n,y_n}\right)}{\sum_{c=1}^{C} \exp\left(x_{n,c}\right)}$$

$$\ell(x, y) = \sum_{n=1}^{N} \frac{1}{\sum_{n=1}^{N} w_{y_n}} l_n$$

$$\mathcal{L} = (1-\alpha)\ell(s\_logits, target) + \alpha D_{\mathrm{KL}}(P, Q)$$

$$P = \mathrm{LogSoftmax}\left(\frac{s\_logits}{T}\right)$$

$$Q = \mathrm{Softmax}\left(\frac{t\_logits}{T}\right)$$

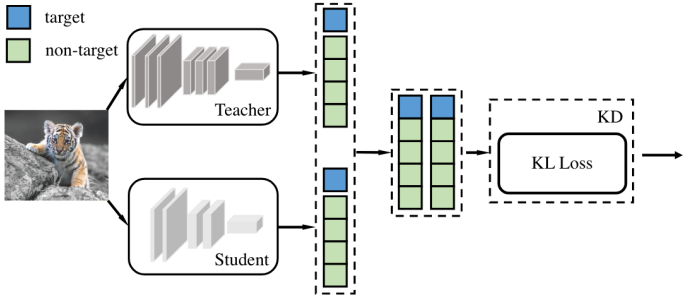


03. Related Work

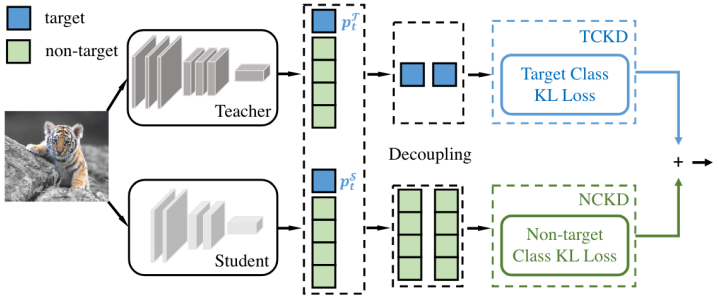
Decoupled Knowledge Distillation

student	TCKD	NCKD	top-1	Δ
ResNet32×4 as the teacher				
ResNet8×4			72.50	-
	✓	✓	73.63	+1.13
	✓		68.63	-3.87
		✓	74.26	+1.76
ShuffleNet-V1			70.50	-
	✓	✓	74.29	+3.79
	✓		70.52	+0.02
		✓	74.91	+4.41
WRN-40-2 as the teacher				
WRN-16-2			73.26	-
	✓	✓	74.96	+1.70
	✓		70.96	-2.30
		✓	74.76	+1.50
ShuffleNet-V1			70.50	-
	✓	✓	74.92	+4.42
	✓		70.62	+0.12
		✓	75.12	+4.62

Table 1. Accuracy(%) on the CIFAR-100 validation set. Δ represents the performance improvement over the baseline.



(a) Classical Knowledge Distillation (KD).

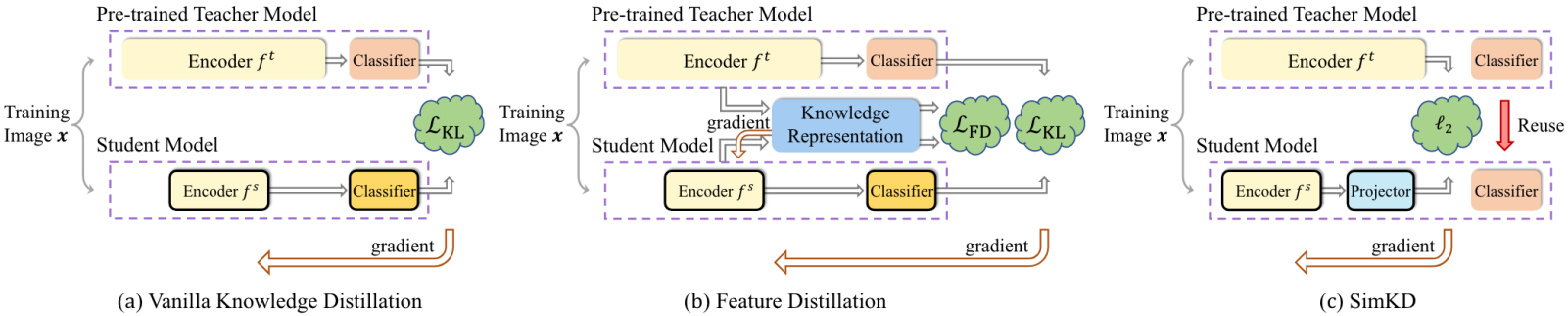


$$\textbf{Classical KD} = TCKD + (1 - p_t^T) * NCKD$$

$$\textbf{DKD(Ours)} = \alpha * TCKD + \beta * NCKD$$

KD with the Reused Teacher Classifier

- Simply reuse the teacher network's classifier
- Add a projector before the classifier to align the input shape of the classifier
- Employ L2 loss to force the input of the classifier to be the same as the teacher

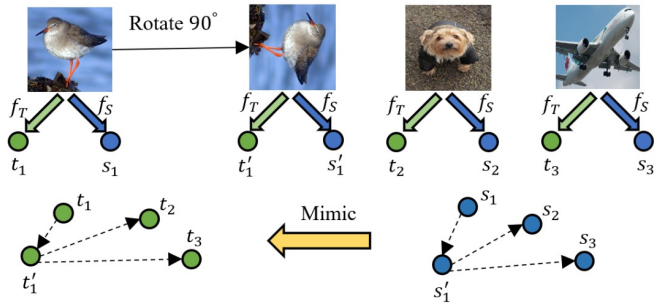


KD with the Reused Teacher Classifier

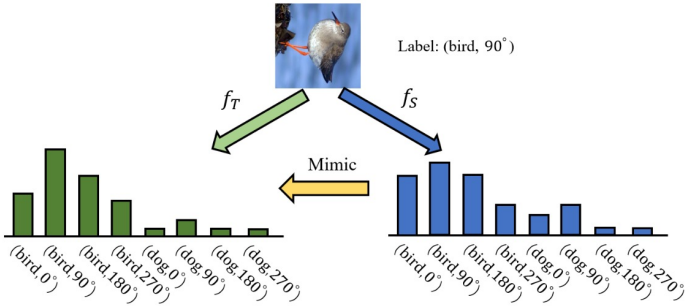
Student	WRN-40-1	ResNet-8x4	ResNet-110	ResNet-116	VGG-8	ResNet-8x4	ShuffleNetV2
	71.92 ± 0.17	73.09 ± 0.30	74.37 ± 0.17	74.46 ± 0.09	70.46 ± 0.29	73.09 ± 0.30	72.60 ± 0.12
KD [24]	74.12 ± 0.29	74.42 ± 0.05	76.25 ± 0.34	76.14 ± 0.32	72.73 ± 0.15	75.28 ± 0.18	75.60 ± 0.21
FitNet [39]	74.17 ± 0.22	74.32 ± 0.08	76.08 ± 0.13	76.20 ± 0.17	72.91 ± 0.18	75.02 ± 0.31	75.82 ± 0.22
AT [53]	74.67 ± 0.18	75.07 ± 0.03	76.67 ± 0.28	76.84 ± 0.25	71.90 ± 0.13	75.74 ± 0.09	75.41 ± 0.10
SP [46]	73.90 ± 0.17	74.29 ± 0.07	76.43 ± 0.39	75.99 ± 0.26	73.12 ± 0.10	74.84 ± 0.08	75.77 ± 0.08
VID [1]	74.59 ± 0.17	74.55 ± 0.10	76.17 ± 0.22	76.53 ± 0.24	73.19 ± 0.23	75.56 ± 0.13	75.22 ± 0.07
CRD [44]	74.80 ± 0.33	75.59 ± 0.07	76.86 ± 0.09	76.83 ± 0.13	73.54 ± 0.19	75.78 ± 0.27	77.04 ± 0.61
SRRL [50]	74.64 ± 0.14	75.39 ± 0.34	76.75 ± 0.14	77.19 ± 0.09	73.23 ± 0.16	76.12 ± 0.18	76.19 ± 0.35
SemCKD [6]	74.41 ± 0.16	76.23 ± 0.04	76.62 ± 0.14	76.69 ± 0.48	75.27 ± 0.13	75.85 ± 0.16	77.62 ± 0.32
SimKD	**75.56 ± 0.27**	**78.08 ± 0.15**	**77.82 ± 0.15**	**77.90 ± 0.11**	**75.76 ± 0.12**	**76.75 ± 0.23**	**78.39 ± 0.27**
Teacher	WRN-40-2	ResNet-32x4	ResNet-110x2	ResNet-110x2	ResNet-32x4	WRN-40-2	ResNet-32x4
	76.31	79.42	78.18	78.18	79.42	76.31	79.42

Table 1. Top-1 test accuracy (%) of various knowledge distillation approaches on CIFAR-100.

Hierarchical Self-supervised Augmented Knowledge Distillation



(a) Self-supervised contrastive relationship [Xu *et al.*, 2020].



Hierarchical Self-supervised Augmented Knowledge Distillation

- Mimicry loss is used to make the student network learn from the teacher network's response to pretext task

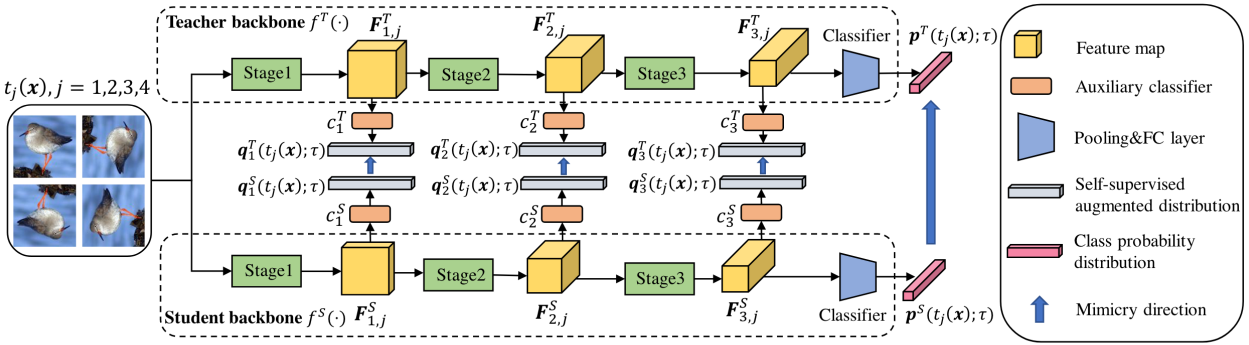


Figure 2: Overview of our proposed HSAKD. Both teacher and student networks are equipped with several auxiliary classifiers after various convolutional stages to capture diverse self-supervised augmented knowledge from hierarchical feature maps. Mimicry loss is applied from self-supervised augmented distributions of the student $\{\boldsymbol{q}_l^S(t_j(\boldsymbol{x});\tau)\}_{l=1}^L$ to corresponding that of the teacher $\{\boldsymbol{q}_l^T(t_j(\boldsymbol{x});\tau)\}_{l=1}^L$ generated from same feature hierarchies in a one-to-one manner. Following the conventional KD, we also force the mimicry from the class probability distribution of student $\boldsymbol{p}^S(t_j(\boldsymbol{x});\tau)$ to that of the teacher $\boldsymbol{p}^T(t_j(\boldsymbol{x});\tau)$. During the inference period, we only retain the student backbone $f^S(\cdot)$ and drop all auxiliary classifiers $\{c^S(\cdot)\}_{l=1}^L$. Therefore there has no extra inference cost compared with the original student network.

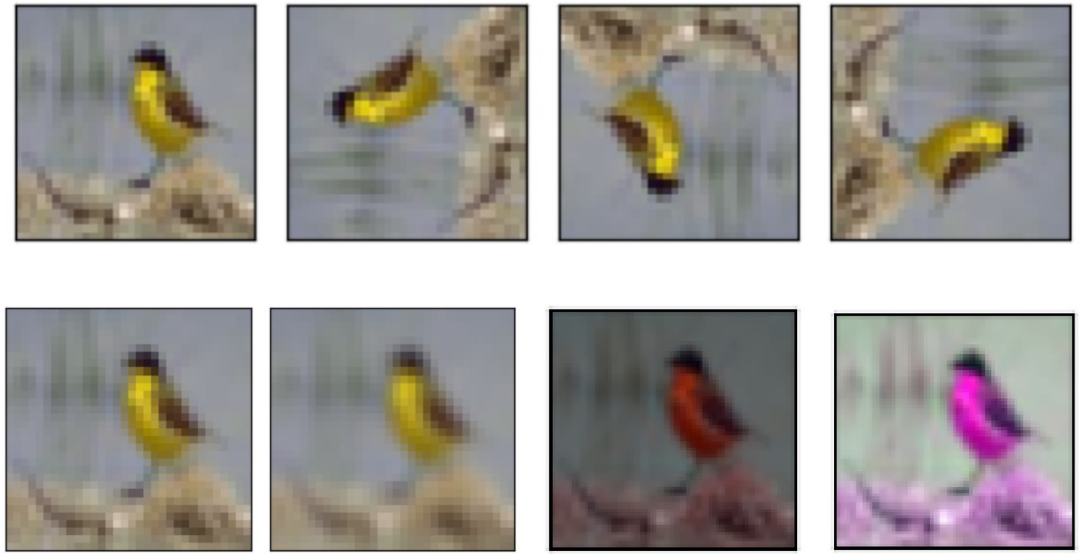
Hierarchical Self-supervised Augmented Knowledge Distillation

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet56 ResNet20	ResNet32×4 ResNet8×4	VGG13 MobileNetV2	ResNet50 MobileNetV2	WRN-40-2 ShuffleNetV1	ResNet32×4 ShuffleNetV2
Teacher	76.45	76.45	73.44	79.63	74.64	79.34	76.45	79.63
Teacher*	80.70	80.70	77.20	83.73	78.48	83.85	80.70	83.73
Student	73.57 (±0.23)	71.95 (±0.59)	69.62 (±0.26)	72.95 (±0.24)	73.51 (±0.26)	73.51 (±0.26)	71.74 (±0.35)	72.96 (±0.33)
KD	75.23 (±0.23)	73.90 (±0.44)	70.91 (±0.10)	73.54 (±0.26)	75.21 (±0.24)	75.80 (±0.46)	75.83 (±0.18)	75.43 (±0.33)
FitNet	75.30 (±0.42)	74.30 (±0.42)	71.21 (±0.16)	75.37 (±0.12)	75.42 (±0.34)	75.41 (±0.07)	76.27 (±0.18)	76.91 (±0.06)
AT	75.64 (±0.31)	74.32 (±0.23)	71.35 (±0.09)	75.06 (±0.19)	74.08 (±0.21)	76.57 (±0.20)	76.51 (±0.44)	76.32 (±0.12)
AB	71.26 (±1.32)	74.55 (±0.46)	71.56 (±0.19)	74.31 (±0.09)	74.98 (±0.44)	75.87 (±0.39)	76.43 (±0.09)	76.40 (±0.29)
VID	75.31 (±0.22)	74.23 (±0.28)	71.35 (±0.09)	75.07 (±0.35)	75.67 (±0.13)	75.97 (±0.08)	76.24 (±0.44)	75.98 (±0.41)
RKD	75.33 (±0.14)	73.90 (±0.26)	71.67 (±0.08)	74.17 (±0.22)	75.54 (±0.36)	76.20 (±0.06)	75.74 (±0.32)	75.42 (±0.25)
SP	74.35 (±0.59)	72.91 (±0.24)	71.45 (±0.38)	75.44 (±0.11)	75.68 (±0.35)	76.35 (±0.14)	76.40 (±0.37)	76.43 (±0.21)
CC	75.30 (±0.03)	74.46 (±0.05)	71.44 (±0.10)	74.40 (±0.24)	75.66 (±0.33)	76.05 (±0.25)	75.63 (±0.30)	75.74 (±0.18)
CRD	75.81 (±0.33)	74.76 (±0.25)	71.83 (±0.42)	75.77 (±0.24)	76.13 (±0.16)	76.89 (±0.27)	76.37 (±0.23)	76.51 (±0.09)
SSKD	76.16 (±0.17)	75.84 (±0.04)	70.80 (±0.02)	75.83 (±0.29)	76.21 (±0.16)	78.21 (±0.16)	76.71 (±0.31)	77.64 (±0.24)
Ours	77.20 (±0.17)	77.00 (±0.21)	72.58 (±0.33)	77.26 (±0.14)	77.45 (±0.21)	78.79 (±0.11)	78.51 (±0.20)	79.93 (±0.11)
Ours*	**78.67** (±0.20)	**78.12** (±0.25)	**73.73** (±0.10)	**77.69** (±0.05)	**79.27** (±0.12)	**79.43** (±0.24)	**80.11** (±0.32)	**80.86** (±0.15)

04. Methodology

SSL Auxiliary Classifier

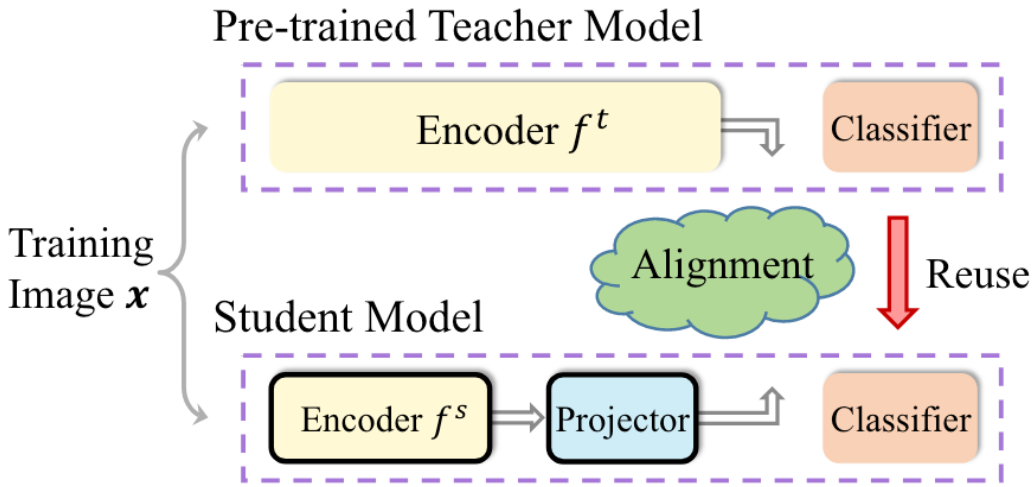
- Eight auxiliary classification branches



Pearson Correlation Loss

- Employ Pearson Correlation of the student logits and the teacher logits to supervise the student training
- Inter-class information: A cat is similar to a dog, but very different from a boat
- Intra-class information: The output of the teacher network that belongs to the same class, the teacher network has three outputs c, d, and b
- Relation c > d > b is also important information.
- An image of a cat: c
- An image of a dog: d
- An image of a boat: b

Reused Classifier



05. Experiments

Image Classification

Semantic Segmentation

	mIoU	Latency (s) (1000 frames)	Params (M)
SegFormer B2	46.51	26.6	27.48
SegFormer B4	50.29	43.98	64.13
EfficientViT L1	48.94	18.08	40.41
EfficientViT L2	**50.48**	**20.1**	**51.46**

Semantic Segmentation (ROS Environment)

	Previous Model (SegFormer)	Current Model (EfficientViT with Model Compression and Inference Optimization)	Δ
mIoU	46.51	50.48	**+ 8.5%**
Latency (s) (per frame)	0.0927	0.0310	**- 66.56%**
FPS	10.79	32.26	**+ 198.98%**

06. Conclusion

Conclusion

- The proposed EfficientKD approach can significantly improve the performance of the student model by mitigating information transfer loss during knowledge distillation
- By leveraging multiple model compression and inference optimization techniques, including knowledge distillation, model quantization, lightweight model design and JIT compilation, we can accelerate the semantic segmentation model's inference speed by 198.98%

Thanks a lot!