Supplementary Material *for* Distribution Shift Matters for Knowledge Distillation with Webly Collected Images

Jialiang Tang[1,2,3], Shuo Chen[4,*], Gang Niu[4], Masashi Sugiyama[4,5], Chen Gong[1,2,3,*]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, China
[3]Jiangsu Key Laboratory of Image and Video Understanding for Social Security, China
[4]Center for Advanced Intelligence Project, RIKEN, Japan
[5]The Graduate School of Frontier Sciences, The University of Tokyo, Japan

Abstract

Here, we offer detailed information about the experiments in Section 1 and show intensive visualization results of our proposed KD^3 and the compared baseline methods in Section 2.

1. Related Information of Experiments

This section introduces detailed information about the datasets and DNNs used in our experiments and describes the compared data-free model compression methods.

1.1. Datasets

In Table 1, we introduce the details of the datasets used in our experiments, including the image shape, the number of categories, and the number of images in training and test sets.

1.2. Calculational and Storage Burdens of Various DNNs

In this section, by following the classic knowledge distillation approaches [10, 17], we infer the calculational requirements of DNNs by their needed floating-point operations (FLOPs) to process an input image in TinyImageNet and measure the storage burdens of DNNs by counting their learnable parameters. As shown in Table 2, we can see that the calculation and memory burdens for student networks are significantly less than teacher networks, demonstrating that our KD^3 can effectively compress the large pre-trained teacher network to obtain a lightweight student network.

*Corresponding authors: Chen Gong (chen.gong@njust.edu.cn), Shuo Chen (shuo.chen.ya@riken.jp).

Dataset	Image shape	#classes	#training set	#test set
MNIST [13]	1×28×28	10	50,000	10,000
MNIST-M [8]	3×32×32	10	59,001	90,001
SVHN [16]	3×32×32	10	73,257	26,032
CIFAR10 [11]	3×32×32	10	50,000	10,000
CIFAR100 [11]	3×32×32	100	50,000	10,000
CINIC [4]	3×32×32	10	90,000	90,000
TinyImageNet [12]	3×64×64	200	100,000	10,000
ImageNet [5]	3×256×256	1,000	1,281,167	50,000

Table 1. The details of datasets used in our experiments, the items with the prefix "#" denote the cardinality. The "1×28×28" in column "Image shape" represents the channel number, height, and width of images in the corresponding dataset are 1, 28, and 28, respectively.

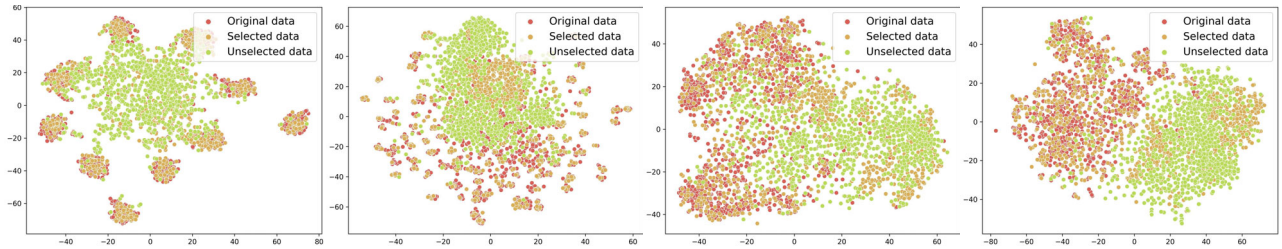
Teacher	Student	#params		FLOPs	
		Teacher	Student	Teacher	Student
ResNet32×4 [9]	ResNet8×4 [9]	7.41M	1.21M	4.35G	0.71G
ResNet32×4 [9]	MobileNetV2 [18]	7.41M	0.81M	4.35G	29.23M
ResNet32×4 [9]	ShuffleV1 [21]	7.41M	0.86M	4.35G	168.59M
ResNet32×4 [9]	ShuffleV2 [14]	7.41M	1.26M	4.35G	186.82M
ResNet110×2 [9]	ResNet110 [9]	6.89M	1.73M	4.06G	1.03G
ResNet110×2 [9]	ResNet116 [9]	6.89M	1.83M	4.06G	1.09G
ResNet110×2 [9]	ShuffleV1 [21]	6.89M	0.86M	4.06G	168.59M
ResNet110×2 [9]	ShuffleV2 [14]	6.89M	1.26M	4.06G	186.82M

Table 2. The parameters (in millions, M) and floating-point operations (FLOPs, in Gigas, G) of various teacher networks and student networks on the TinyImageNet [12] dataset.

1.3. Introduction for Compared Methods

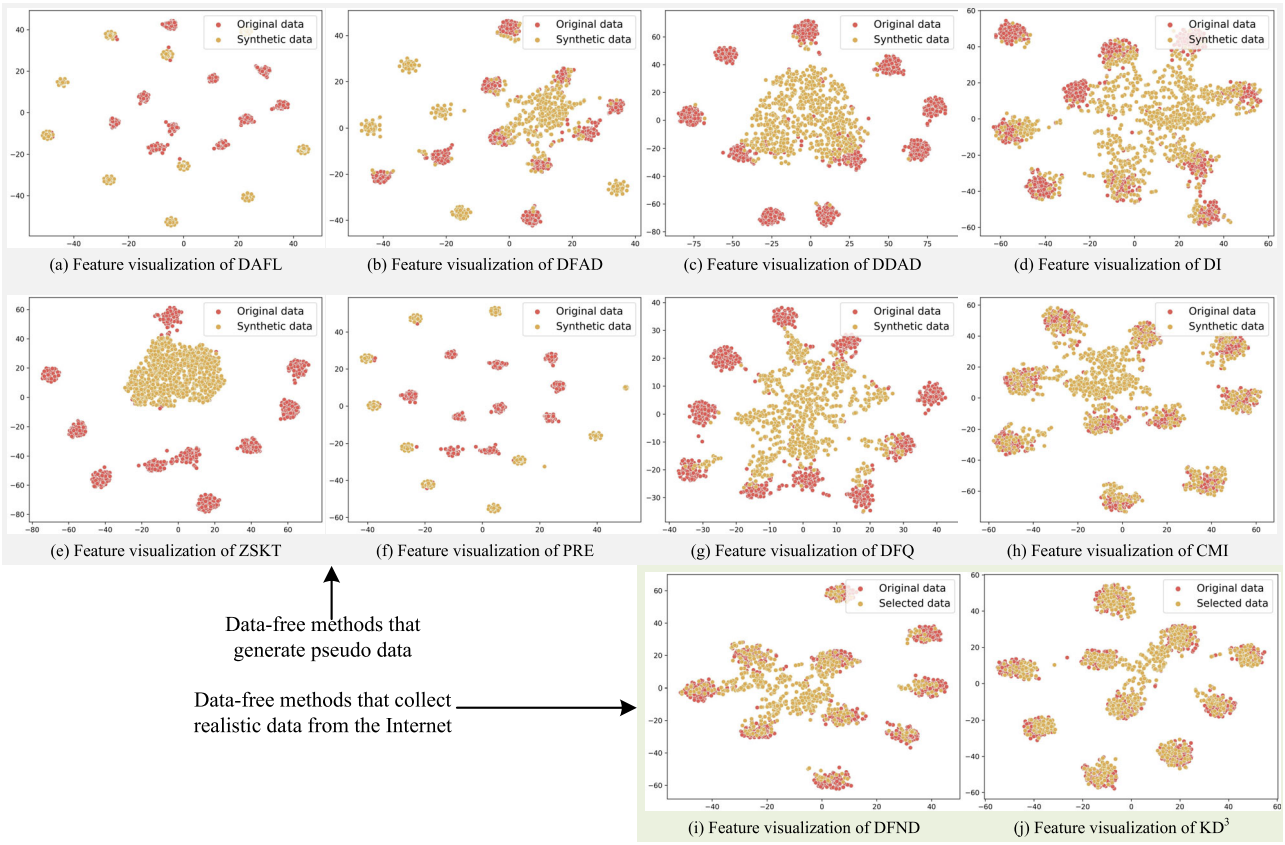
In this paper, we introduce the compared data-free model compression methods, which utilize the generated pseudo data or data collected from the Internet to train student network without using the original training data. The compared methods are described as follows:

- Data-Free Learning (DAFL) [2], which utilizes a sim-



(a) Feature visualization on CIFAR10 (b) Feature visualization on CIFAR100 (c) Feature visualization on CINIC (d) Feature visualization on TinyImageNet

Figure 1. Visualization of features produced by ResNet18 (trained by our proposed KD^3) using t-SNE [19]. The original images are randomly chosen from (a) CIFAR10 [11], (b) CIFAR100 [11], (c) CINIC [4], and (d) TinyImageNet [12], and the selected/unselected images are sampled from ImageNet [5]. Here, each item contains 1,000 images.



(a) Feature visualization of DAFL (b) Feature visualization of DFAD (c) Feature visualization of DDAD (d) Feature visualization of DI

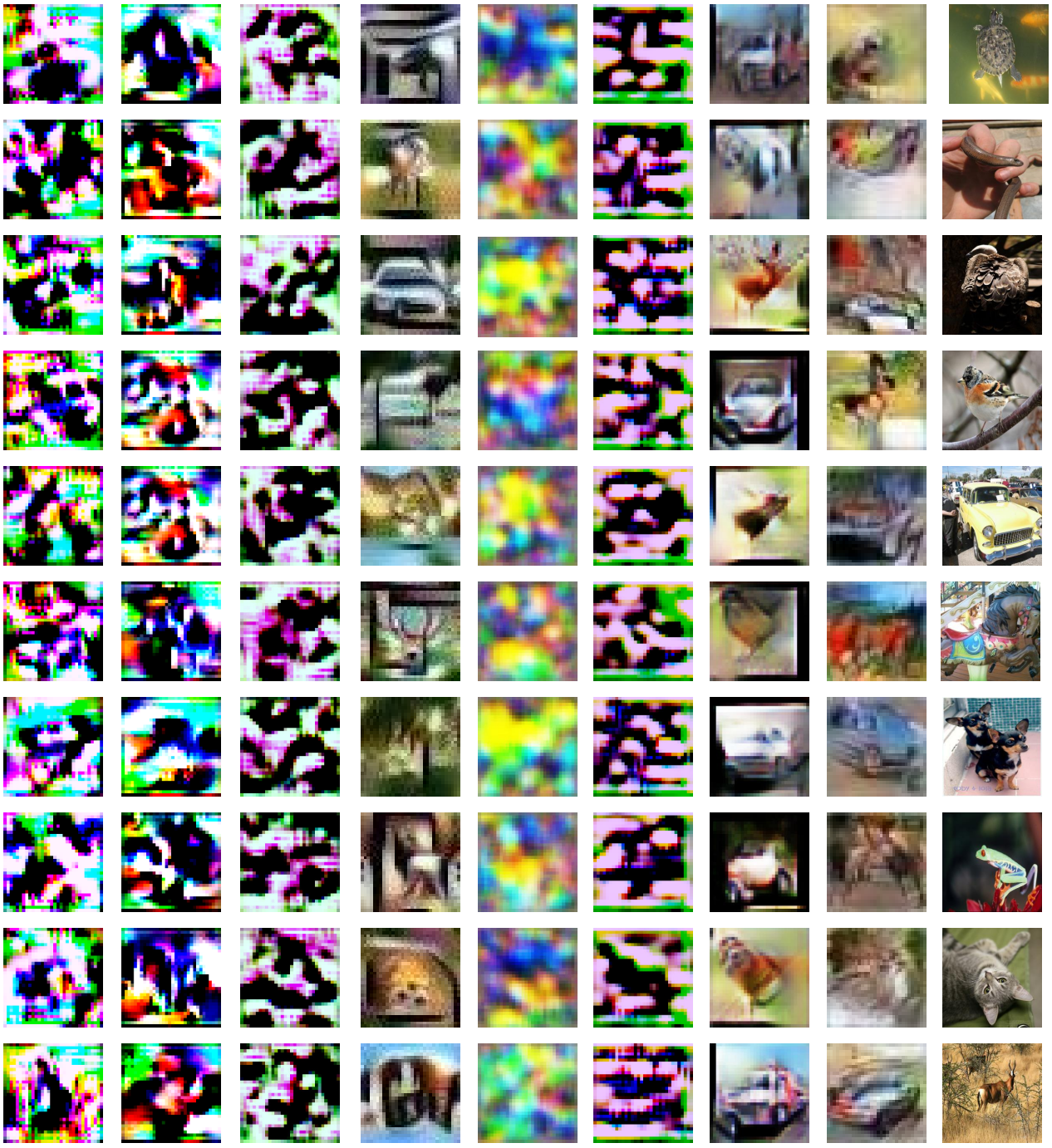
(e) Feature visualization of ZSKT (f) Feature visualization of PRE (g) Feature visualization of DFQ (h) Feature visualization of CMI

(i) Feature visualization of DFND (j) Feature visualization of KD^3

Figure 2. Visualization of ResNet34-produced features by t-SNE [19]. There are 1,000 images in the original data and synthetic/selected data, respectively. The original images are randomly sampled from CIFAR10 [11], and the selected data in subfigures (i) and (j) belongs to ImageNet [5].

ple generator to produce pseudo data under the supervision of teacher network.

- Data-Free Adversarial Learning (DFAD) [6], which uses teacher network together with student network to guide the data approximation of generator.
- DeepInversion (DI) [20], which directly inverts the means and variances in the features of teacher network to reconstruct the training images.
- Dual Discriminator Adversarial Distillation (DDAD) [22], which further extracts the means and variances stored in batch normalization layers of teacher network to improve the visual authenticity of synthetic images produced by generator.
- Zero-Shot Knowledge Transfer (ZSKT) [15], which encourages student network to predict similarly to teacher network on the "hard instances" produced by

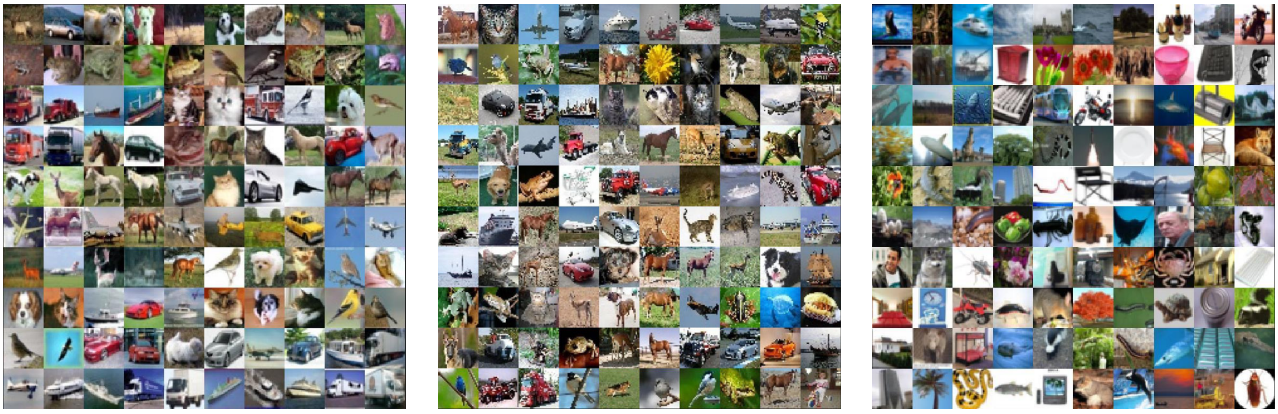


(a) DAFL (b) DFAD (c) DDAD (d) DI (e) ZSKT (f) PRE (g) DFQ (h) CMI (i) DFND

Figure 3. Visualization of the pseudo images (as shown in subfigures (a)-(h)) and webly collected images (as shown in subfigure (i)) generated/selected by the compared data-free model compression methods.

an adversarial generator, thus improving the performance of student network.

- Pseudo Replay Enhanced Data-Free Knowledge Distillation (PRE) [22], which utilizes multiple rounds of pseudo data to train student network, since the student network can memorize the previously learned knowledge.



(a) Original images (b) Selected images (c) Unselected images

Figure 4. Visualization of original images in CIFAR10 (subfigure (a)), webly collected images in ImageNet selected/unselected by our proposed KD^3 (subfigure (b)/(c)).

- Data-Free Quantization (DFQ) [3], which mimics the activations between multi-layers of the pre-trained model and quantized model to recover the performance of quantized model without finetunes on the original data.
- Contrastive Model Inversion (CMI) [7], which promotes generator to produce distinguishable instances by contrastive learning.
- Data-Free Noisy Distillation (DFND) [1], the only existing data-free knowledge distillation method that utilizes the webly collected data, which searches confident instances from the Internet to train student networks.

2. Additional Visualization Results

This section offers many visualization results to further demonstrate the effectiveness of our proposed KD^3.

2.1. Visualization Features of Student Network

We visualize the features produced by student network (ResNet18) trained by our KD^3. The original images are randomly sampled from CIFAR10, CIFAR100, CINIC, and TinyImageNet. Meanwhile, the selected and unselected images are provided by ImageNet. As shown in Fig. 1, the distribution of images selected by student network is close to that of original images in the feature space. This visualization results demonstrate that the student network trained by our KD^3 can select proper instances from the webly collected data for itself training.

2.2. Visualization Features of Compared Methods

We visualize the features of pseudo images generated by the compared data-free model compression methods, including DAFL, DI, DDAD, ZSKT, PRE, DFQ, and the real-world images selected by DFND. More specifically, the original data is CIFAR10, and ResNet34 is employed to process original images, synthetic images, and real-world images (selected from the ImageNet by DFND and our KD^3).

The visualization results are shown in Fig. 2. We can observe that the images selected by our proposed KD^3 are closer to the original data distribution than those generated or selected by the compared data-free methods. These visualization results further demonstrate the effectiveness of the data selection method in our KD^3, which can select more useful training instances for student network than other baseline methods.

2.3. Visualization of Synthetic and Selected Data

We visualize the pseudo images generated by DAFL, DI, DDAD, ZSKT, PRE, DFQ, and the webly collected images selected by DFND and our proposed KD^3. We assume the original data is CIFAR10 and the teacher-student pair is ResNet34-ResNet18. As shown in Fig. 3, we can observe that the visual quality of the synthetic images (*i.e.*, subfigures (a)-(h)) is poor. As a result, the performance of student networks trained on these pseudo data is suboptimal. Meanwhile, we can see that the webly collected images selected by DFND contain some images with different distributions of the original data (*e.g.*, the images of "turtle" and "trojan" in subfigure (i)). Therefore, the student network trained by DFND still can not compare with the same one trained on the original data because it ignores the distribution shift. In contrast, Fig. 4 shows the webly collected images selected by our proposed KD^3. In practice, the selected images are similar to those in the original data yet contain some images with different distributions. Fortunately, our proposed KD^3 can effectively solve this distribution shift. Thus, the student network trained by our proposed KD^3 can achieve comparable performance to that trained on the orig-

inal data.

References

[1] Hanting Chen, Tianyu Guo, Chang Xu, Wenshuo Li, Chunjing Xu, Chao Xu, and Yunhe Wang. Learning student networks in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6428–6437, 2021. 4

[2] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3514–3522, 2019. 1

[3] Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, pages 710–711, 2020. 4

[4] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018. 1, 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 1, 2

[6] Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv:1912.11006*, 2019. 2

[7] Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*, 2021. 4

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research (JMLR)*, 17(1):2096–2030, 2016. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 1

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 2

[12] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 1, 2

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[14] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 1

[15] Paul Micaelli and Amos J Storkey. Zero-shot knowledge transfer via adversarial belief matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 2

[16] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 1

[17] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv:1412.6550*, 2014. 1

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. 1

[19] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(11), 2008. 2

[20] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8715–8724, 2020. 2

[21] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, 2018. 1

[22] Haoran Zhao, Xin Sun, Junyu Dong, Milos Manic, Huiyu Zhou, and Hui Yu. Dual discriminator adversarial distillation for data-free model compression. *International Journal of Machine Learning and Cybernetics (IJMLC)*, 13(5):1213–1230, 2022. 2, 3