Knowledge Diffusion for Distillation

Tao Huang[1], Yuan Zhang[3], Mingkai Zheng[1], Shan You[2], Fei Wang[4], Chen Qian[2], Chang Xu[1]

[1]The University of Sydney [2]SenseTime Research [3]Peking University [4]University of Science and Technology of China

Motivation for Better Knowledge Distillation

Representation Gap in KD

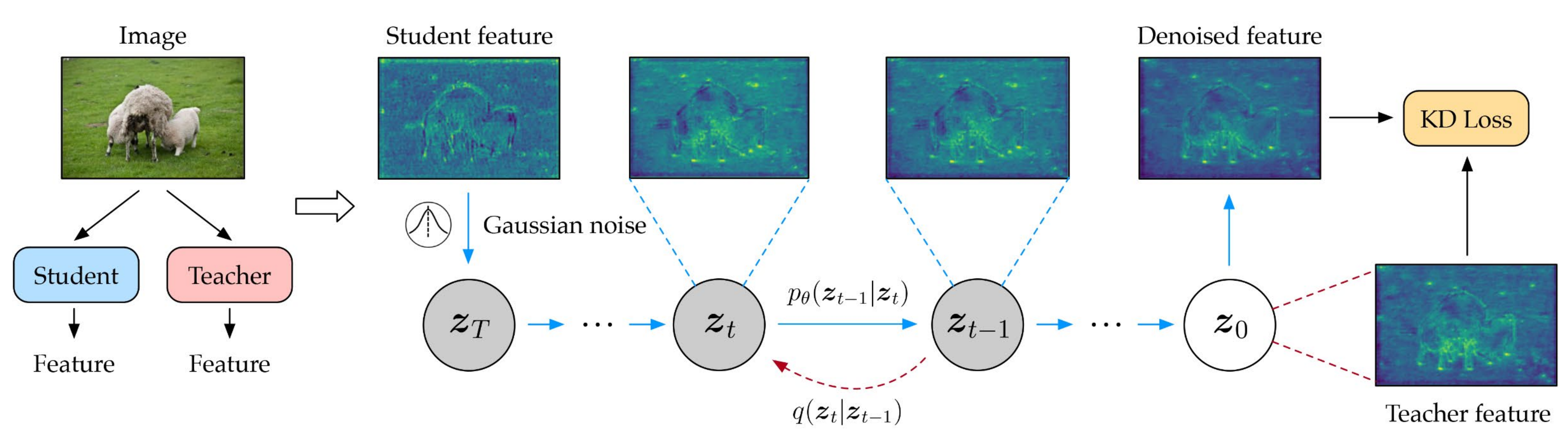
- Challenge of the teacher-student model capacity gap.
- Existing methods are often complex and task-specific.

Noise in Distillation Features

- Student features are noisier due to the limited capacity.
- The noise leads to suboptimal distillation and performance.

⇩

Our DiffKD Approach



- A novel method using diffusion models for denoising the student features.
- Distillation on denoised student features with simple losses such as MSE.

Innovations

- Lightweight diffusion model with linear autoencoder.
- Adaptive noise matching for precise denoising.

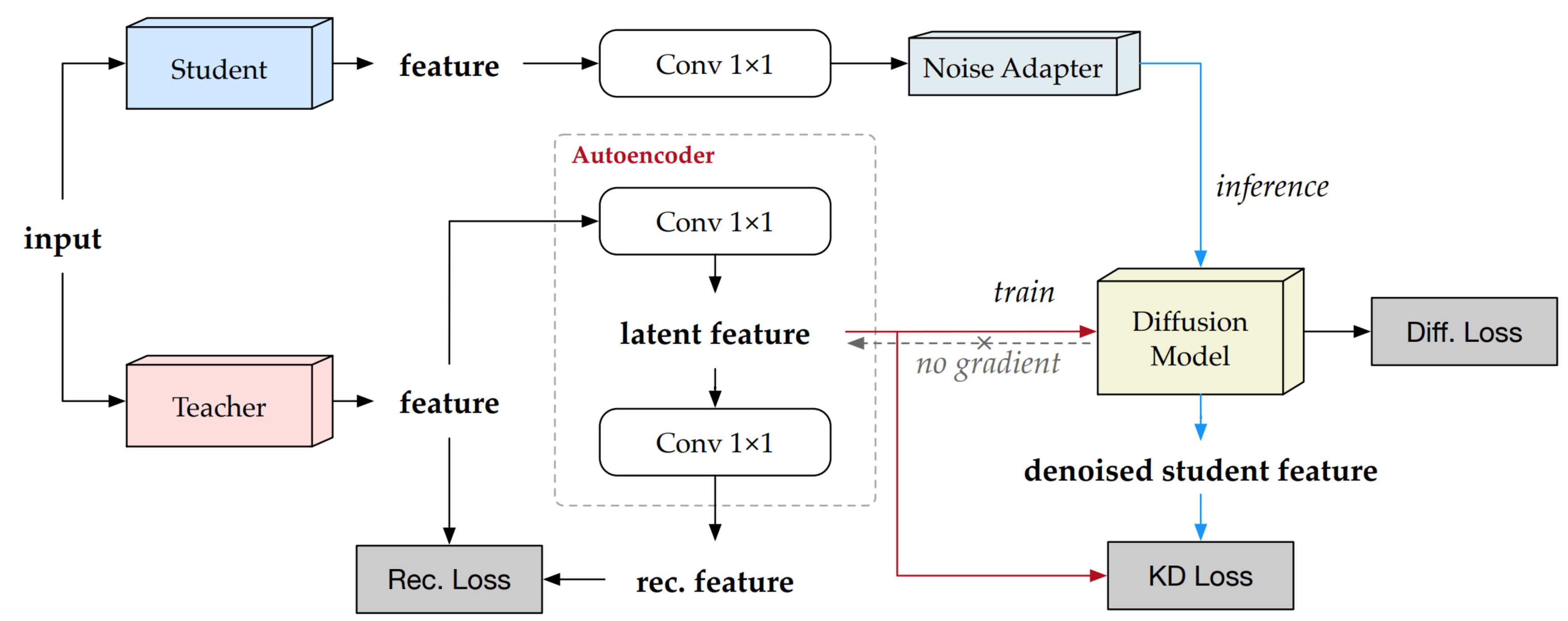
Effectiveness

- Applicable to various feature types.
- Superior performance in multiple tasks and settings.

PAPER

CODE

Method

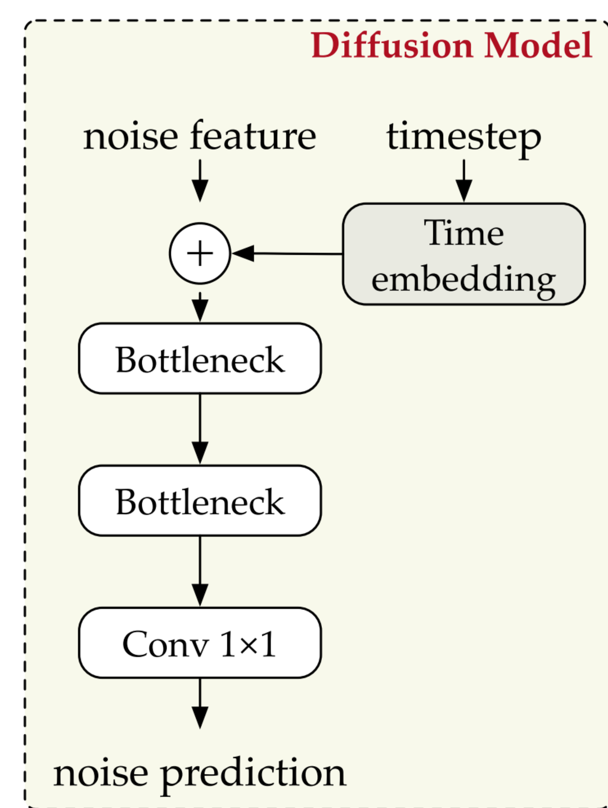


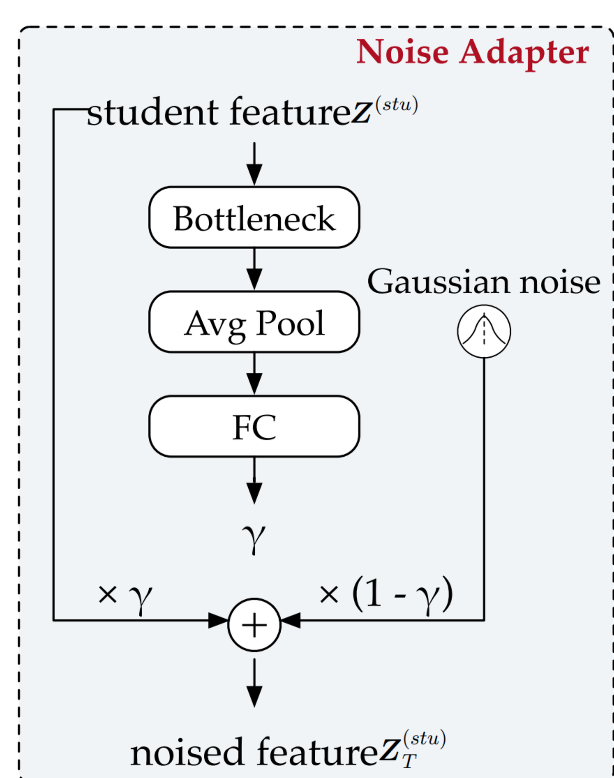
Simultaneous Optimization with

- *Task loss* for training the student.
- *KD loss* for training the student & noise adapter.
- *Diffusion loss* for training the diffusion model.
- *Reconstruction loss* for training the linear autoencoder.

Diffusion Model

- A lightweight model with ResNet Bottleneck blocks.
- Trained with teacher features.
- Leveraged for denoising student features.





Noise Adapter

- Addresses the challenge of inexact noisy levels in student features.
- Measures the noisy level of feature.
- Complements additional Gaussian noise to feature to match the noisy level.

Experiments

ImageNet

†: with advanced DIST loss

Student (teacher)		Tea.	Stu.	KD [13]	Review [6]	DKD [50]	DIST [16]	MSE	DiffKD	DiffKD†
R18 (R34)	Top-1	73.31	69.76	70.66	71.61	71.70	72.07	70.58	72.22	**72.49**
	Top-5	91.42	89.08	89.88	90.51	90.41	90.42	89.95	90.64	**90.71**
MBV1 (R50)	Top-1	76.16	70.13	70.68	72.56	72.05	73.24	72.39	73.62	**73.78**
	Top-5	92.86	89.49	90.30	91.00	91.05	91.12	90.74	91.34	**91.48**

ImageNet with Stronger Teachers

Teacher	Student	Top-1 ACC (%)						
		Tea.	Stu.	KD [13]	RKD [30]	SRRL [46]	DIST [16]	DiffKD
ResNet-50	ResNet-34	80.1	76.8	77.2	76.6	76.7	77.8	**78.1**
	MobileNetV2		73.6	71.7	73.1	69.2	74.4	**74.9**
	EfficientNet-B0		78.0	77.4	77.5	77.3	78.6	**78.8**
Swin-L	ResNet-50	86.3	78.5	80.0	78.9	78.6	80.2	**80.5**
	Swin-T		81.3	81.5	81.2	81.5	82.3	**82.5**

COCO

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage detectors						
T: Faster RCNN-R101	39.8	60.1	43.3	22.5	43.6	52.8
S: Faster RCNN-R50	38.4	59.0	42.0	21.5	42.1	50.3
DiffKD	40.6	60.9	43.9	**23.0**	44.5	**54.0**
T: CM RCNN-X101	45.6	64.1	49.7	26.2	49.6	60.0
S: Faster RCNN-R50	38.4	59.0	42.0	21.5	42.1	50.3
DiffKD	42.2	62.8	46.0	**24.2**	46.6	55.3

Cityscapes

Method	Params (M)	FLOPs (G)	mIoU (%) Val	mIoU (%) Test
T: DeepLabV3-R101	61.1	2371.7	78.07	77.46
S: DeepLabV3-R18	13.6	572.0	74.21	73.45
DiffKD	13.6	572.0	**77.78**	**76.24**
T: DeepLabV3-R101	61.1	2371.7	78.07	77.46
S: PSPNet-R18	12.9	507.4	72.55	72.29
DiffKD	12.9	507.4	**75.83**	**75.61**

Visualizations

Image | Student | Denoised student | Teacher