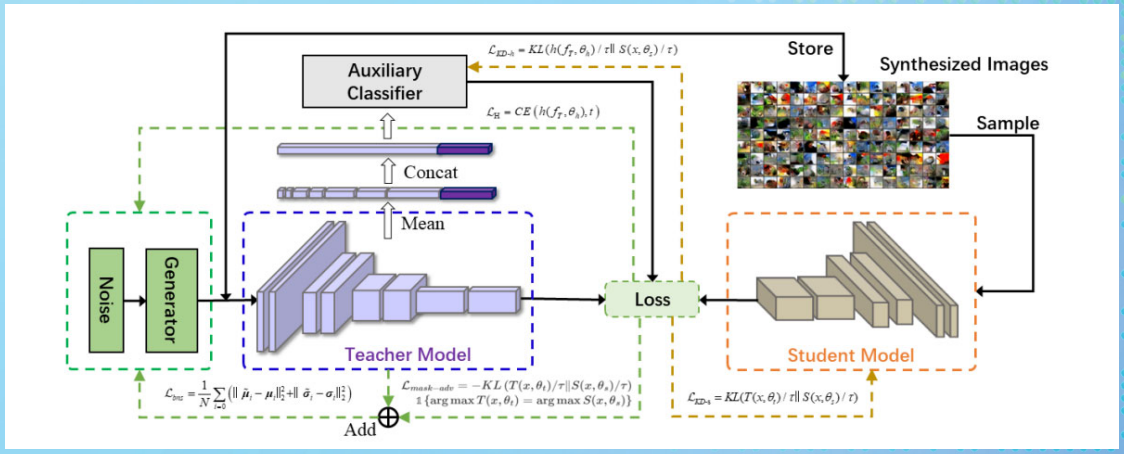

60
WHERE
INNOVATION
BEGINS
DESIGN
AUTOMATION
CONFERENCE
JULY 9–13, 2023
MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

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, Chongqing, China
[2]Midea Group, Beijing, China
*Equal Contributions. ※Corresponding Author.
{chen yi, ren.ao, yangtao, liuduo}@cqu.edu.com, ningliu1220@gmail.com



Outline

1. Background
2. Preliminary
3. Motivation
4. IFHE: Intermediate-Feature Heterogeneity Enhancement
5. Experimental Result
6. Conclusion
7. References

Background: Knowledge Distillation

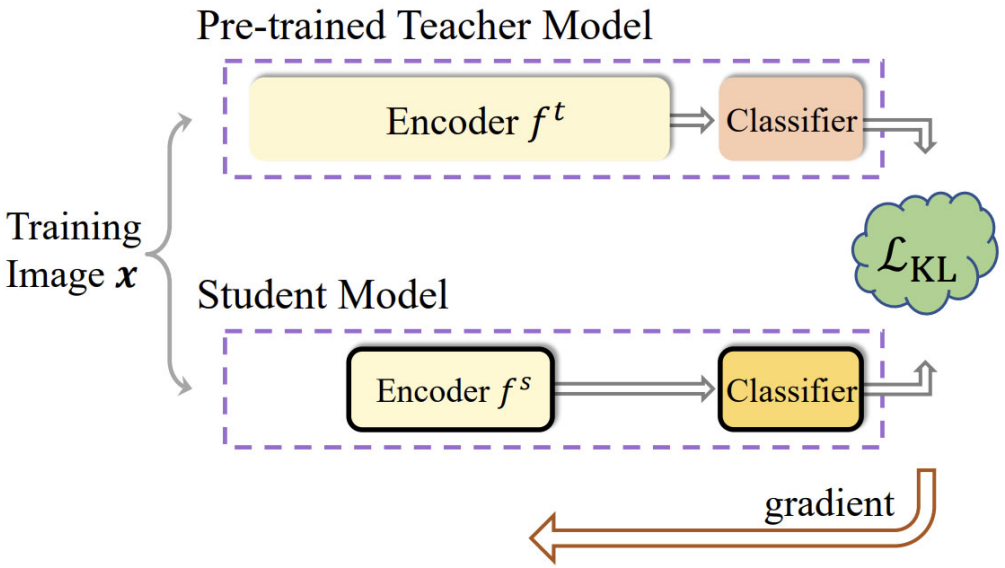
Train a compact and high-performance model by distilling knowledge from a large model

Pros

- Retain a tiny student model.
- Reduce inference time, storage cost.
- Hardware friendly.
- Higher performance than training scratch.

Cons

- Lower compression rate (than pruning and quantization)
- More time-consuming (than simply train from scratch)
- **Data required.**



$$\mathcal{L}_{KD}(x) = KL(T(x, \theta_t)/\tau \| S(x, \theta_s)/\tau)$$

Background: Data-Free Knowledge Distillation

Pros

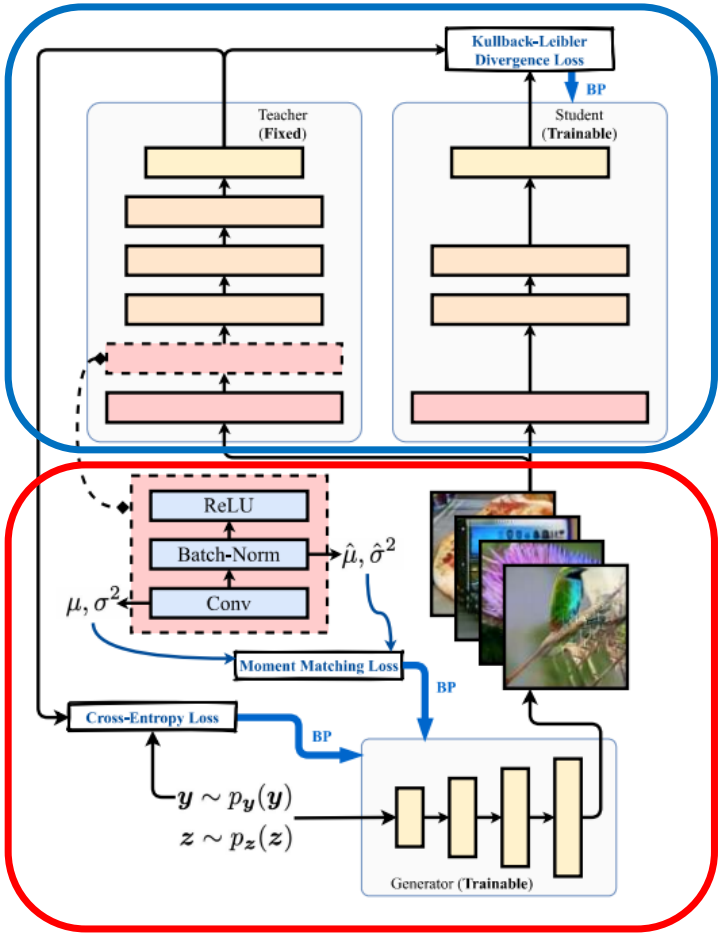
- Almost inherits all the advantages of KD
- Safe and private

Cons compared to traditional KD

- More accuracy loss and lower compression rate
- More time-consuming and resource-consuming

Application situations

- Provide a pretrained model but without data.
- biometric identification and medical image recognization.



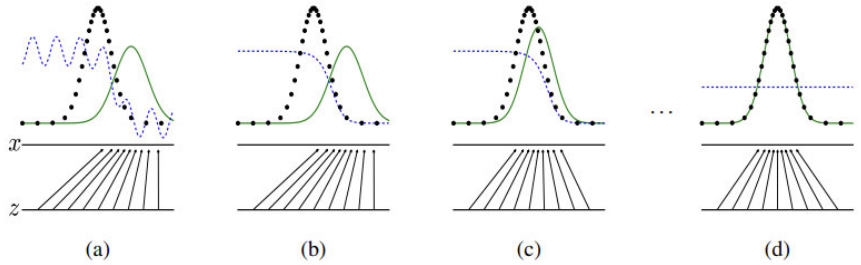
Knowledge Distillation

Image Synthesis

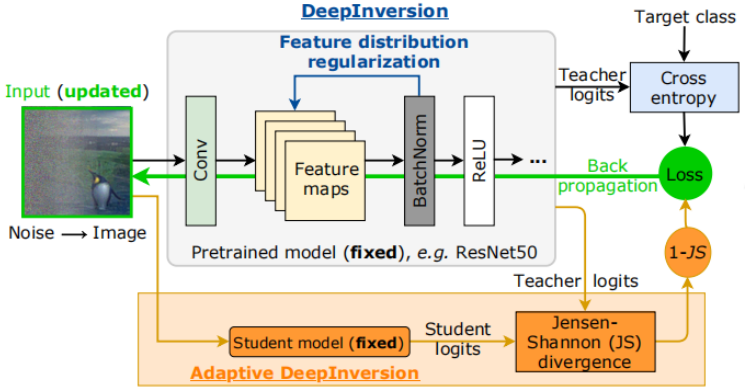
Background: Image Synthesis

- Generative Adversarial Network
  - Adversarial Training by the discriminator and generator
  - Train a generator to generate images
  - Pros: save time and hardware resources
  - Cons: lower performance

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$



- Model Inversion
  - Extract information from a well-trained network.
  - Train noises to images.
  - Pros: higher performance
  - Cons: more time and hardware resources consumed.



Preliminary: synthesize image from the teacher

1. Make the Image Confident.

$$\mathcal{L}_{oh}(x) = CE(T(\hat{x}, \theta_t), t)$$

2. Make the Image Reality.

$$\mathcal{L}_{\text{bns}}(x) = \frac{1}{N}\sum_{i=0}(|\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i|_2^2 + |\tilde{\boldsymbol{\sigma}}_i - \boldsymbol{\sigma}_i|_2^2)$$

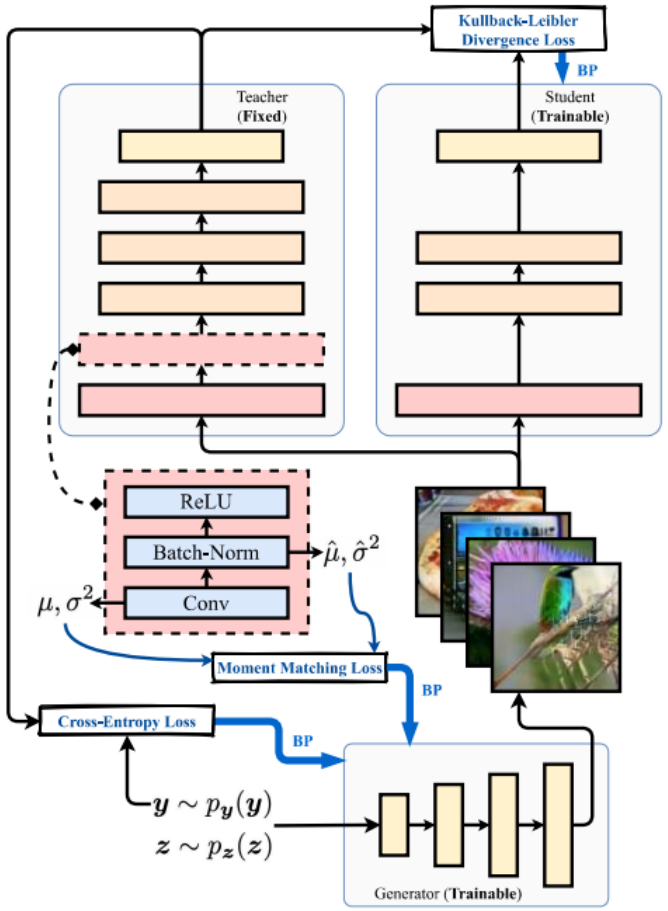
3. Make the Image "Useful".

$$\mathcal{L}_{adv}(x) = -KL(T(\hat{x}, \theta_t)/\tau \,|S(\hat{x}, \theta_s)/\tau)$$

Q: Train generator or noises?

A: Both

Higher performance, less time-consuming



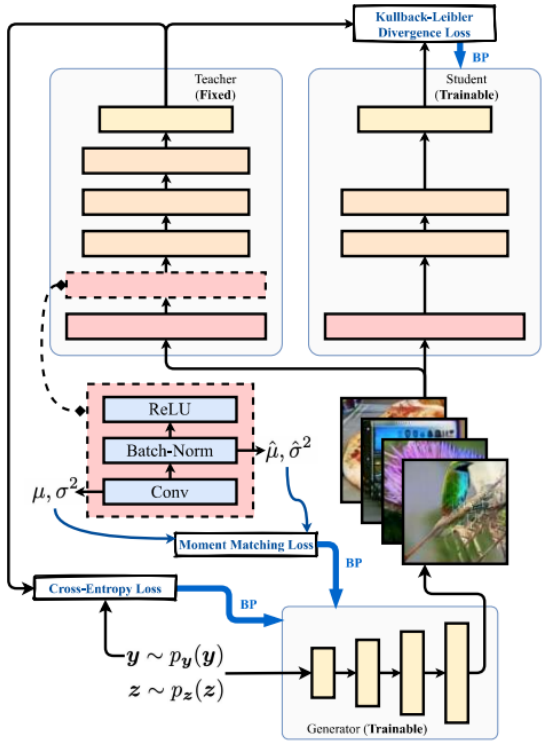
Preliminary: Framework of DFKD

Step1: Image Synthesis

$$\mathcal{L}_{inv}(x) = \alpha \cdot \mathcal{L}_{oh}(x) + \beta \cdot \mathcal{L}_{bns}(x) + \gamma \cdot \mathcal{L}_{adv}(x)$$

Step2: Knowledge Distillation

$$\mathcal{L}_{KD}(x) = KL(T(x, \theta_t)/\tau \,|S(x, \theta_s)/\tau)$$



Motivation

- Accuracy gap between KD and DFKD
  - Make the synthesized images closer to the real images.

- Differences between synthesized images & real images?
  - From the loss term of image synthesis.
  - no difference between the intermediate features of different categories of images

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

Makes images different.
But only Constrained on the last layer.

$$\mathcal{L}_{\text{bns}}(x) = \frac{1}{N}\sum_{i=0}(|\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i|_2^2 + |\tilde{\boldsymbol{\sigma}}_i - \boldsymbol{\sigma}_i|_2^2) \longrightarrow$$

Makes all BNS the same

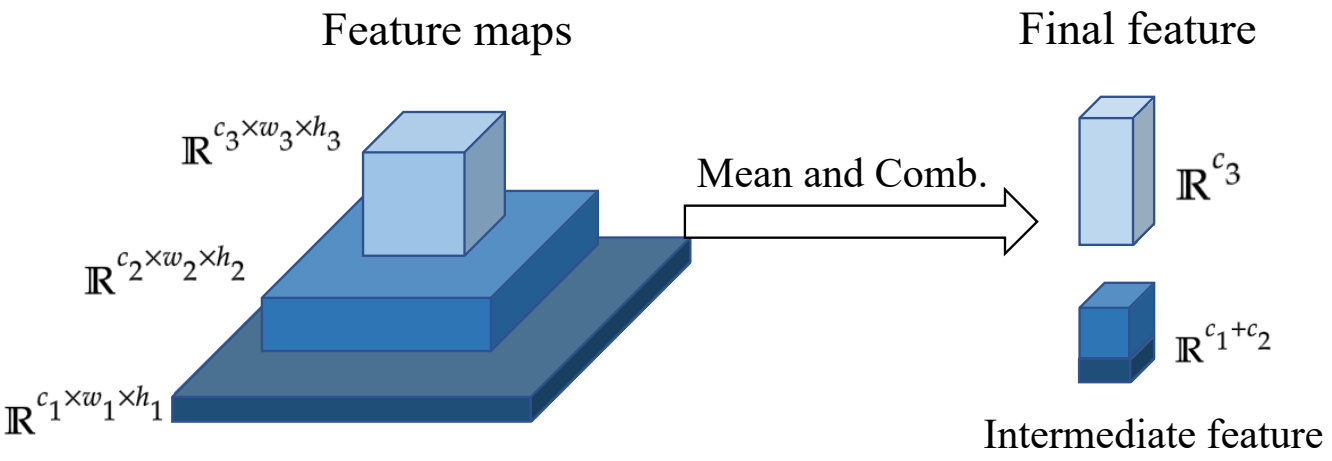
$$\mathcal{L}_{adv}(x) = -KL(T(\hat{x}, \theta_t)/\tau \,|S(\hat{x}, \theta_s)/\tau) \longrightarrow$$

Well, none of your business...

Motivation

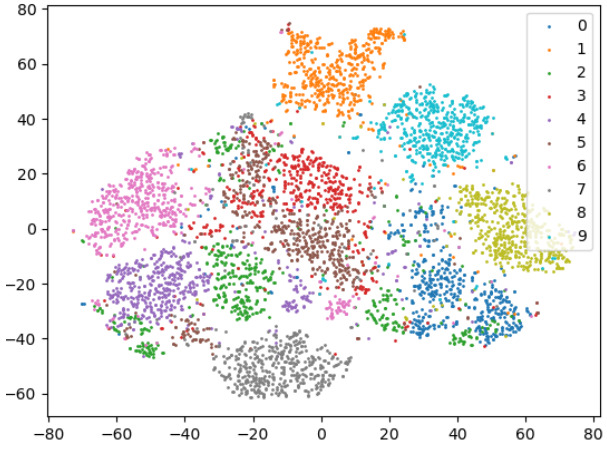
- Differences between synthesized images & real images?
  - No difference between the intermediate features
  - GoogleNet: Features in the intermediate layer are discriminative!

The definition of the intermediate feature

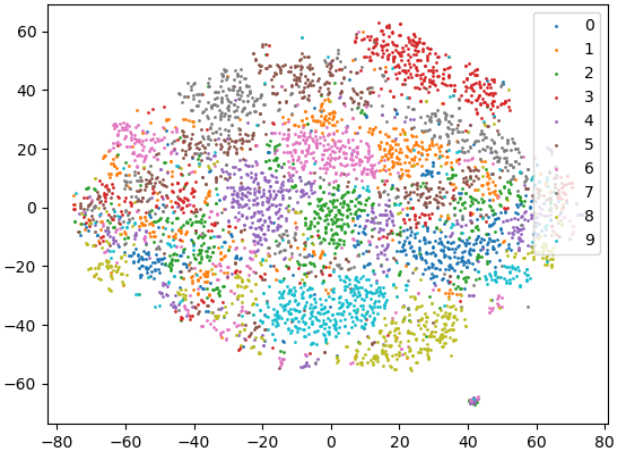


Motivation

- Differences between synthesized images & real images?
  - No difference between the intermediate features
  - GoogleNet: Features in the intermediate layer are discriminative!



Real images



synthesized images

t-sne visualization results of inter-mediate features on wrn40-2, CIFAR-10

- Real images: **Heterogeneous**
- Synthesized images: Chaotic.

Intermediate-Feature Heterogeneity Enhancement

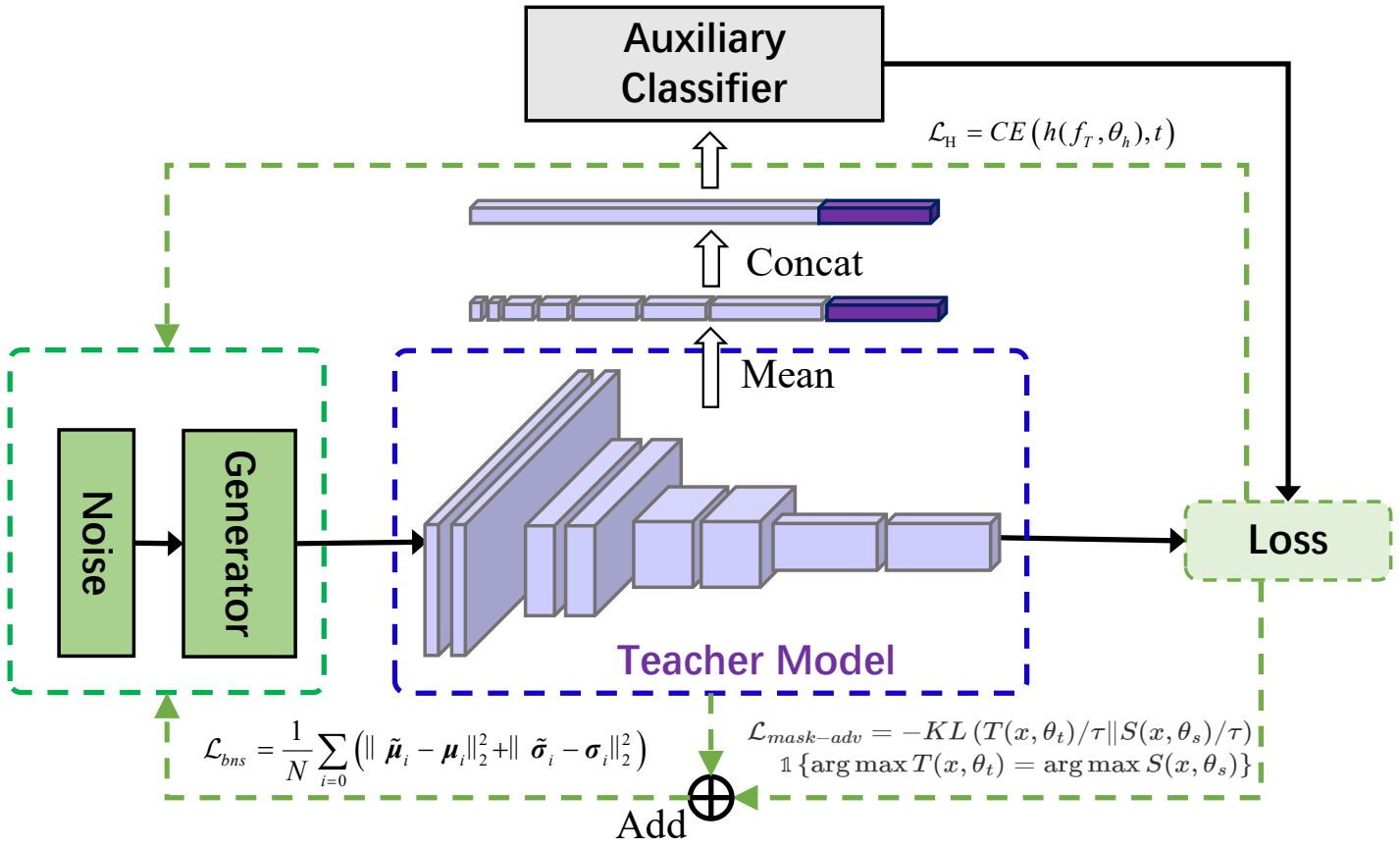
- How to make the intermediate features heterogeneous?
  - Use an auxiliary classifier to enhance heterogeneity of intermediate-features

- Curse of $\mathcal{L}_{oh}$: too confident is not beneficial to knowledge transfer.
  - Other works[1,2] : soft target of $\mathcal{L}_{oh}$, target mix.
  - **Ours**: merge $\mathcal{L}_{oh}$ into feature heterogeneity enhancement.

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

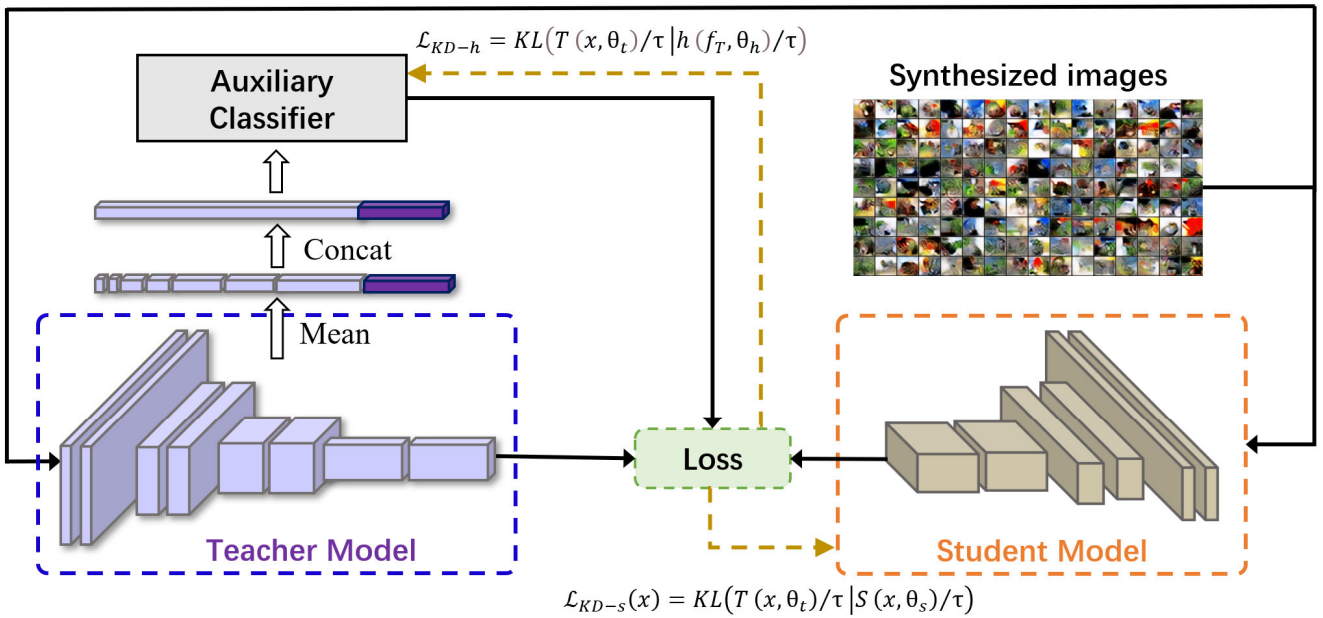
IFHE: Image Synthesis

- Concat features.
- Auxiliary classifier use features to predict classes.
- Train noises and generator by $\mathcal{L}_H$, $\mathcal{L}_{bns}$, and $\mathcal{L}_{adv-mask}$.
- Save images.



IFHE: Knowledge Distillation

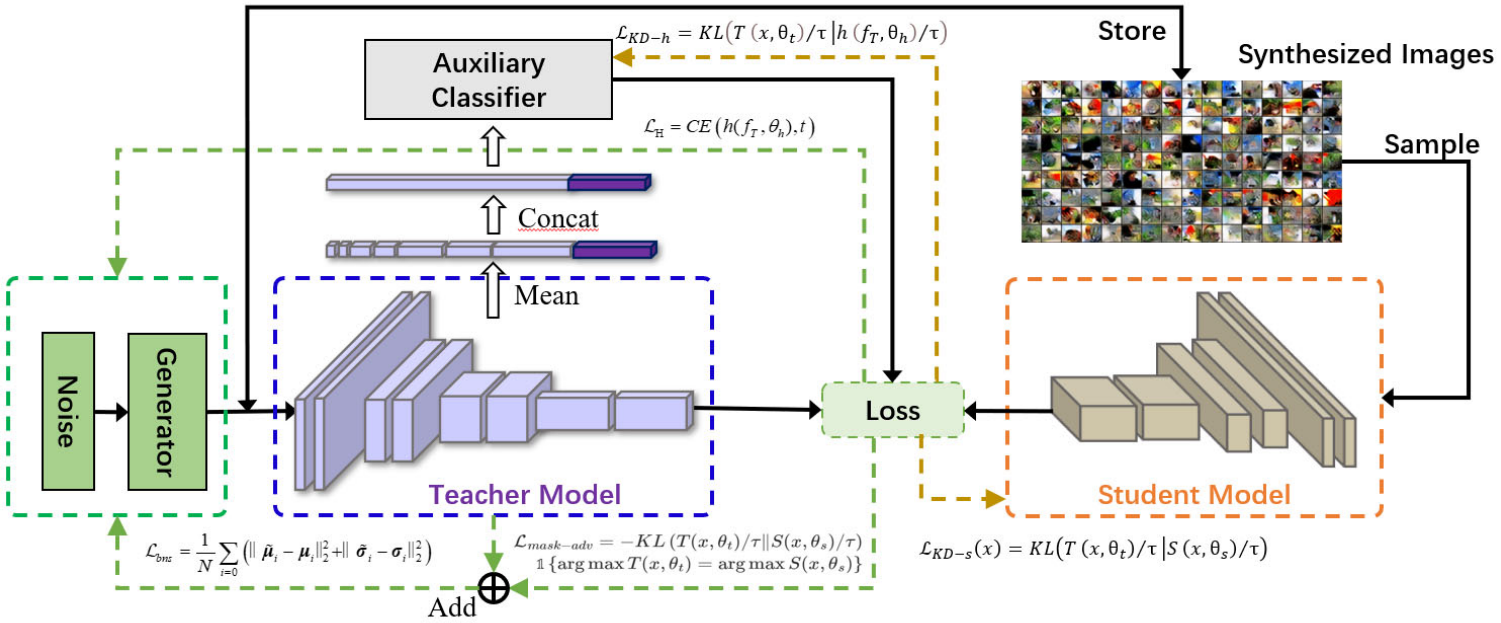
- Sample Images
- Put images into teacher and student.
- Train student and auxiliary classifier by $\mathcal{L}_{KD-s}$ and $\mathcal{L}_{KD-h}$ respectively.



Framework of IFHE Data-Free Knowledge Distillation

Step1: Image Synthesis

$$\mathcal{L}_{\text{inv}}(z, g_\theta) = \alpha \cdot \mathcal{L}_{\text{H}} + \beta \cdot \mathcal{L}_{\text{bns}} + \gamma \cdot \mathcal{L}_{\text{mask-adv}}$$

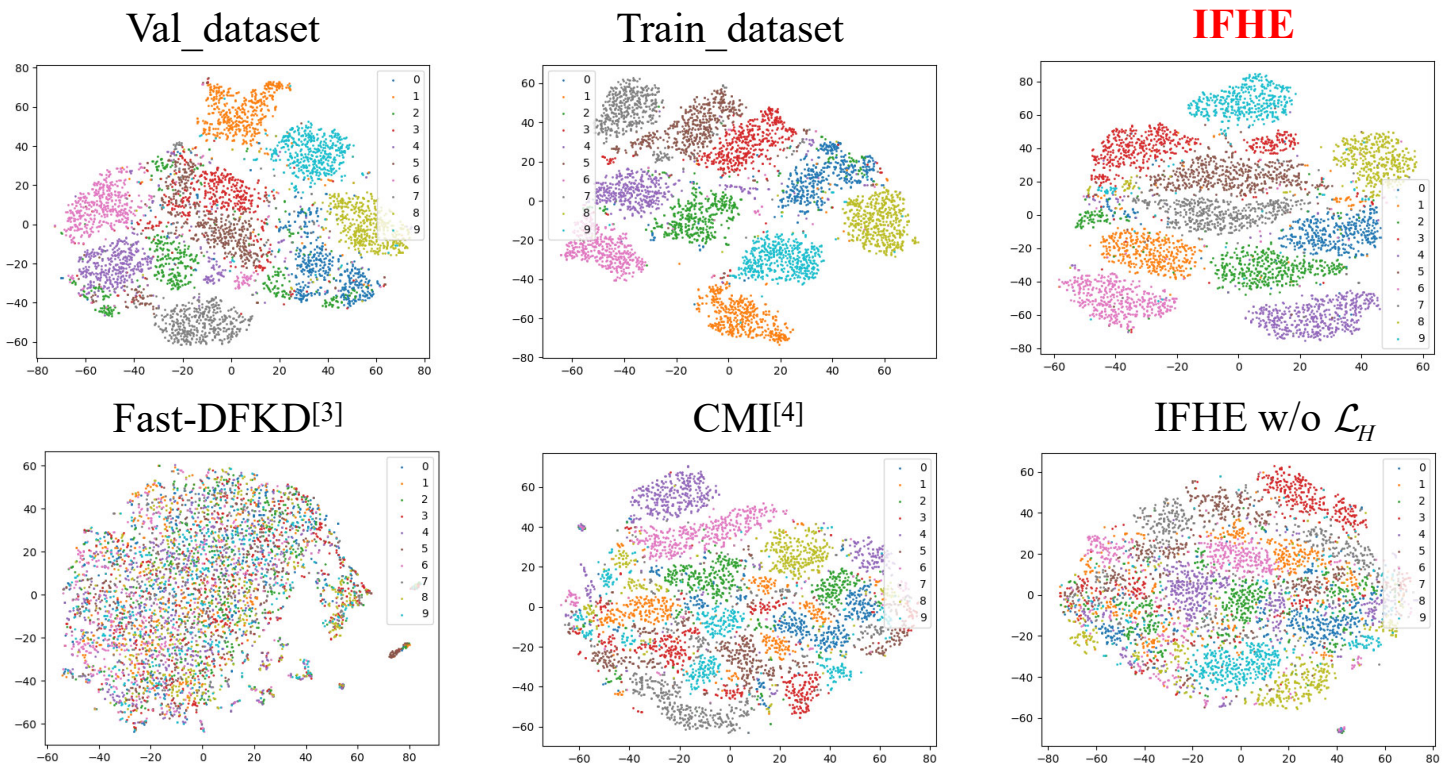


Step2: Knowledge Distillation

$$\mathcal{L}_{KD-h} = KL(T(x,\theta_t)/\tau \,|h(f_T,\theta_h)/\tau) \qquad \mathcal{L}_{KD-s}(x) = KL(T(x,\theta_t)/\tau \,|S(x,\theta_s)/\tau)$$

Experimental Results: Intermediate Features

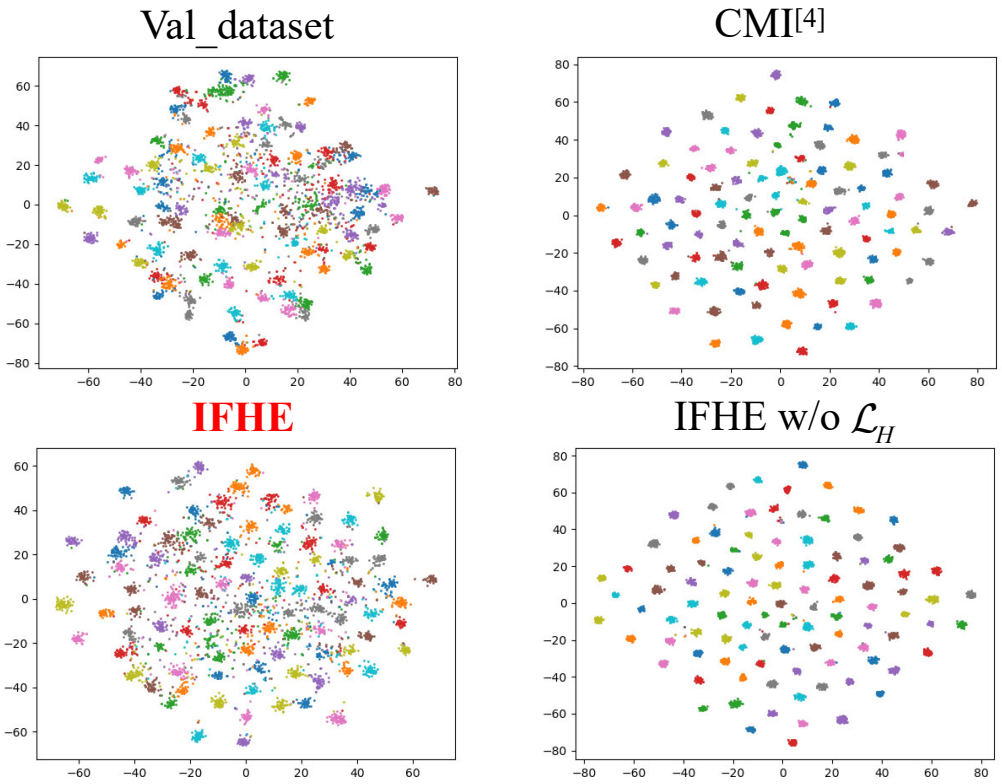
- Intermediate features by IFHE is more discriminative than other DFKD methods



t-sne visualization results of intermediate features

Experimental Results: Final Features

- Final features by IFHE contains more category information, which is more likely to real images.



t-sne visualization results of final features

Experimental Results: CIFAR10/100

CIFAR10

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**

CIFAR100

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**

- Experiment on CIFAR-10/100 dataset, various teacher-student pairs.
- Outperform all other DFKD methods.
- WRN40-2/WRN16-1 pair even exceed train scratch with original data.

Experimental Results: Ablation study

- Different scaling factors of $\mathcal{L}_H$
- Because CIFAR100 with more categories, the best α is larger than CIFAR10's

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

- $\mathcal{L}_H$ has good 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**

Conclusion

- Intermediate features on real images and synthesized images
- Propose IFHE method to enhance intermediate features heterogeneous
- Consistently outperforms other DFKD methods

Reference

[1] Wang Z. Data-free knowledge distillation with soft targeted transfer set synthesis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(11): 10245-10253.

[2] Li Y, Zhu F, Gong R, et al. Mixmix: All you need for data-free compression are feature and data mixing[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4410-4419.

[3] Fang G, Mo K, Wang X, et al. Up to 100x faster data-free knowledge distillation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(6): 6597-6604.

[4] Fang G, Song J, Wang X, et al. Contrastive model inversion for data-free knowledge distillation[J]. arXiv preprint arXiv:2105.08584, 2021.

[5] Chen H, Wang Y, Xu C, et al. Data-free learning of student networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3514-3522.

[6] Micaelli P, Storkey A J. Zero-shot knowledge transfer via adversarial belief matching[J]. Advances in Neural Information Processing Systems, 2019, 32.

[7] Yin H, Molchanov P, Alvarez J M, et al. Dreaming to distill: Data-free knowledge transfer via deepinversion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8715-8724.

[8] Choi Y, Choi J, El-Khamy M, et al. Data-free network quantization with adversarial knowledge distillation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 710-711.

Thank you!