IFHE: Intermediate-Feature Heterogeneity Enhancement for Image Synthesis in Data-Free Knowledge Distillation

Yi Chen[1], Ning Liu[2], Ao Ren[1], Tao Yang[1], Duo Liu[1]

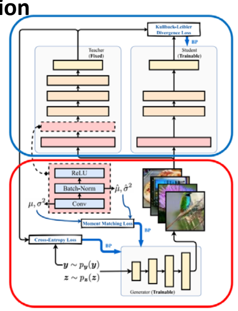
[1]School of Computer Science, Chongqing University, [2]Midea Group

Motivation

❖ Data-Free Knowledge Distillation

Train a compact model without original data

- Safe and private.
- Reduce inference time.
- More accuracy loss than traditional KD.
- Applications: Biometric identification, medical image recognition.



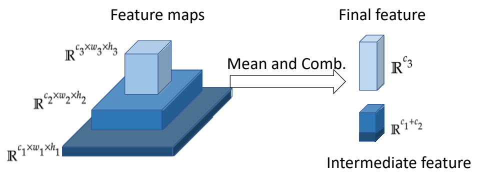
❖ Analysis of accuracy gap

Intermediate features of Synthesized images

- Only constrained on the last layer.
  $\mathcal{L}_{oh}(x) = CE(T(\hat{x}, \theta_t), t)$
- Makes all BNS the same.
  $$\mathcal{L}_{bns}(x) = \frac{1}{N}\sum_{i=0}(|\tilde{\mu}_i - \mu_i|_2^2 + |\tilde{\sigma}_i - \sigma_i|_2^2)$$
- Nothing to do with adversarial loss term.
  $\mathcal{L}_{adv}(x) = -KL(T(\hat{x}, \theta_t)/\tau \,|S(\hat{x}, \theta_s)/\tau)$

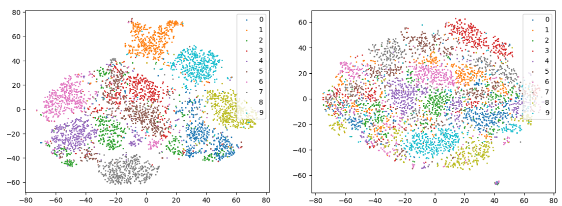
❖ Definition of the intermediate feature

mean and combination of all intermediate feature maps



❖ Analysis of intermediate features

- Real images: Heterogeneous
- Synthesized images: Chaotic



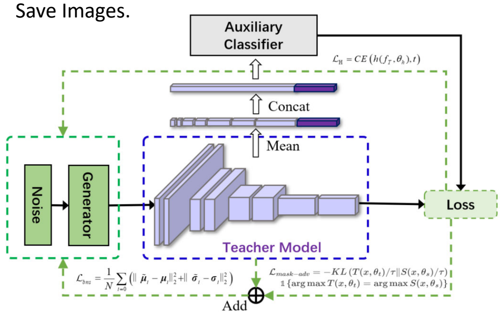
Real images

synthesized images

Method - IFHE

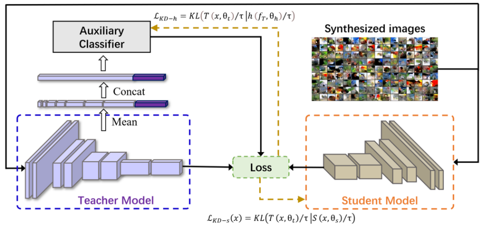
❖Image Synthesis

- Use auxiliary classifiers to predict labels.
- Train noises and generator by loss terms.
- Save Images.



❖Knowledge Distillation

- Sample Images and put them into teacher and student.
- Train both student and auxiliary classifier.



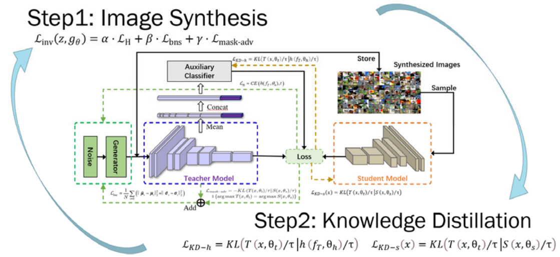
❖ Replace one-hot loss

Curse of $\mathcal{L}_{oh}$: too confident is not beneficial to KD.

- Merge $\mathcal{L}_{oh}$ into feature heterogeneity enhancement.

$$\mathcal{L}_{oh}(x) = CE(T(\hat{x}, \theta_t), t) \longrightarrow \mathcal{L}_H = CE(h(f_T, \theta_h), t)$$

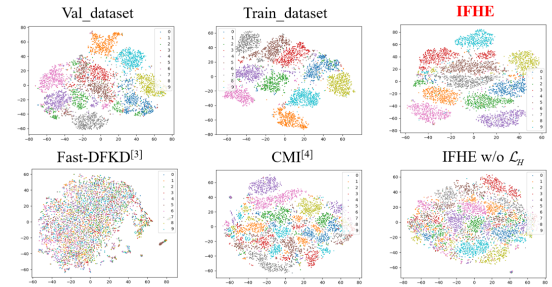
❖ Framework of IFHE-DFKD



Results

❖ Heterogeneity of Intermediate Features

- Intermediate features by IFHE is more discriminative than other DFKD methods.



❖ Experimental results

➢ CIFAR-10

Model	WRN40-2 (%) WRN16-1 (%)	WRN40-2 (%) WRN40-1 (%)	WRN40-2 (%) WRN16-2 (%)	VGG11 (%) ResNet18 (%)	ResNet34 (%) ResNet18 (%)
Teacher	94.87	94.87	94.87	92.25	95.70
Student	91.12	93.94	93.95	95.20	95.20
DAFL[5]	65.71	81.33	81.55	81.10	92.22
ZSKT[6]	83.74	86.07	89.66	89.46	93.32
ADI[7]	83.04	86.85	89.72	90.36	93.26
DFQ[8]	86.14	91.69	92.01	90.84	94.61
CMI[4]	90.01	92.78	92.52	91.13	94.84
Ours	**91.80**	**93.71**	**93.59**	**92.01**	**95.09**

➢ CIFAR-100

Model	WRN40-2(%) WRN16-1(%)	WRN40-2(%) WRN40-1(%)	WRN40-2(%) WRN16-2(%)	VGG11(%) ResNet18(%)	ResNet34(%) ResNet18(%)
Teacher	75.83	75.83	75.83	71.32	78.05
Student	65.31	72.19	73.56	77.10	77.10
DAFL[5]	22.50	34.66	40.00	57.29	74.47
ZSKT[6]	30.15	29.73	28.44	34.72	67.74
ADI[7]	53.77	61.33	61.34	54.13	61.32
DFQ[8]	54.77	62.92	59.01	68.32	77.01
CMI[4]	57.91	68.88	68.75	70.56	77.04
Ours	**61.55**	**69.95**	**70.61**	**70.98**	**77.11**

- Outperform all other DFKD methods
- Wrn40-2/wrn16-1 pair on CIFAR-10 even exceed train scratch with original data

❖ Ablation Study

➢ Different scaling factor of $\mathcal{L}_H$

Dataset	$\alpha=0.0$	$\alpha=0.3$	$\alpha=0.5$	$\alpha=0.8$	$\alpha=1.0$	$\alpha=1.3$	$\alpha=1.5$
CIFAR-10	91.23	91.64	**91.80**	91.61	91.66	91.33	90.91
CIFAR-100	59.66	59.79	60.03	60.54	**61.55**	61.07	61.01

➢ Orthogonality with other generator-based DFKD methods

Method	Acc. (%)	Method	Acc. (%)	Method	Acc. (%)
DAFL[5]	66.43	DFQ[8]	84.86	IFHE w/o L_h	91.23
DAFL + IFHE	**69.18**	DFQ + IFHE	**85.96**	IFHE	**91.80**