Data-Free Knowledge Distillation via Feature Exchange and Activation Region Constraint

Shikang Yu[1,2*] Jiachen Chen[1,2*] Hu Han[1,2,3] Shuqiang Jiang[1,2]
[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
[3] Peng Cheng Laboratory, Shenzhen 518055, China
shikang.yu@vipl.ict.ac.cn; chenjiachen20@mails.ucas.edu.cn; {hanhu, sqjiang}@ict.ac.cn

Abstract

Despite the tremendous progress on data-free knowledge distillation (DFKD) based on synthetic data generation, there are still limitations in diverse and efficient data synthesis. It is naive to expect that a simple combination of generative network-based data synthesis and data augmentation will solve these issues. Therefore, this paper proposes a novel data-free knowledge distillation method (SpaceshipNet) based on channel-wise feature exchange (CFE) and multi-scale spatial activation region consistency (mSARC) constraint. Specifically, CFE allows our generative network to better sample from the feature space and efficiently synthesize diverse images for learning the student network. However, using CFE alone can severely amplify the unwanted noises in the synthesized images, which may result in failure to improve distillation learning and even have negative effects. Therefore, we propose mSARC to assure the student network can imitate not only the logit output but also the spatial activation region of the teacher network in order to alleviate the influence of unwanted noises in diverse synthetic images on distillation learning. Extensive experiments on CIFAR-10, CIFAR-100, Tiny-ImageNet, Imagenette, and ImageNet100 show that our method can work well with different backbone networks, and outperform the state-of-the-art DFKD methods. Code will be available at: https://github.com/skgyu/SpaceshipNet.

1. Introduction

Knowledge distillation (KD) aims to train a lightweight student model that can imitate the capability of a pre-trained complicated teacher model. In the past decade, KD has been studied in a wide range of fields such as image recognition, speech recognition, and natural language processing. Traditional KD methods usually assume that the whole or part of the training set used by the teacher network is accessible by the student network [17, 24, 34]. But in practical applications, there can be various kinds of constraints in accessing the original training data, e.g., due to privacy issues in medical data [1, 2, 5, 20, 28, 35] and portrait data [3], and copyright and privateness of large data volume such as JFT-300M [40] and text-image data [37]. The traditional KD methods no longer work under these scenarios. Recently, data-free knowledge distillation (DFKD) [4,7,10,13,14,22,25,27,29,43,46] seeks to perform KD by generating synthetic data instead of accessing the original training data used by the teacher network to train the student network. Thus, the general framework of DFKD consists of two parts: synthetic data generation that replicates the original data distribution and constraint design between student and teacher network during distillation learning. Synthetic data generation methods in DFKD mainly consist of noise image optimization-based methods [4, 26, 31, 44] and generative network-based methods [8, 9, 12, 13, 27, 29, 45]. The former approaches optimize randomly initialized noise images to make them have the similar distribution to the original training data. These methods can theoretically generate an infinite number of independent and identically distributed images for student network learning, but they are usually extremely time-consuming, and thus are difficult in generating sufficient synthetic data with high diversity. The later approaches learn a generator to synthesize images that approximate the distribution of the original training data. These methods can be much faster than the image

[*] Equal contribution

This research was supported in part by the National Key R&D Program of China (grant 2018AAA0102500), and the National Natural Science Foundation of China (grant 62176249).

optimization-based approach, but the diversity of the synthesized data is usually limited because the generation of different images are not completely independent with each other.

Despite the encouraging results achieved, DFKD remains a challenging task, because the synthetic data may have a different distribution from the original data, which could potentially result in bias in student network learning. The possible reason is that the noises in the synthesized images can easily lead to the bias of the network's region of interest. In addition, the widely used KL divergence constraint between student and teacher networks in existing DFKD methods may not work well with synthetic data [4].

This paper proposes a novel DFKD method that utilizes channel-wise feature exchange (CFE) and multi-scale spatial activation region consistency (mSARC) constraint to improve knowledge transfer from the teacher network to the student network. The proposed method enhances the diversity of synthetic training data and the robustness to unwanted noises in synthetic images during distillation. Unlike previous generative network-based methods that employed multiple generators to synthesize images [27] or reinitialization and retraining of generator [13] to enhance the synthetic training data diversity, our method improves the synthetic training data diversity by using the features of early synthetic images to perform CFE. When our generative network and those of other methods have learned to generate the same number of synthetic images, the proposed method can produce more diverse training data for distillation. However, CFE also amplifies unwanted noise in synthetic images, which may hinder distillation learning (traditional data augmentation methods also suffer from this problem, e.g., CutMix [47] and Mixup [50]). To address this issue, we propose the mSARC constraint, which enables the student network to learn discriminative cues from similar regions to those used by the teacher network, effectively overcoming the limitations of the traditional KL divergence loss when applied to synthetic images during distillation learning. Moreover, combining our mSARC with traditional data augmentation methods [47,50] can still significantly improve distillation learning with synthetic data.

We evaluate our method on a number of datasets, including CIFAR-10 [19], CIFAR-100 [19], and Tiny-ImageNet [21]. Our approach demonstrates superior performance compared to the state-of-the-art DFKD methods. Moreover, we observe that the student networks trained using our DFKD method achieve comparable performance to those trained using original training data. Additionally, we evaluate our method on subsets of ImageNet [11] with 10 and 100 classes and a resolution of 224 × 224, validating the efficacy of our method on generating high-resolution synthetic images for distillation learning. In our ablation study, we verify the effectiveness of the key components (CFE and mSARC), we find that mSARC plays a particularly important role when strong data augmentation is applied to the synthetic images.

2. Related Work

2.1. Data-free Knowledge Distillation

DFKD aims at transferring knowledge from pre-trained teacher model (usually a big model) to a student model (usually a lightweight) without access to the original training data. Lopes et al. [26] first tried to utilize the teacher model and its metadata to reconstruct the original training data for KD, but it is unlikely to obtain the metadata in practice. Unlike [26], many subsequent works studied DFKD without relying on metadata prior. Some methods optimized randomly initialized noise images to generate synthetic data [4,31,44]. Nayak et al. [31] modeled the teacher network's output as a Dirichlet distribution and used it as a constraint when optimize the noise images to obtain synthetic images. Yin et al. [44] regularized the distribution of the synthesized images according to the statistics stored in the batch normalization layers of the teacher network. Since the optimization process of each image is independent the computations of different images are not shared. Thus, these methods can be extremely slow in synthetic image generation [44].

Instead of optimizing randomly initialized noise images, some method obtain synthetic images based on generative network to synthesize training data, i.e., methods based on generative networks [8, 9, 12, 13, 27, 29, 45]. According to different types of network input, these methods can be grouped into two categories, i.e., non-conditional generative network-based methods [8,9,12,13,29] and conditional generative network-based methods [27, 45]. The former typically generates synthetic images by sampling random noises in a generative network. The latter combines random noises and a conditional vector to generate synthetic images, and thus is able to better control the class of the synthetic images. While generative network-based approaches are more efficient, the diversity of the synthetic images can be limited [12].

CDFKD-MFS [14] aimed to solve different DFKD problem, i.e., distilling knowledge from multiple teacher networks (teachers with different parameters) without accessing original data. They use a student network with additional parameters. The student network uses multi-level feature-sharing to learn from multiple teachers, and the predictions of multi-student headers are aggregated to improve performance.

2.2. Data Augmentation via Data Mixing

Data mixing (DM) in image domain [39, 41, 47, 50] and DM in feature domain [6, 42] has been widely explored to improve model generalization ability.

Image domain data mixing usually include linear interpolation [50] and spatial level substitution [39, 41, 47] between different images. Compared with DM method in image domain, there are some differences expected in the objectives of our method. Our approach is to increase the diversity of training data by mixing the data of synthetic images through channel mixing at the feature level. Moreover, in terms of mechanisms of DM in image domain, these methods may work better for natural images than synthetic data. CutMix [47] has been shown to be effective in directing the model to focus on the less discriminative parts of the object. However, since our image synthesis is guided by the classification network only, the details contained in the synthesized images are consistent with the information of interest to the classification network.

Compared with some DM methods in feature domains [6, 42], they do share some similarities with our approach in that we are both operating at the feature level. While most DM in feature domains method is an interpolation operation on the features of two samples, except Cao et al.'s [6] proposed method, which is a mixing operation on the feature channels just like ours. But their work differs from our approach in terms of motivation and implementation. In terms of motivation, our feature exchange aims to enrich the diversity of synthetic images used for DFKD, their work is more of a regularization approach embedded inside the classification network (like dropout [38] does). In terms of implementation, since their method requires the help of shallow parameters of the classification network for feature generation, this makes their method unable to make the shallow part of the classification network benefit from this regularization approach. In contrast, our feature exchange is performed in the generative network, so there is no such problem. It is worthwhile to discuss additionally that the feature mixing in Mixmix [23] is different from the aforementioned DM in feature domain, where "Mix" means loss-optimized mixing, i.e., using multiple pre-trained models to invert the same image through the loss function.

2.3. Feature-level Knowledge Distillation

Some KD work [16, 18, 29, 32, 49] differs from previous KD via logit distillation, in that they impose consistency constraints on the middle layer features of the network. These methods, like ours, are constrained for the middle layer of the network to achieve KD. However, there are differences between these methods and our approach in terms of implementation and objectives, especially in terms of objectives. Among them, Fitnets [32], OFD [16] and FT [18] are more to constrain the intermediate layers of student and teacher network to have similar responses to the same image, rather than constraining the spatial region concerned by the network to be consistent in the spatial region of attention. The common feature-layer knowledge distillation is a stronger consistency constraint compared to logit distillation, and they not only fail to improve the robustness of knowledge distillation to noise but even have negative effects (the results in Table 3 also confirms this). We find that when data augmentation is performed on synthetic images, the noise in the synthetic data is further amplified to the extent that knowledge distillation cannot be performed properly, and our method can solve this issue, because the optimal region of interest should always be aligned with the edges of the classified objects.

3. Proposed Method

3.1. Problem Formulation

Let $D = \{\mathcal{X} \in R^{c\times h\times w}, \mathcal{Y} = 1, 2, ..., K\}$ denote a training dataset, in which $x^i \in \mathcal{X}$ is an image, and $y^i \in \mathcal{Y}$ is its label. Let $T(x;\theta_T)$ denote a pre-trained teacher network on D. The goal of KD is to learn a lightweight student classification network $S(x, \theta_S)$ that can imitate the classification capability of $T(x;\theta_T)$ using D. Hinton et al. [17] first proposed knowledge refinement to learn the weights of student network with the object of $\min_{\theta_s} \mathcal{L}_{cls} + \mathcal{L}_{KL}$, where $\mathcal{L}_{cls} = \mathrm{E}_{x,y\sim p(\mathcal{X},\mathcal{Y})} \lambda_{CE}(S(x;\theta_S), y)$ is a cross-entropy loss and $\mathcal{L}_{KL} = \mathrm{E}_{x,y\sim p(\mathcal{X},\mathcal{Y})} \mathcal{L}_{KL}(T(x;\theta_T)||S(x;\theta_S))$ is the Kullback-Leibler divergence (KL loss) between the outputs of the student and the teacher network. Different from KD, the goal of DFKD is to learn a lightweight student classification network $S(x, \theta_S)$ that can imitate the classification capability of $T(x;\theta_T)$ without using D.

We follow a common solution of DFDK by generating synthetic images and using them to learn a student network. Specifically, our method contains three essential parts: optimizing the generative network using the teacher network, training image generation using channel-wise feature exchange (CFE), and knowledge distillation using multi-scale spatial activation region consistency (mSARC) constraint (see Fig. 1). Using our CFE, the generator can effectively generate diverse synthetic images for distillation learning. Using mSARC, the student network can learn to imitate not only the logit output but also the visual cues of the teacher network. We detail our method as follows.

3.2. Optimizing the Generative Network Using Teacher Network

We expect to learn a generative network $G : z \rightarrow X$ to generate synthetic data $\{\hat{x}, y\}$ whose distribution is similar to the data $D = \{\mathcal{X}, \mathcal{Y}\}$. We first randomly initialize a mini-batch of noise-label pairs (z, y) per epoch, where z is a random noise sampled from a Gaussian distribution and

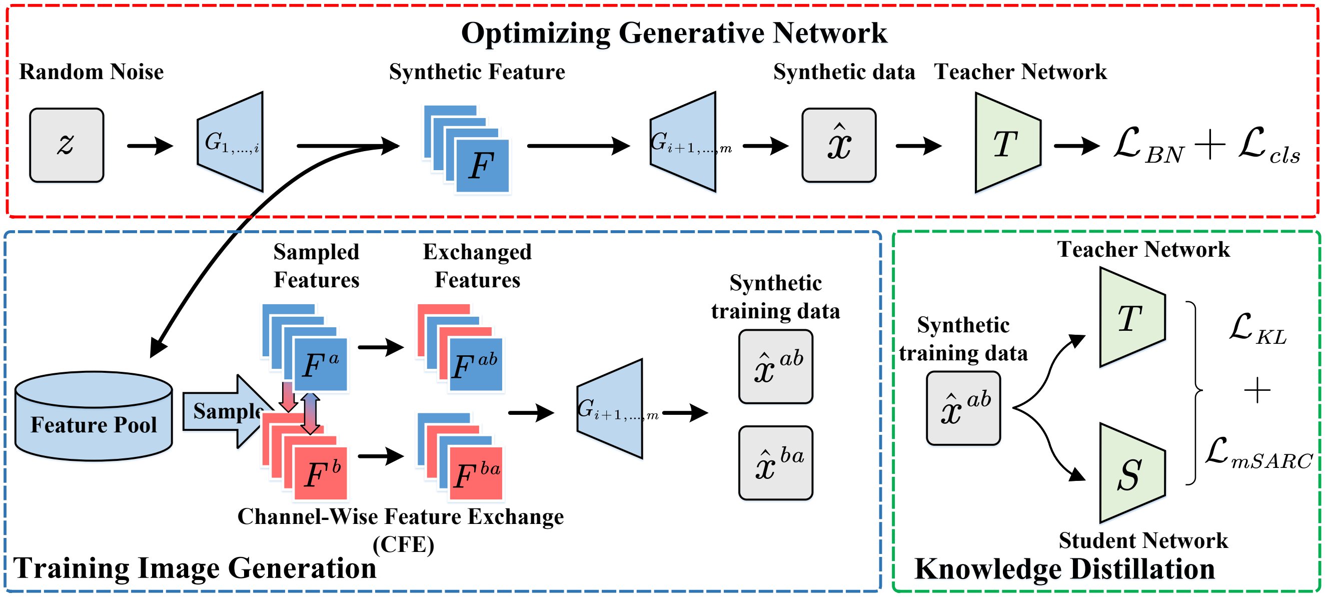


Figure 1. The framework of our data-free knowledge distillation using proposed CFE and mSARC. $G_{1,...,i}$ and $G_{i+1,...,m}$ represent the layers of the generative network G ranging from 1 to i and $i+1$ to m, respectively.

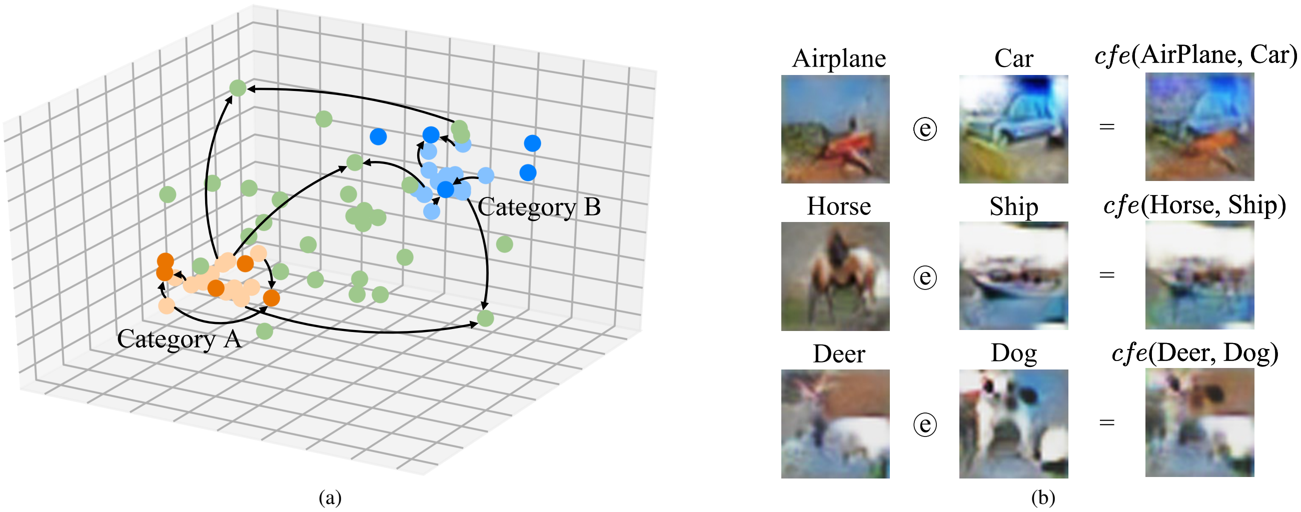


Figure 2. (a) Two basic cases of using channel-wise feature exchange (CFE) to increase the diversity of the synthesized images, i.e., increasing synthetic image diversity by performing CFE between intra-class samples and inter-class samples. (b) Image generation using channel-wise feature exchange, in which ⓔ denotes CFE. The first two images in each row are images obtained by passing the randomly sampled features from the feature pool through $G_{i+1,...,m}$. The last image of each row is obtained by performing channel-wise feature exchange on the sampled features and then feeding the exchanged feature to $G_{i+1,...,m}$.

$y \in \{1, 2, ...K\}$ is a random class label sampled from a uniform distribution. We feed the noise z into the generative network G and obtain $\hat{x} = G(z)$, in which a cross-entropy loss $\mathcal{L}_{cls}$ between the generated image $\hat{x}$ and the label y is used to force the generated image $\hat{x}$ be classified into a specific class by the teacher network T. In addition to using the cross-entropy loss $\mathcal{L}_{cls}$, we also use BN regularization $\mathcal{L}_{BN}$ [30, 44], a commonly used loss in DFKD, to constrain the mean and variance of the feature at the batch normalization layer to be consistent with the running-mean and running-variance of the same layer. After training the generative network G with $\mathcal{L}_{cls}$ and $\mathcal{L}_{BN}$, the synthetic data is expected to approximate the distribution of the original data.

3.3. Training Image Generation using Channel-wise Feature Exchange

As shown in the top of Fig. 1, we use the generative network G to generate synthetic images to be used for distillation learning. Unlike previous DFKD methods, which store synthetic images [7, 10, 13, 27, 29, 43], we use a feature pool to store the hidden features of the synthetic images when optimizing G and use them to improve the diversity of later training image generation. We propose a channel-wise feature exchange (CFE) to make use of the stored hidden features to improve the diversity of training image generation.

Specifically, to generate diverse synthetic images for training the student network S, we first randomly sample the features from the feature pool, e.g., F_i^a and F_i^b, in which a and b are indexes of different synthetic images when optimizing G, and the stored feature is from layer i of G. After sampling the features, we perform channel-wise feature exchange by randomly swapping half of the feature channels between F_i^a and F_i^b to obtain two mixed features, i.e., $F_i^{ab} = cfe(F_i^a, F_i^b)$, and $F_i^{ba} = cfe(F_i^b, F_i^a)$ (see bottom left part of Fig. 1). Next, we feed the mixed features F_i^{ab} and F_i^{ba} to the deep layers of G ranging from $i+1$ to m. Finally, we can obtain two synthetic images that differ from the conventional ones, i.e., $\hat{x}_i^{ab} = G_{i+1,...,m}(F_i^{ab})$ and $\hat{x}_i^{ba} = G_{i+1,...,m}(F_i^{ba})$. F_i^{ab} and F_i^{ba} can be expected to expand the diversity than the original features, resulting in generated images being more diverse than those produced by a conventional generation model. Fig. 2a illustrates how CFE can improve the synthetic image diversity by obtaining more dense and diverse sampling in the latent space. Fig. 2b shows several examples of synthetic images from the features of F^a, F^b, as well as F^{ab}. We can see that CFE does improve data diversity in image generation.

3.4. Knowledge Distillation using Multi-scale Spatial Activation Region Consistency Constraint

Given the synthetic training images $\hat{x}$ (such as $\hat{x}^{ab}$ and $\hat{x}^{ba}$), we minimize the Kullback-Leibler (KL) divergence between the logits of the student network S and the logits of the teacher network T using $\mathcal{L}_{KL}$.

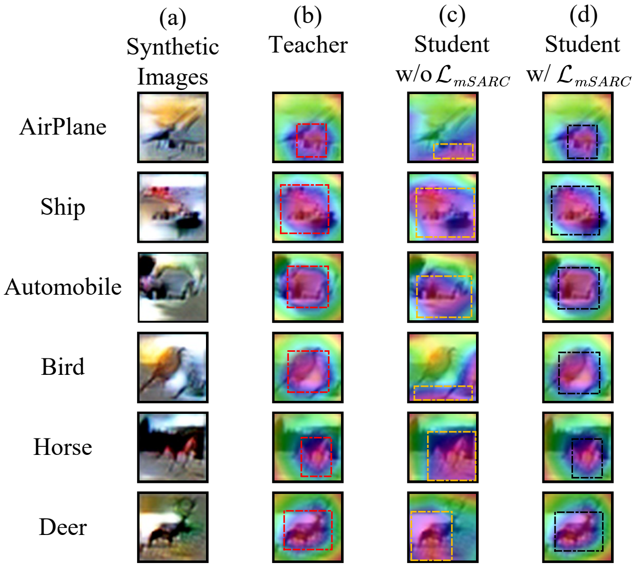


Figure 3. Comparisons of the CAMs of the teacher and student networks on synthetic images with and without using our mSARC: (a) synthetic images, (b) CAMs of the teacher network, (c) CAMs of the student network with KL loss alone, and (d) CAMs of the student network with both KL loss and our mSARC. (b) and (c) show big inconsistencies in learning visual cues for classification when only KL loss is used. (b) and (d) become much close to each other when using both KL loss and our mSARC loss. All displayed images are correctly classified by these three networks.

Algorithm 1 Proposed Method for DFKD.

```
Input: Pre-trained teacher network $T$ and student network $S$.
Output: An optimized student network $S$.
Initialize a generative network $G$;
for epoch=0,1,...,max_epoch do
    Sample a minibatch of noise $z$;
    for iteration=0,1,...,max_g_iterations do
        Generate synthetic data $\hat{x} = G(z)$;
        Optimize generative network $G$ using $\mathcal{L}_{cls}$ and $\mathcal{L}_{BN}$;
    Store the features after the $i_1$-th, $i_2$-th, and $i_n$-th layers into the feature pool;
    for iteration=0,1,...,max_kd_iterations do
        Sample features $F_i^1, F_i^2, ..., F_i^n$ from the feature pool;
        Perform CFE on $F_i^1, F_i^2, ..., F_i^n$ to obtain new features $F_i^{1*}, F_i^{2*}, ..., F_i^{n*}$ with probability $p$;
        Obtain synthetic training data $\hat{x}$ from $G_{i+1,...,m}(F_i^*)$ and feed them to $S$ and $T$;
        Optimize student network $S$ using $\mathcal{L}_{KL}$ and $\mathcal{L}_{mSARC}$;
return the student network $S$.
```

However, if we only use the conventional KL divergence constraint between the student and teacher network, we notice the two networks may learn different image cues even though their final predicted labels are the same. Fig. 3 gives such an example, in which the teacher network learns cues from red rectangle in the image, but the student network under KL loss may learn cues from the orange rectangle of the same image. This can lead to "shortcut learning" of the model, and reduce the student network's generalization ability in unseen domains. Therefore, we introduce a multi-scale spatial activation region consistency (mSARC) constraint, to encourage the hidden layers at different stages of the student network S to "focus on" the same spatial regions as the hidden layers of the teacher network T at the corresponding stages. Specifically, we constrain the class activation maps (CAMs) [51] of the feature at layer i_k of the student network S to be consistent with CAMs of the

Dataset	Teacher	Student	Test accuracy (%)								
			T.	**S.**	**DAFL** [7]	**ZSKT** [29]	**ADI** [43]	**DFQ** [10]	**LS-GDFD** [27]	**CMI** [13]	**SpaceshipNet**
CIFAR-10	ResNet-34	ResNet-18	95.70	95.20	92.22	93.32*	93.26	94.61	95.02	94.84 (94.39*)	95.39
	VGG-11	ResNet-18	92.25	95.20	81.10	89.46	90.36	90.84	N/A	91.13 (90.93*)	92.27
	WRN-40-2	WRN-16-1	94.87	91.12	65.71	83.74	83.04	86.14	N/A	90.01 (89.27*)	90.38
	WRN-40-2	WRN-40-1	94.87	93.94	81.33	86.07	86.85	91.69	N/A	92.78 (92.08*)	93.56
	WRN-40-2	WRN-16-2	94.87	93.95	81.55	89.66	89.72	92.01	N/A	92.52 (92.00*)	93.25
CIFAR-100	ResNet-34	ResNet-18	78.05	77.10	74.47	67.74	61.32	77.01	77.02	77.04 (74.25*)	77.41
	VGG-11	ResNet-18	71.32	77.10	57.29	34.72	54.13	68.32	N/A	70.56 (68.45*)	71.41
	WRN-40-2	WRN-16-1	75.83	65.31	22.50	30.15	53.77	54.77	N/A	57.91 (57.44*)	58.06
	WRN-40-2	WRN-40-1	75.83	72.19	34.66	29.73	61.33	61.92	N/A	68.88 (65.33*)	68.78
	WRN-40-2	WRN-16-2	75.83	73.56	40.00	28.44	61.34	59.01	N/A	68.75 (66.09*)	69.95
Tiny-ImageNet	ResNet-34	ResNet-18	66.44	64.87	N/A	N/A	N/A	63.73	N/A	64.01	64.04

Table 1. DFKD results on CIFAR-10, CIFAR-100, and Tiny-ImageNet. 'T.' and 'S.' denote that the teacher and student network trained using the labeled data in the training set, respectively, and this applies to other tables below. The results of DAFL, ZSKT, ADI, DFQ, and LS-GDFD are from [13]. The results marked by "*" comes from running the code of [13]* under the setting used in their paper.

feature at layer j_k in the teacher network T. The constraint can be formulated as

$$\mathcal{L}_{mSARC} = \mathrm{E}_{\hat{x}\sim p(\hat{x})} \sum_{k=1}^{t} ||CAM_{i_k}(S,\hat{x}) - CAM_{j_k}(T,\hat{x})||_2^2, \tag{1}$$

where $k = 1, 2, ..., t$, i_k and j_k denotes the layer index, and $CAM(\cdot,\cdot)$ is computed using [33, 51] and a detailed description can be found in the supplementary material. As shown in Fig. 3, after using our mSARC, the CAMs of the student and teacher networks become much close with other. Such a property is important for avoiding the student network to learn non-inherent cues for classification.

3.5. Overall Loss

The overall loss function for optimizing the generative network G can be expressed as

$$\mathcal{L}_G = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{BN}\mathcal{L}_{BN}, \tag{2}$$

in which λ_{cls} and λ_{BN} are parameters balancing two loss terms. The overall loss function for distillation learning can be expressed as

$$\mathcal{L}_{KD} = \lambda_{KL}\mathcal{L}_{KL} + \lambda_{mSARC}\mathcal{L}_{mSARC}. \tag{3}$$

in which λ_{KL} and λ_{mSARC} are parameters balancing two loss terms. Specific coefficient settings and more training details can be found in the supplemental materials. The whole algorithm of the proposed DFKD method via CFE and mSARC is given in Alg. 1.

4. Experiments

4.1. Settings

We evaluate our method using several different backbone networks, i.e., ResNet [15], VGG [36], and Wide ResNet [48] on three classification datasets, including CIFAR-10 [19], CIFAR-100 [19], and Tiny-ImageNet [21]. For CIFAR-10 and CIFAR-100, each contains 60,000 images, of which 50,000 images are used for training, and 10,000 images are used for testing. CIFAR-10 and CIFAR-100 contain 10 and 100 categories, respectively, and the images in both datasets have a resolution of 32×32. Tiny-ImageNet contains 100,000 training images and 10,000 validation images, with a resolution of 64 × 64. The dataset has 200 image categories.

4.2. Comparisons with SOTA DFKD methods

Table 1 shows the DFKD results by our method and several state-of-the-art (SOTA) methods, i.e., DAFL [7], ZSKT [29], ADI [43], DFQ [10], LS-GDFD [27], and CMI [13]) when using the same teacher network. ADI is a noise image optimization-based method, while all other methods are generative network-based methods, including the proposed method. ADI optimizes images slowly and has difficulty generating large amounts of data with limited computational resources. In contrast, generative network-based methods may lack sample diversity because the generated samples can be highly correlated and limited in variation. LS-GDFD and CMI only partially address the diversity issues by using multiple generators. By using CFE instead of hundreds of networks, our approach provides highly diverse training images. In addition, we use mSARC that is important to our disllation learning when using CFE. Using our proposed DFKD method, the accuracy on the CIFAR-10 test set of ResNet-18 even surpasses the ResNet-18 trained by simple supervised learning methods using labeled data, i.e., the CIFAR-10 training set.

4.3. Additional Experiments at Higher Resolution

To evaluate our method on datasets with higher image resolutions (here, we use 224 × 224), we conduct experiments on Imagenette and ImageNet100. Imagenette[*] is a subset of 10 easily classified classes from ImageNet [11]. ImageNet100[†] is a subset of 100 random classes from ImageNet-1k dataset. Since the existing complete open-source SOTA methods do not report results on ImageNet or

[*] https://github.com/fastai/imagenette

[†] https://www.kaggle.com/datasets/ambityga/imagenet100

Dataset	Test accuracy (%)			
	T.	CMI [13]	ZSKT [29]	SpaceshipNet
Imagenette	81.30	77.55	41.91	80.59
ImageNet100	70.74	52.60	6.54	65.40

Table 2. DFKD results on high-resolution datasets (Imagenette and ImageNet100) when using ResNet-34 as the teacher network and ResNet-18 as the student network.

its subsets in their papers, we use two open-sourced methods CMI and ZSKT for comparison. As shown in Table 2, when evaluated on high-resolution datasets, CMI and ZSKT-trained student networks perform significantly worse than the teacher network, while the student network trained by our method only has a small performance gap with the teacher network. This suggests our method generalizes well to DFKD with high-resolution images.

4.4. Ablation Study

To investigate the effectiveness of CFE and mSARC, we conduct several ablation studies.

How Important is CFE? We show the effect of CFE by discarding it during training. Results of (b) and (c) in Table 3 show that the accuracy degrades from 77.41% to 62.05% when dropping CFE. This result shows that CFE is helpful for improving distillation learning when using mSARC. Replacing CFE with CutMix and Mixup leads to an accuracy of 76.47% (Table 3 (d)) and 75.72% (Table 3 (e)), which are higher than the result without using any diversity method (62.05%) but slightly lower than our results. This suggests that CFE, CutMix, and Mixup are all helpful for improving the diversity of the synthetic images, but CFE may be more suitable for DFKD when using mSARC.

How important is mSARC? We verify the effectiveness of our mSARC from two aspects. We first discard mSARC while still using CFE, CutMix, and Mixup. The results in Table 3 (f, g, h) show discarding mSARC will greatly harm the performance when using CutMix, Mixup, and CFE. This is possibly because using these diversity methods alone can severely amplify the unwanted noises produced in the synthesized images, thereby limiting their effectiveness in improving distillation learning. Next, we compare our mSARC with the feature-level constraint used in several feature-level constraints used in recent KD methods [16, 18, 29, 32, 49]. These are strong constraints that force the feature or attention map of the student and teacher to be consistent, whereas mSARC only constrains them to focus on the same spatial region. Replacing mSARC with AT [49] and FitNets [32] lead to accuracy of 36.09% and 25.81% (Table 3 (i,j)), both of them are lower than our method. We conjecture that this is because when using CFE to generate training images, the noise in the synthetic data is amplified to the extent that these strong feature-level constraints cannot be performed properly. The above results not only illustrate the importance of mSARC for improving the robustness of DFKD, but also show that it is not feasible to replace our mSARC by simply using the common feature-level constraint.

Method	Diversity method	Distillation constraint	Teacher: 78.05%
			Student (%)
(a)	None	None	60.17
(b)	CFE	mSARC	77.41
(c)	None	mSARC	62.05
(d)	CutMix	mSARC	76.47
(e)	Mixup	mSARC	75.72
(f)	CFE	None	48.18
(g)	CutMix	None	65.24
(h)	Mixup	None	61.40
(i)	CFE	AT	36.09
(j)	CFE	FitNets	25.81

Table 3. Ablation experiments about CFE and mSARC in our method on CIFAR-100 when using ResNet-34 as the teacher network and ResNet-18 as the student network.

5. Further Analysis

5.1. Impact of mSARC on CAMs

In the ablation study above, we show that mSARC is able to improve the test accuracy of the student network. In this section, we show that $\mathcal{L}_{mSARC}$ promotes the student network to focus on the same spatial region with the teacher network. We compute the CAM difference between the teacher network and the student networks trained with and without $\mathcal{L}_{mSARC}$ for the 10,000 images in the CIFAR-100 test set. The results are shown in Table 4. As can be observed, the CAM difference between the teacher network and the student network drastically declines after using $\mathcal{L}_{mSARC}$. We also compute the CAM differences for the 3,309 images in the CIFAR-100 test set that can be correctly classified to their ground-truth category by both the student networks trained by our method and our method without using $\mathcal{L}_{mSARC}$. The result are shown in Table 4. The CAM difference still drastically declines after using $\mathcal{L}_{mSARC}$. These results positively indicate that mSARC can effectively reduce the CAM difference between the student and teacher network. In other words, it promotes the student network to learn discriminative cues from the same spatial region with the teacher network. More details and results on CIFAR-10 are given in the supplementary material.

Method	(a)		(b)	
	MAE	MSE	MAE	MSE
SpaceshipNet	23.40	2419.44	18.77	1579.34
SpaceshipNet w/o $\mathcal{L}_{mSARC}$	56.04	8892.64	46.78	6744.43

Table 4. Difference of CAMs between the teacher network T and the student networks S trained by our method with and without using $\mathcal{L}_{mSARC}$ for (a) the complete 10,000 images in the CIFAR-100 test set and (b) a subset of 3,309 images in the CIFAR-100 test set that can be correctly classified into their ground-truth category by both student networks trained by our method with and without using $\mathcal{L}_{mSARC}$.

Fraction of swapped channels from F_i^a *	Test accuracy (%)
0% (w/o CFE)	62.05
10%	77.04
30%	77.11
50%	77.41

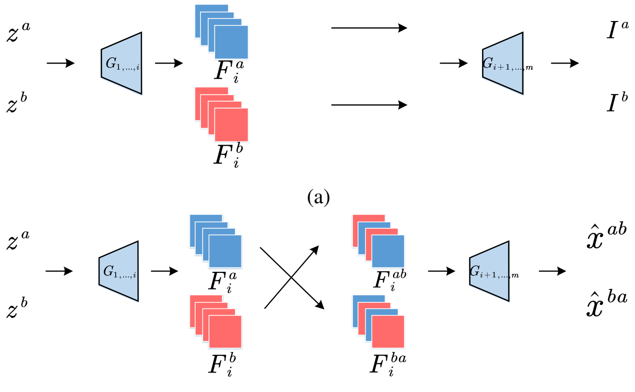
Table 5. The influence of performing CFE with different fractions of the swapped channels. Columns marked with * represent the proportion of channels in the feature F_i^{ab} that are from F_i^a, when performing $F_i^{ab} = cfe(F_i^a, F_i^b)$. In other words, they are the fractions of channels in F_i^a used for performing $cfe(F_i^a, F_i^b)$.

5.2. Impact of Fraction of Swapped Channels

In this section, we examine the impact of the number of swapped channels in CFE on the performance of our method. In previous experiments, we set the fraction of the swapped feature channels as 0.5. Here, we consider different fractions of swapped channels in features F_i^a, i.e., 0, 0.1, and 0.3. All the other settings remain the same with Section 4.2. The results on CIFAR-100 are shown in Table 5. As can be observed, the proposed CFE can stably improve the test accuracy using various fractions of swapped channels (i.e., 77.04%, 77.11%, and 77.41% vs. 62.05%). When 10%-90% of channels of a feature are used for swapping, the accuracy of the trained student network remains quite stable, i.e., ranging from 77.04% to 77.41%. The results indicate that the proposed CFE is effective for DFKD while not sensitive to different fractions of exchanged channels.

5.3. Why CFE Improves DFKD?

When we have n noises $z^1, z^2, ..., z^n$, we can get n images $I^1 = G(z^1), I^2 = G(z^2), ..., I^n = G(z^n)$ after feeding z to the layers of generative network G, and then have n corresponding features $F_i^1 = G_{1,...,i}(z^1), F_i^2 = G_{1,...,i}(z^2), ..., F_i^n = G_{1,...,i}(z^n)$ from the i-th layer of generative network G. Assume each feature F_i^k contains nc channels. Fig. 4a and Fig. 4b show the pipelines of generating images I from noise z without and with using CFE, respectively. Theoretically, when we perform CFE between F_i^p and F_i^q $(p + q = n + 1)$, the upper bound of the number of features we can obtain is $C_{nc}^{\frac{nc}{2}}$. If we feed these $C_{nc}^{\frac{nc}{2}}$ features to the $(i + 1)$-th to the m-th layers of the generative network G, we can get $C_{nc}^{\frac{nc}{2}}$ different images. These images constitute the image sampling space used to train the student network. The number of samples in this space is $C_{nc}^{\frac{nc}{2}}$, which far exceeds the size of the image sampling space without using CFE (i.e., n). Therefore using CFE can improve the result of data-free knowledge distillation.



(a)

(b)

Figure 4. The pipeline of training image generation (a) without using CFE and (b) using CFE.

6. Conclusion

Although tremendous progress has been made in data-free knowledge distillation (DFKD), existing DFKD methods still have limitations in diverse and efficient data synthesis due to the limitations of their image generation methods. In this work, we propose a novel DFKD method (SpaceshipNet) based on channel-wise feature exchange (CFE) and multi-scale spatial activation region consistency (mSARC) constraint to address these issues. Specifically, CFE allows our generative network to better sample from the feature space and efficiently generate diverse images for learning the student network. However, we found that using strong data augmentation methods (e.g., the commonly used CutMix and MixUp or our CFE) alone can severely amplify the unwanted noises generated in the synthesized images, and degrade distillation learning. Hence, we propose mSARC to assure the student network can imitate not only the logit output but also the spatial activation region of the teacher network to alleviate the influence of unwanted noises in diverse synthetic images on distillation learning. Combined with our mSARC, using traditional data augmentation on synthetic data can still significantly improve distillation learning. We obtained the best results on all three datasets commonly used for DFKD and validated our method on high-resolution datasets (two subsets of ImageNet).

References

[1] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017. 1

[2] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 1

[3] Himanshu S Bhatt, Samarth Bharadwaj, Richa Singh, and Mayank Vatsa. Memetically optimized mcwld for matching sketches with digital face images. *IEEE Trans. Inf. Forensics Security.*, 7(5):1522–1535, 2012. 1

[4] Kuluhan Binici, Shivam Aggarwal, Nam Trung Pham, Karianto Leman, and Tulika Mitra. Robust and resource-efficient data-free knowledge distillation by generative pseudo replay. *AAAI*, 2022. 1, 2

[5] Andrea Borghesi and Roberto Maroldi. Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression. *La radiologia medica*, 125:509–513, 2020. 1

[6] Xu Cao, HuanXin Zou, XinYi Ying, RunLin Li, ShiTian He, and Fei Cheng. Channelmix: A mixed sample data augmentation strategy for image classification. In *ICSP*, pages 269–274. IEEE, 2021. 3

[7] 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 *ICCV*, pages 3513–3521, 2019. 1, 5, 6

[8] 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 *ICCV*, pages 3514–3522, 2019. 1, 2

[9] Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *CVPRW*, pages 710–711, 2020. 1, 2

[10] Yoojin Choi, Jihwan P. Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *CVPR*, pages 3047–3057, 2020. 1, 5, 6

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2, 6

[12] Gongfan Fang, Kanya Mo, Xinchao Wang, Jie Song, Shitao Bei, Haofei Zhang, and Mingli Song. Up to 100x faster data-free knowledge distillation. *arXiv preprint arXiv:2112.06253*, 2021. 1, 2

[13] Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. *IJCAI*, 2021. 1, 2, 5, 6, 7

[14] Zhiwei Hao, Yong Luo, Zhi Wang, Han Hu, and Jianping An. Cdfkd-mfs: Collaborative data-free knowledge distillation via multi-level feature sharing. *IEEE Transactions on Multimedia*, 24:4262–4274, 2022. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[16] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 3, 7

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 3

[18] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *NeurIPS*, 31, 2018. 3, 7

[19] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 2, 6

[20] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, 2019. 1

[21] Ya Le and Xuan Yang. Tiny imagenet visual recongnition challenge. *Technical Report*, 2015. 2, 6

[22] Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Xin Dong, Fengwei Yu, Shaoqing Lu, and Shi Gu. Mixmix: All you need for data-free compression are feature and data mixing. In *ICCV*, pages 4410–4419, 2021. 1

[23] Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Xin Dong, Fengwei Yu, Shaoqing Lu, and Shi Gu. Mixmix: All you need for data-free compression are feature and data mixing. In *ICCV*, pages 4410–4419, October 2021. 3

[24] Jit Yan Lim, Kian Ming Lim, Shih Yin Ooi, and Chin Poo Lee. Efficient-prototypicalnet with self knowledge distillation for few-shot learning. *Neurocomputing*, 459:327–337, 2021. 1

[25] Yuang Liu, Wei Zhang, Jun Wang, and Jianyong Wang. Data-free knowledge transfer: A survey. *arXiv preprint arXiv:2112.15278*, 2021. 1

[26] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017. 1, 2

[27] Liangchen Luo, Mark Sandler, Zi Lin, Andrey Zhmoginov, and Andrew Howard. Large-scale generative data-free distillation. *arXiv preprint arXiv:2012.05578*, 2020. 1, 2, 5, 6

[28] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 1

[29] Paul Micaelli and Amos J. Storkey. Zero-shot knowledge transfer via adversarial belief matching. In *NeurIPS*, pages 9547–9557, 2019. 1, 2, 3, 5, 6, 7

[30] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *ICCV*, pages 1325–1334, 2019. 4

[31] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *ICML*, pages 4743–4751. PMLR, 2019. 1, 2

[32] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 3, 7

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 6

[34] Chengchao Shen, Xinchao Wang, Youtan Yin, Jie Song, Sihui Luo, and Mingli Song. Progressive network grafting for few-shot knowledge distillation. *AAAI*, 35(3):2541–2549, May 2021. 1

[35] Alberto Signoroni, Mattia Savardi, Sergio Benini, Nicola Adami, Riccardo Leonardi, Paolo Gibellini, Filippo Vaccher, Marco Ravanelli, Andrea Borghesi, Roberto Maroldi, and Davide Farina. Bs-net: learning covid-19 pneumonia severity on a large chest x-ray dataset. *Medical Image Analysis*, page 102046, 2021. 1

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[37] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR*, pages 2443–2449, 2021. 1

[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 3

[39] Cecilia Summers and Michael J Dinneen. Improved mixed-example data augmentation. In *WACV*, pages 1262–1270. IEEE, 2019. 3

[40] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852, 2017. 1

[41] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Ricap: Random image cropping and patching data augmentation for deep cnns. In *ACML*, pages 786–798. PMLR, 2018. 3

[42] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, pages 6438–6447. PMLR, 2019. 3

[43] 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 deep-inversion. In *CVPR*, pages 8712–8721, 2020. 1, 5, 6

[44] 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 deep-inversion. In *CVPR*, pages 8715–8724, 2020. 1, 2, 4

[45] Jaemin Yoo, Minyong Cho, Taebum Kim, and U Kang. Knowledge extraction with no observable data. *NeurIPS*, 32, 2019. 1, 2

[46] Xinyi Yu, Ling Yan, Yang Yang, Libo Zhou, and Linlin Ou. Conditional generative data-free knowledge distillation. *Image and Vision Computing*, page 104627, 2023. 1

[47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 2, 3

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 6

[49] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 3, 7

[50] H Zhang, M Cisse, Y Dauphin, and D Lopez-Paz. mixup: Beyond empirical risk management. In *ICLR*, pages 1–13, 2018. 2, 3

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 5, 6