

Dual Discriminator Adversarial Distillation for Data-free Model Compression

Haoran Zhao ¹ · Xin Sun ^{1,2} · Junyu Dong ¹ · Milos Manic ³ · Huiyu Zhou ⁴ · Hui Yu ⁵

Abstract Knowledge distillation has been widely used to produce portable and efficient neural networks which can be well applied on edge devices for computer vision tasks. However, almost all top-performing knowledge distillation methods need to access the original training data, which usually has a huge size and is often unavailable. To tackle this problem, we propose a novel data-free approach in this paper, named Dual Discriminator Adversarial Distillation (DDAD) to distill a neural network without the need of any training data or meta-data. To be specific, we use a generator to create samples through dual discriminator adversarial distillation, which mimics the original training data. The generator not only uses the pre-trained teacher’s intrinsic statistics in existing batch normalization layers but also obtains the maximum discrepancy from the student model. Then the generated samples are used to train the compact student network under the supervision of the teacher. The proposed method obtains an efficient student network which closely approximates its teacher network, without using the original training data. Extensive experiments are conducted to demonstrate the effectiveness of the proposed approach on CIFAR, Caltech101 and ImageNet datasets for classification tasks. Moreover, we extend our method to semantic segmentation tasks on several public datasets such as CamVid, NYUv2, Cityscapes and VOC 2012. To the best

Xin Sun
E-mail: sunxin@ouc.edu.cn
Junyu Dong
E-mail: dongjunyu@ouc.edu.cn

- ¹ College of Information Science and Engineering, Ocean University of China, Qingdao, PR China
- ² Department of Aerospace and Geodesy, Technical University of Munich, Germany
- ³ College of Engineering, Virginia Commonwealth University, US
- ⁴ School of Informatics, University of Leicester, UK
- ⁵ School of Creative Technologies, University of Portsmouth, Portsmouth, UK

of our knowledge, this is the first work on generative model based data-free knowledge distillation on large-scale datasets such as ImageNet, Cityscapes and VOC 2012. Experiments show that our method outperforms all baselines for data-free knowledge distillation.

Keywords Deep neural networks · image classification · model compression · knowledge distillation · data-free

1 Introduction

Deep convolutional neural networks (CNNs) have attained state-of-the-art performance in various computer vision tasks, e.g., image classification [1–4], object detection [5–8] and semantic segmentation [9–13]. Whereas, along with high-performance, CNNs with tens of millions of parameters, hundreds of layers often require heavy computation and storage. Thus, one prominent obstacle is that it is impossible to deploy such complicated CNNs into resource-constrained devices, such as mobile-phone and autonomous cars.

To this end, a large body of works have been published [14–19], which intend to compress and speed-up the cumbersome CNNs into more lightweight ones. The popular compression techniques are divided into five categories: quantization [20, 21], low rank factorization, pruning [22, 23] and sharing weights, compact convolutional filter, and knowledge distillation [24]. For example, Guo et al. [23] propose a novel pruning framework which can compress the cumbersome model by pruning the channels and automatically decide the network structure after channel pruning. Kang et al. [25] introduce a novel pruning framework that reflects accelerator architectures. Liu et al. [26] propose to compress a compact student by distilling the cross-modal representations of the pre-trained multi-modal teacher. Among these methods, knowledge distillation [24] is one of the most popular paradigms for learning a portable student model from the pre-trained complicated teacher by directly imitating its outputs.

Knowledge distillation (KD) is proposed by Hinton et al. [24] for supervising the training of a compact yet efficient student model by capturing and transferring the knowledge of a large teacher model to a compact one. Note that, the student model pays more attention to the probability correlation between categories rather than the one-hot labels by softening the outputs of the teacher in softmax. Further researches propose various kinds of distillation loss functions in the field of CNNs optimization. For example, FitNets [27] introduces the hints loss to guide the training of intermediate layers of the student network. AT [28] transfers the attention features of intermediate layers from the teacher network to student network. FSP [29] designs the flow distillation loss to force the student model to mimic flow matrices of the teacher model among the feature maps between two layers. RKD [30] transfers mutual relations of data examples by the distance-wise and angle-wise distillation losses. SP [31] preserves the pairwise similarities in student’s representation space instead to mimic the representation space of the teacher.

The aforementioned knowledge distillation approaches have achieved high efficiency and accuracy on public datasets when the original training dataset can be obtained. However, we cannot access the original training dataset in many scenarios such as security and privacy. Thus, we are not able to directly use these existing knowledge distillation approaches for distilling a compact student model.

Nevertheless, a few works [32–38] have been proposed to obtain the portable student model by knowledge distillation without the need of any original training data. For example, Chen et al. [32] and Micaelli et al. [33] conduct pilot research on data-free knowledge distillation. In Chen’s work [32], the one-hot samples are generated to replace original samples, which can highly activate the neurons of the teacher network. Micaelli et al. [33] propose to search images that poorly match the teacher network. Besides, Lopes et al. [35] utilize the layer spectral activations or the layer activations from the teacher model to reconstruct the training samples. Nayak et al. [34] propose to model the softmax space using the parameters of the teacher model to transfer data. However, these methods only take account of information from the student but ignores the valuable information of the pre-trained teacher model. Besides, there has been another perspective [37] which starts with an arbitrary input and is iteratively updated via back-propagation. Nayak et al. [34] synthesize the training samples named Data Impressions via modelling the data distribution in the softmax space. Another two related works [37, 39] extend the Inceptionism method [40] with some regularizers. Although the generated images from these approaches are realistic, they leave the student information constraint-free.

In summary, works mentioned above have obtained impressive results on classification tasks but still suffer from several limitations. One kind of data-free knowledge distillation methods, which exploit the generative adversarial networks, only aim to train the generator for synthesizing images that could fool the student. Another kind of works which update the arbitrary input via back-propagation, only take the pre-trained teacher model into account but ignore the information from the student. The generated samples obtained by existing works cannot well balance the realistic of images and the generalization of student model. Thus, these generated images are not very useful for distilling the compact student model efficiently.

In this work, we propose a novel data-free framework to fully utilize both the intrinsic statistics from the pre-trained teacher model and the information from the student model. We train an adversarial generator through dual discriminator adversarial distillation to produce images, which mimic the training data distribution. It could also improve the generalization of the student model under the supervision of the teacher. To be specific, we use two discriminators to jointly train the generator. One fixed discriminator is the individual pre-trained teacher model which recovers the training data distribution from the existing teacher model. We utilize the valuable knowledge that encoded in the batch normalization layers which is widely used in high-performance CNNs such as ResNet and DenseNet. The other is consisted of both the teacher

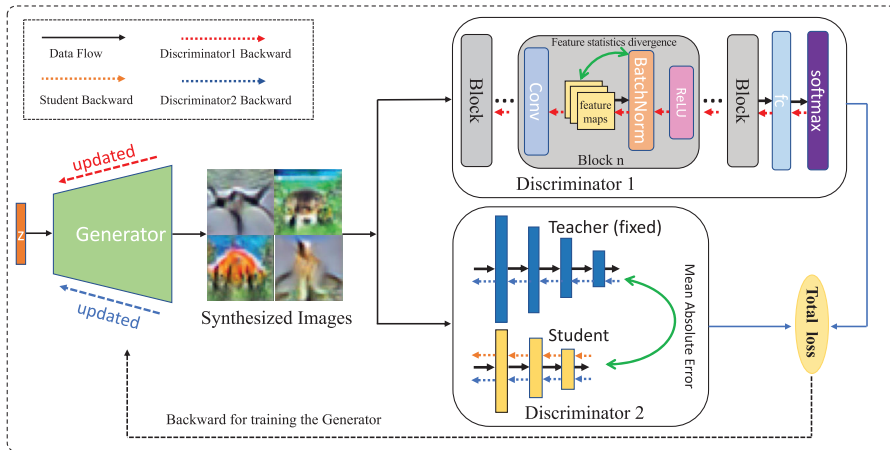


Fig. 1 The overall framework of the proposed dual discriminator adversarial distillation method. The training procedure is divided into two stage. First, we train a generator to produce samples instead of the original training data. Two different discriminators are employed for training the generator. Second, we use the generated images to transfer the knowledge from the teacher model to the student model.

and student model which is customized for the student to produce samples. It could improve the student model’s generalization. The main contributions in this paper are summarized as follows:

- We introduce a novel data-free knowledge distillation framework by leveraging a two-stage dual discriminator adversarial distillation approach, which trains the portable and efficient student model without the need of training data.
- The proposed DDAD method not only utilizes the pre-trained teacher’s intrinsic statistics in existing batch normalization layers but also obtains the maximum discrepancy from the student model.
- Extensive experiments demonstrate the effectiveness of the proposed method on several classification and semantic segmentation benchmarks. To the best of our knowledge, this is the first work for generative based data-free knowledge distillation on large-scale datasets such as ImageNet, Cityscapes and VOC 2012.

The remaining part of this paper is organized as follows. Section 2 reviews related work and Section 3 presents our dual discriminator adversarial distillation architecture. After experimental results are presented in Section 4, we conclude this paper in Section 5.

2 Related Works

In this section, we briefly review some recent researches which are closely related to our approach. First, we present existing knowledge distillation meth-

ods which are divided into two categories based on the requirements of data, i.e., data-driven knowledge distillation and data-free knowledge distillation. Then, we review some mainstream methods about generative adversarial networks.

2.1 Data-driven Knowledge Distillation

Caruana et al. [41] first confirm that one ensemble of networks could transfer the knowledge to the single network. Then Ba et al. [42] propose to teach the student network by penalizing the difference of logits between the teacher and student. Later, the concept of Knowledge Distillation (KD) is introduced by Hinton et al. [24] to solve model compression problems. It transfers the dark knowledge from the pre-trained teacher network to the compact student network by mimicking the class probabilities outputs, which are softened by setting a temperature hyperparameter in softmax. Then the performance of compact student can be retained and close to the performance of the teacher. Note that, KD requires original training data to capture the valuable knowledge from the teacher network.

However, the knowledge contained in the soft-labels is insufficient when the teacher network goes deeper. To tackle this issue, some improvements have been made, which extend KD by utilizing the intermediate representation as supervision. For example, Fitnets [27] forces the student to learn the similar intermediate features as teacher’s which are defined as hints. Zagoruyko et al. [28] define the attention maps from the intermediate features as the knowledge and then obtain a better performance compared to the one using original feature itself. Moreover, FSP [29] designs the flow distillation loss to force the student to mimic flow matrices of teacher among the feature maps between two layers. RKD [30] transfers mutual relations of data examples by the distance-wise and angle-wise distillation losses. SP [31] preserves the pairwise similarities in student’s representation space instead to mimic the representation space of the teacher. CTKD [43] combines the knowledge from different teacher models to improve the student’s performance in KD. Due to excellent performance, knowledge distillation has been used to solve a variety of complex applications such as object detection [26, 44], semantic segmentation [45], lane detection [46], face recognition [47–49] and action recognition [50].

Nevertheless, the above traditional data-driven knowledge distillation methods need full of original training data, which are difficult to be obtained in real world. Thus, several few-shot knowledge distillation approaches are proposed to distill knowledge from teacher to student using few samples. Li et al. [51] propose to compress the teacher network by combining block-wise distillation with network pruning using few training data. Bai et al. [52] introduce a few-shot cross distillation method which employs network pruning to reduce the layer-wise accumulated errors. The work of [53] introduces an approach which combines mixup and active learning to distill knowledge for the student network using a small number of examples as possible. Although achieving

promising results, these methods also need some original training samples and still cannot be used when the original training dataset is unavailable. Therefore, the data-free methods need to be studied for knowledge distillation.

2.2 Data-free Knowledge Distillation

Although most of the existing works highly rely on original training data, a few attempts have been made to distill the compact student when original training data is not accessible [39]. One kind of methods employs the generative adversarial networks to synthesize samples, which are used to distill knowledge from the teacher network to the student network. For example, Chen et al. [32] propose to employ a generator to approximate the data distribution of training dataset for supervising the training of student without original training dataset. Micaelli et al. [33] also utilize a simple adversarial fashion to train a generator for searching images on which the student network poorly matches the teacher network. Then the generated samples are used for training the student. Yoo et al. [36] propose to generate the artificial data points as original training data by training the generator and decoder networks.

However, the methods mentioned above only take the information from the student into account, but ignore the valuable information of the pre-trained teacher model. That means the generated images from the generator are customized for the student instead of considering intrinsic statistics contained in the pre-teacher network. Thus, there has been another kind of methods for data-free knowledge distillation which obtains artificial data by updating random noise images. These methods optimize the random inputs by some regularizers while the pre-trained teacher is fixed during the training phase. For example, Nayak et al. [34] synthesize the training samples named Data Impressions via modelling the data distribution in the softmax space. Yin et al. [37] optimize the random inputs to obtain synthesize class-conditional input images by inverting a pre-trained teacher network. During training, the knowledge contained in the batch normalization layers of the teacher is utilized for regularizing the distribution of intermediate feature maps. Haroush et al. [39] extend the Inceptionism method [40], which is a feature visualization approach based on the pre-trained model, and generate samples according to low order statistics captured by existing batch normalization layers. Unfortunately, though the images from these approaches are realistic, these methods leave the student information constraint-free. That means the generator do not know what images the student network need in the distilling procedure. Moreover, the images synthesizing phase and knowledge distillation process are separated into two stages among these methods.

2.3 Adversarial Distillation

Adversarial learning attracts more and more attention due to its effectiveness in generative adversarial networks (GANs) [54]. It is used to capture the

underlying distribution of real data sets by a generative model. Inspired by this, some adversarial knowledge distillation methods employ GANs to further promote the student model to perfectly mimic the teacher model.

Chen et al. [32] propose to utilize an adversarial generator to synthesize fake samples as training data. Moreover, Liu et al. [55] employ the generator to augment the training dataset. Micaelli et al. [33] further distill knowledge by generated samples from an adversarial generator. Different from the above methods, in which the teacher network is frozen during the training of the student network, Chung et al. [56] propose to jointly optimize the teacher and student in each iteration. Therefore, combining GAN and KD is an effective way to improve the student performance in KD and overcome the limitations of unaccessible data. Different from existing works, our proposed dual discriminator adversarial distillation (DDAD) employs two different discriminators to train a generator, which can produce more efficient and diversity samples for distilling knowledge from the teacher model to the student model.

3 Proposed Method

In this section, we will describe our DDAD method for data-free model compression in detail. As Figure 1 shown, we train the generator network by employing two different discriminators in the first stage. Then we use the samples generated from the generator network to train the compact student under the supervision of teacher network. Note that the pre-trained teacher network is regarded as one discriminator, which utilizes the intrinsic statistics of existing batch normalization layer from the CNNs. The other discriminator is consisted of the teacher and the student.

3.1 Teacher-Student Learning Paradigm

We define the teacher model as T with parameters W_t , and the student model as S with parameters W_s . Furthermore, z_t and z_s are logits of teacher and student. There has been a temperature hyperparameter τ in vanilla knowledge distillation [24] for softening the output of the teacher model. So we define $P_t^\tau = \text{softmax}(z_t/\tau)$ and $P_s^\tau = \text{softmax}(z_s/\tau)$ as the softened output predictions of the teacher and student respectively. Then the student network is trained to mimic the behavior of the teacher network by optimizing the following loss function:

$$\mathcal{L}_{KD}(W_s) = \mathcal{H}(y_{true}, P_s) + \lambda KL(P_t^\tau, P_s^\tau) \quad (1)$$

where \mathcal{H} is the cross-entropy loss between the true labels and outputs of student network, KL is the Kullback-Leibler divergence between the softened outputs of teacher and student models. λ is hyper parameter that adjust the balance between these two terms.

3.2 Training of Generator Network

Although GANs have the powerful ability to generate images. However, we cannot directly use vanilla GANs to generate training samples due to the absence of original ground truth in data-free scenario. Thus, we introduce a novel dual discriminator adversarial distillation to produce training samples instead of original training data. In vanilla GANs [54], the input noise vector z is mapped to the desired data x by the Generator G , i.e. $G : z \rightarrow x$, while the Discriminator D is used for distinguishing the real samples from generating samples G as following:

$$\mathcal{L}_{GAN} = \mathbb{E}_{y \sim p_{data}(y)}[\log D(y)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (2)$$

When discriminator D cannot classify the image as being generated from generator G or from p_{data} , the two player min-max optimization converges.

Furthermore, D2GAN [57] employs two discriminators together with a generator to tackle the problem of mode collapse in GAN following the objective function:

$$\begin{aligned} \mathcal{L} = & \alpha \mathbb{E}_{y \sim p_{data}(y)}[\log D_1(y)] + \mathbb{E}_{z \sim p_z(z)}[-D_1(G(z))] \\ & + \mathbb{E}_{y \sim p_{data}(y)}[-D_2(z)] + \beta \mathbb{E}_{z \sim p_z(z)}[\log(D_2(G(z)))] \end{aligned} \quad (3)$$

By Eq. 3, the D_1 , D_2 and G are trained alternatively. However, the discriminators of D2GAN require real training data for training. Similar to D2GAN, there is also a generator and two discriminators in our framework. The generator \mathcal{G} is used to generate samples which satisfies the intrinsic statistics from the teacher and also could help the teacher to better transfer knowledge. The first discriminator \mathcal{D}_1 is the pre-trained teacher model itself with fixed parameters. And the other discriminator \mathcal{D}_2 is consisted of the teacher model and student model together.

For training the generator, we continuously optimize \mathcal{G} depending on \mathcal{D}_1 and \mathcal{D}_2 by the following function:

$$\begin{aligned} \mathcal{G} = & \arg \min_G \mathbb{E}_{z \sim p_z(z)}[\log(1 - \mathcal{D}_1(G(z)))] \\ & + \arg \min_G \mathbb{E}_{z \sim p_z(z)}[\log(1 - \mathcal{D}_2(G(z)))] \end{aligned} \quad (4)$$

The discriminator \mathcal{D}_1 encourages the generator to produce samples which could mimic the distribution of the original training data based on internal statistics' divergence. As we all known, almost all top-performance CNNs employ the Batch-Normalization (BN) layers which are beneficial to the performance and speed of convergence. When we feed the samples produced by the generator to \mathcal{D}_1 , all BN layers calculate the running means and running variances of its inputs. Thus, we employ \mathcal{D}_1 to force the BN layers from the teacher model to mimic original means and variances stored in pre-trained teacher.

More formally, we refer the running means and variances as $\hat{\mu}$ and $\hat{\sigma}$, respectively. And the history means and variances stored in pre-trained teacher are defined as $\tilde{\mu}$ and $\tilde{\sigma}$. Then we assume that the internal statistics follow the Gaussian distribution.

$$\mathcal{L}_{\mathcal{D}_1} = \mathbb{E}_{z \sim p_z(z)} KL(\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2) || \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)) \quad (5)$$

In addition, we utilize the pre-trained teacher model, together with the student model to act as the discriminator \mathcal{D}_2 .

$$\mathcal{L}_{\mathcal{D}_2} = -\mathbb{E}_{z \sim p_z(z)} \left[\frac{1}{n} \|P_t^\tau(G(z)) - P_s^\tau(G(z))\|_1 \right] \quad (6)$$

Note that, we use the negative MAE loss for this objective function. By Eq. 6, we force G to produce hard samples which produce the maximum discrepancy between teacher and student networks. It means that the student model do not master these samples which need to be learned. In other words, \mathcal{D}_2 is employed to customize samples for student network and get better distillation performance.

By combining the aforementioned two loss functions, i.e., Eq. 5 and 6, we obtain the final objective function for the generator:

$$\mathcal{L}_G = \delta \mathcal{L}_{\mathcal{D}_1} + \gamma \mathcal{L}_{\mathcal{D}_2} \quad (7)$$

where δ, γ are hyperparameters that adjust the balance between the loss terms. By minimizing the loss function in Eq. 7, we train and optimize the generator. We aim to use the generator to generate samples which could mimic the data distribution of the original training data when the training loss converge.

3.3 Training of the Student Network

In this stage, we utilize the samples produced from the generator G to train the student under the supervision of the teacher model. Due to the absence of real data, the student could not get the useful knowledge from the teacher model by the soft targets. Thus, we use the MAE loss instead of KL in Eq. 1. And the loss function for training the student is as follows:

$$\mathcal{L}_S = \mathbb{E}_{z \sim p_z(z)} \left[\frac{1}{n} \|P_t^\tau(G(z)) - P_s^\tau(G(z))\|_1 \right] \quad (8)$$

Note that, we only update the student S and fix the generator G on this phase. The generated samples are from the generator G with a batch of random noise z from Gaussian distribution.

This phase is similar to vanilla KD which distills the valuable knowledge from teacher network to student network without hesitation due to the access to real data. However, the original training data is not available in our data-free scenario. Due to the absence of ground truth labels, we cannot use the first term in Eq. 1. Thus, the generated samples by the generator are vital for our

Algorithm 1 Dual Discriminator Adversarial Distillation

Input: A pre-trained teacher model T with parameters W_t .

Output: A compact student model S with parameters W_s .

- 1: Initialization: The generator G , the student network S and training hyper-parameters.
 - 2: **Repeat:**
 - 3: **Stage 1: Train the Generator.**
 - 4: Generate vector z as one batch randomly.
 - 5: Synthesize samples as training images by $x \leftarrow G(z)$.
 - 6: Compute the loss function \mathcal{L}_G in Eq. 7.
 - 7: Updating the weights in G by back-propagation.
 - 8: **Stage 2: Train the student model.**
 - 9: Generate vector z as one batch randomly.
 - 10: Generate samples x from z with $G(z)$.
 - 11: Compute the softened output of teacher network and
 - 12: student network: $P_t^r(x)$, $P_s^r(x)$
 - 13: Compute the loss function for KD in Eq. 1.
 - 14: Updating the weights by the gradient in W_s .
 - 15: **Until:** converge.
-

data-free distillation performance. As aforementioned, we train the generator in dual discriminator adversarial manner to produce samples which could be well used for distilling knowledge.

3.4 Optimization

The complete data-free distillation algorithm for our two-stage adversarial training is summarized in Algorithm 1. Similar to GANs, the whole training procedure can be divided into two stages.

In the first stage, we fix the discriminators and only update the generator. To be specific, we employ two different discriminators to jointly train the generator G , which could generate samples mimicking the original training data distribution.

Note that Eq. 7 is the total loss function for optimizing the generator G . The first term of \mathcal{L}_G is used for discriminator D_1 , which forces G to generate samples satisfying the intrinsic statistics of teacher model. The second term in Eq. 7 is for discriminator D_2 consisted with the teacher model and student model. It take into account what kind of generated samples are most useful for distilling the student model.

In the second stage, we fix the generator G and only update the student S in discriminators. Note that, the student S , together with the teacher T jointly consist of our discriminator D_2 . Thus, the weights of discriminator D_2 will be updated in this stage. In addition, the pre-trained teacher network has the ability to extract semantic features, since it has been trained on the original

dataset. Thus, we employ the pre-trained teacher as the fixed discriminator D_1 instead of training a new discriminator. We froze its weights in this stage.

Therefore, we optimize the student model with parameters W_s by the KD loss in Eq. 1. Then we distill the compact student model using generated samples instead of the original training data. Our proposed method maintains diversity of generated images due to the dual discriminator adversarial distillation manner.

4 Experiments and Results

In the following sections, we conduct extensive experiments to verify the effectiveness of DDAD method. We implement our approach with Pytorch on NVIDIA 2080Ti GPUs. For baselines, we adopt the vanilla KD [24] method and several very recent data-free knowledge distillation works that achieve strong performance, including two non-generative based methods ZSKD [34] and BNS [39], and some generative based data-free methods, i.e., ZSKT [33], KEGNET [36], DAFL [32], DFAD [38], DFQ [58] and CMI [59]. For example, DAFL proposes to generate images that highly activate the neurons of pre-trained teacher network, and DFAD obtains a compact student network by optimizing the upper bound of the discrepancy between the student and the teacher.

4.1 Datasets.

CIFAR. The CIFAR-10 and CIFAR-100 datasets contain 60K RGB images from 10 and 100 categories, where each class has 6K and 600 images respectively. The training set contains 50K images and rest of the 10K images are used for testing set. Due to complex and advanced of ResNet [60], we employ ResNet-34, ResNet-18 as the teacher network and student network respectively as suggested in [32].

Caltech101. The Caltech101 dataset contains images of objects belonging to 101 classes, which are collected for classification. Each class contains about 50 images. We randomly divide the dataset into training set and testing set, which has 6982 images and 1695 images respectively. Note that we randomly resize and crop the images to 128×128 for the training procedure.

ImageNet. The ImageNet dataset [61] consists of 1.28M training images with 1000 categories and the rest of 50K images is used as the testing set. We explore the full resolution of 224×224 for training and testing procedures.

CamVid. The CamVid dataset [62] is an automotive dataset for road scene segmentation. It contains 367 training and 233 testing RGB images. We use the 11 categories in CamVid such as building, tree, sky, car, road, etc. and ignore the 12th categories contains unlabeled data. We train the teacher network using the images with 128×128 random crop.

NYUv2. The NYUv2 dataset [63] is consisted of indoor scenes images for indoor segmentation. We use the 1449 labeled RGB-D images with 13 categories in NYUv2 and also crop the images to 128×128 resolution.

Cityscapes. The Cityscapes dataset [64] collects 5000 images in street scenes from 50 different cities, where 19 semantic categories are used for evaluation. It contains high quality pixel-level finely annotated images with 1024×2048 pixels. These images are divided into training/validation/testing with 2,975/500/1,525 images respectively. Training images for teacher are randomly cropped into 512×512 .

Pascal VOC. The Pascal VOC dataset [65] contains 20 foreground objects categories and an extra background category. These images are divided into 1,464/1,449/1,456 groups for training/validation/testing respectively. We train the teacher network with a crop size of 512×512 .

4.2 Experimental Settings

Network architecture. For classification experiments, we adopt the ResNet-34 for the teacher model, the ResNet-18 for the student model, which stacks the basic residual blocks to achieve the state-of-the-art performance. It is complex and advanced to be adopted for investigating the effectiveness of our approach. For semantic segmentation experiments, the DeepLabV3 model is adopted as teacher with an ImageNet pre-trained ResNet-50 as backbone. Furthermore, we adopt the Mobilenet-V2 as backbone and train the students from scratch.

Moreover, we adopt two different generator networks for classification and semantic segmentation tasks. Specifically, we employ the generator, which uses nearest neighbor interpolation for up sampling on CIFAR dataset. For other datasets, we use a strong generator proposed in [66], which uses the deconvolutions instead of the interpolations. In our experiments, we set the learning rate as $1e-3$ and optimize the generators with Adam.

Implementation details. Firstly, we train the teacher networks via the standard back-propagation. We adopt ResNet-34 as the teacher network and train it for 200 epochs using SGD with momentum 0.9. We set the weight decay to $5e-4$. We use $\delta = 0.01$, $\gamma = 0.1$ and set the initial learning rate to 0.1 and then divide it by 10 at 80 and 120 epochs for CIFAR datasets. As for Caltech101, we use the same training hyperparameters and set the learning rate to 0.05 and divided by every 100 epochs. Secondly, we train student and generator networks using the proposed method. Specifically, we update the student network using SGD with momentum 0.9 and weight decay $5e-4$. For generator networks, we use Adam with a learning rate of $1e-3$. The student and generator networks are trained for 500 epochs on CIFAR dataset, and 300 epochs for other datasets. For ImageNet dataset, we set the hyperparameters $\delta = 0.01$, $\gamma = 0.5$. For semantic segmentation tasks, we use the same training hyperparameters and modify the learning rate to 0.05 and reduce the weight decay to $5e-5$.

Note that, we use the pre-trained ResNet-34 as the teacher network for ImageNet. Due to the mode collapse problem in such large-scale dataset, we employ multiple generators to alleviate this issue. To be specific, we employ 10 generators, each of which is trained to produce 100 classes. In another word, all the classes are divided into 100 groups, i.e., classes 1–100 in the first group, classed 101–200 in the second group, etc. We sample 1000 images per group in a data-free manner, which uses Equation 5 as the optimization objective [39] to get several batches of images, to estimate the per-group statistics for generator training.

Evaluation Metrics. For classification experiments, we take the accuracy of prediction for the metric and present the model size by the number of network parameters. We use different seeds for 3 times and adopt the median of classification accuracy as the final results. For segmentation tasks, the Intersection over Union (IoU) means the ratio of interval and union between the ground truth and the predicted segmentation for each category. In our experiments, we take the mean Intersection over Union (mIoU) for the segmentation metric. Furthermore, we calculate the sum of floating point operations (FLOPs) as the model complexity for one forward on a fixed input size.

4.3 Experimental Results

4.3.1 CIFAR and Caltech101

We first conduct experiments on CIFAR dataset to demonstrate the effectiveness of DDAD approach. For CIFAR dataset, we set the batch size to 256 and train the generator and student models from scratch for 500 epochs, which are randomly initialized.

The performance of classification on CIFAR-10, CIFAR-100 datasets are concluded in Table 1. We get the pre-trained teacher network that achieves the 95.54%, 77.50% accuracy on CIFAR-10 and CIFAR-100 respectively using standard back-propagation. However, the student network only achieves a 95.12%, 76.53% accuracy on CIFAR-10 and CIFAR-100 when trained individually, which are obviously lower than those of the pre-trained teacher network. Then we train the student network using the vanilla knowledge distillation, which improves the classification accuracy of the student model from 95.12%, 76.53% to 95.43% 76.87% on CIFAR-10 and CIFAR-100 respectively using the original training data.

The rest rows of Table 1 reports the classification results of several recent data-free knowledge distillation approaches. These methods could be divided into two kinds, i.e., non-generative based data-free methods and generative based data-free methods. We compare our approach with the state-of-the-art data-free methods using their released codes. First, we compare our method with two non-generative based methods ZSKD [34] and BNS [39]. As can be seen from Table 1, ZSKD and BNS achieve test accuracies of 91.61% and 93.02% on CIFAR-10, 70.21% and 73.25% on CIFAR-100 respectively. The

Table 1 Classification accuracy (%) on CIFAR-10, CIFAR-100, Caltech101 and ImageNet datasets (3 runs). We use the standard back-propagation to train the student individually. DDAD means the performance of the student model from our method.

Algorithm	Network	Required data	CIFAR-10	CIFAR-100	Caltech101	ImageNet
Student	ResNet-18	✓	95.12%	76.53%	75.69%	69.64%
Teacher	ResNet-34	✓	95.54%	77.50%	76.60%	73.27%
Knowledge Distillation	ResNet-18	✓	95.43%	76.87%	76.12%	70.87%
Zero-Shot KD (ZSKD) [34]	ResNet-18	–	91.61%	70.21%	69.01%	–
Knowledge within (BNS) [39]	ResNet-18	–	93.02%	73.25%	72.71%	51.32%
Zero-Shot Transfer (ZSKT) [33]	ResNet-18	–	85.95%	66.29%	65.82%	–
KEGNET [36]	ResNet-18	–	91.83%	73.91%	72.29%	–
Data-Free Learning (DAFL) [32]	ResNet-18	–	92.22%	74.47%	–	–
Data-free Adversarial (DFAD) [38]	ResNet-18	–	93.30%	69.43%	73.50%	20.63%
Data-Free Quantization (DFQ) [58]	ResNet-18	–	93.26%	67.01%	73.61%	–
Contrastive Inversion (CMI) [59]	ResNet-18	–	94.08%	74.01%	73.92%	–
Our DDAD	ResNet-18	–	94.81%	75.04%	75.01%	59.84%

other generative based data-free methods, i.e., ZSKT [33], KEGNET [36], DAFL [32], DFAD [38], DFQ [58] and CMI [59], achieve the test accuracies of 85.95%, 91.83%, 92.22%, 93.30%, 93.26% and 94.08% on CIFAR-10 dataset, 66.29%, 73.91%, 74.47%, 69.43%, 67.01% and 74.01% on CIFAR-100 dataset. The results of our proposed DDAD method are shown in the bottom of Table 1. We obtain the classification accuracy of 94.81% on CIFAR-10 dataset, which is 2.59% and 1.51% better than DAFL’s and DFAD’s performance. This is very close to 95.43% which is obtained in KD using the original training data. For CIFAR-100, the accuracy of our method is 75.04%, exceeding the DAFL and DFAD by 0.57% and 5.61%.

Furthermore, we conduct experiments on the complex Caltech101 dataset which has 101 classes and images of size 128×128 . The teacher network obtains the accuracy of 76.60% and the student network achieves the accuracy of 75.69% when trained individually. DFAD achieves only 73.50% accuracy without using any original training data and DAFL fails in this case. Our approach gets the highest accuracy among these data-free methods. The student network trained using our method achieves 75.01% accuracy, which is comparable with that of the pre-trained teacher network.

4.3.2 ImageNet

To further investigate the effectiveness of our method, we extend our method to ImageNet dataset. We set the batch size to 128 and train the ensemble of 10 generators and student network from scratch for 100K steps.

We present the testing accuracy of classification on ImageNet in Table 1. The pre-trained weights from torchvision are used for the teacher network which achieves an accuracy of 73.27%. However, the student network only achieves a 69.64% accuracy when trained individually. We first use the vanilla knowledge distillation based on data-driven to train the student, which outperforms the student trained individually by 1.23%. Then we compare our method with the state-of-the-art data-free approaches. To the best of our knowledge, there are few previous methods that work for data-free distillation on ImageNet, especially for generative based methods. We compare our method with

BNS [39] and DFAD [38], which belong to non-generative and generative based data-free distillation methods respectively. Note that, the other state-of-the-art baselines fail on such large-scale dataset.

As can be seen from Table 1, the student of the proposed DDAD method achieves an accuracy of 59.84%, which outperforms the ones trained by BNS and DFAD by 8.52% and 39.21% respectively under the same steps and setting.

Visualization of Generated Samples. We present a set of generated samples by applying our method to a CIFAR-10/CIFAR-100 pre-trained ResNet-34 in Figure 2. Note that, these images are generated from the generators under our framework. Comparing with the generated samples on CIFAR-10 and CIFAR-100, we can find that the generator on CIFAR-100 produces more complicated samples than that on CIFAR-10. As the difficulty of classification increases, the teacher model becomes more knowledgeable so that, in adversarial learning, the generator can recover more complicated images. Furthermore, our method produces some images features and textures, e.g., the flowers, airplanes and cars as presented in the right of Figure 2. Meanwhile, we could observe that the generated images are almost different from each other. We owe the perfect image diversity to the dual discriminators used in our framework.

The left of Figure 3 shows the generated samples of size 128×128 by applying the proposed approach to a Caltech-101 pre-trained ResNet-34. And the right side of Figure 3 shows the real samples from the original training dataset, which are used to train the ResNet-34 teacher network. As can be seen from Figure 3, the generated images, which could not be recognized by humans, significantly improve the performance of student model. Hence, it implies that the realistic images are vital but not the indispensable way for knowledge distillation.

We also present the samples produced from the generators using the proposed DDAD method in Figure 4. Figure 4 shows that the images produced by the generators of our method seem to be less closer to the original training images, but indeed have abundant knowledge for the distillation task. In other words, due to lack of original training images, the generators of our DDAD method are unable to produce the images which can fool humans' eyes as real ones, but they could produce the real samples by the deep neural networks' perception. It also indicates the gap between deep neural networks and the human.

4.3.3 CamVid and NYUv2

In addition to verifying our method on the above described classification task, we also extend the proposed approach to semantic segmentation tasks. We use the resolution of 128×128 cropped images for CamVid and NYUv2 in the following experiments. The teacher model is initialized by the pre-trained ResNet-50 on ImageNet dataset. Furthermore, we train the student network from scratch using the generated samples with the resolution of 128×128 . The batch size is set to 32 in our experiments. For CamVid dataset, we use



CIFAR10



CIFAR100

Fig. 2 Generated samples on CIFAR-10 and CIFAR-100. Given only a pre-trained ResNet-34 teacher network, a randomly initialized ResNet-18 student network.

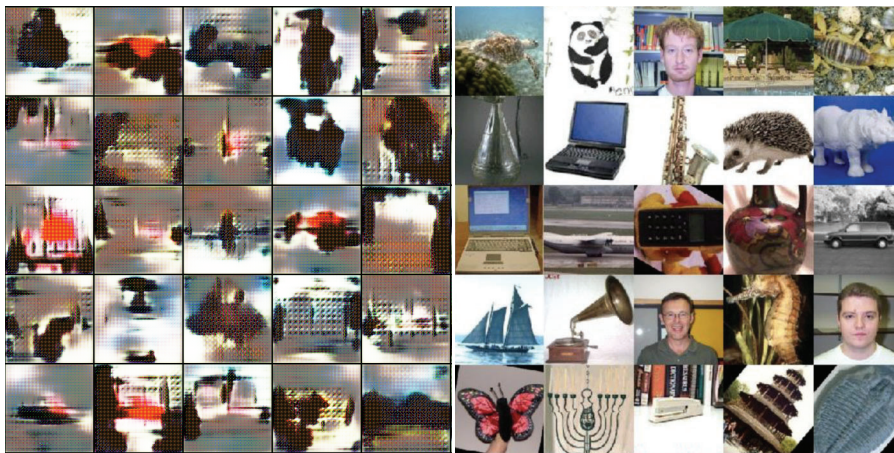


Fig. 3 Generated samples (left) of size 128×128 and real samples (right) from Caltech101. Given only a Caltech-101 pre-trained ResNet-34 teacher network and a randomly initialized ResNet-18 student network.



Fig. 4 Generated samples of size 224×224 on ImageNet. Given only an ImageNet pre-trained ResNet-34 teacher network and a randomly initialized ResNet-18 student network.

Table 2 The mIoU of DeepLabv3 model on CamVid, NYUv2, Cityscapes and VOC 2012 datasets. The teacher model are pre-trained on ImageNet when the student are randomly initialized.

	CamVid		NYUv2		Cityscapes		VOC 2012	
Method	FLOPs	mIoU	FLOPs	mIoU	FLOPs	mIoU	FLOPs	mIoU
Teacher	41.0G	0.819	41.0G	0.731	51.4G	0.749	51.4G	0.720
ZSKT [33]	5.54G	0.210	5.54G	0.173	6.0G	0.063	6.0G	0.108
DAFL [32]	5.54G	0.010	5.54G	0.105	6.0G	0.301	6.0G	0.293
DFAD [38]	5.54G	0.494	5.54G	0.319	6.0G	0.394	6.0G	0.357
Ours	5.54G	0.518	5.54G	0.363	6.0G	0.424	6.0G	0.407

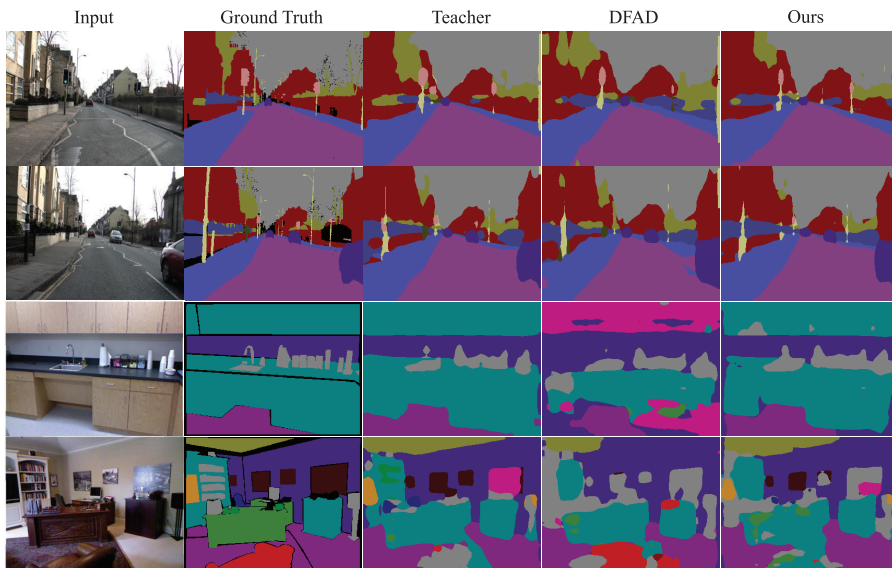


Fig. 5 Qualitative segmentation results on CamVid and NYUv2 produced from MobileNet-V2. Given only the ImageNet pre-trained ResNet-50, which is adopted as backbone. The segmentation results are obtained on generated samples. Our method achieves the best performance when the original training data is not available. Moreover, our method is comparable with that of the teacher network.

SGD with a learning rate of 0.1 and a weight decay of $5e-4$. The generator and student networks are trained for 300 epochs and the learning rate is decayed every 100 epochs. For NYUv2, we set the learning rate as 0.05 and we multiply it by 0.3 on 150 epochs and 250 epochs. We also train the generator and student networks for 300 epochs and the weight decay is reduced to $5e-5$.

Table 2 presents the results of the student model (Mobilenet-V2) on our framework and other state-of-the-art results. We assess the complexity by Flops that is computed on the resolution 128×128 . As can be seen from Table 2, our approach yields mIoU 51.8% on CamVid dataset, which boosts the

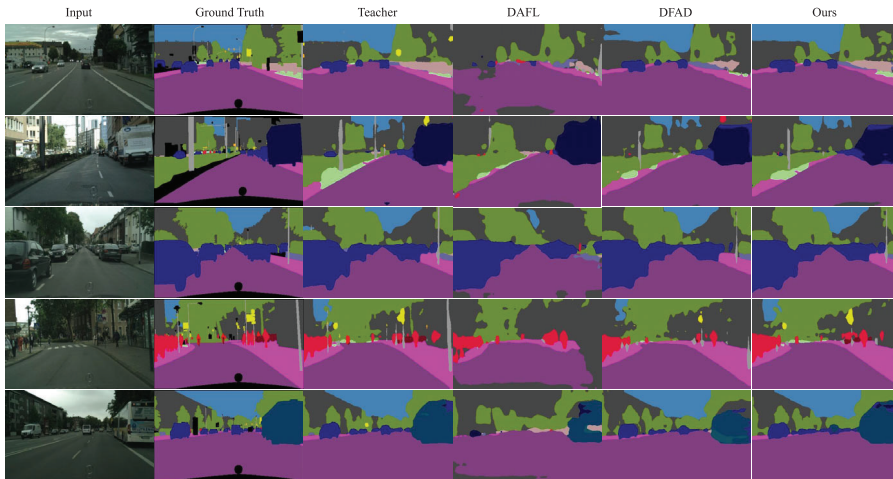


Fig. 6 Qualitative segmentation results on CityScapes produced from MobileNet-V2. Given only the ImageNet pre-trained ResNet-50, which is adopted as backbone. The segmentation results are obtained on generated samples. Our method achieves the best performance when the original training data is not available. Moreover, our method is comparable with that of the teacher network.

performance of DFAD by 2.4 point. For NYUv2 dataset, our method achieves 36.3% mIoU, which improves the performance of DFAD by 4.4%.

Furthermore, we conduct the visualization experiments to verify the effectiveness of our dual discriminator adversarial distillation method and show the qualitative segmentation results are shown in Figure 5. As can be seen from Figure 5, some objects such as traffic signs are clearer than those in DFAD [38] method and even close to the performance of teacher.

4.3.4 Cityscapes and VOC 2012

To further verify the effectiveness of our method, we extend our method to the large-scale and challenging semantic segmentation datasets, such as Cityscapes and VOC 2012. In the following experiments, we use the resolution of 512×512 cropped images for Cityscapes and VOC 2012. Furthermore, we train the student model from scratch using the images with the resolution of 512×512 produced by the trained generators.

We also list the results of the student network under our method and other state-of-the-art baselines in Table 2. The complexity is assessed by Flops that is computed on the resolution 512×512 . As can be seen, our approach achieves 42.4% mIoU on Cityscapes, which gains the improvements of DFAD, DAFL and ZSKT by 3%, 12.3% and 36.1%, respectively. For VOC dataset, our method yields mIoU 40.7%, which boosts the performance of DFAD, DAFL and ZSKT by 5%, 11.4% and 29.9%, respectively.

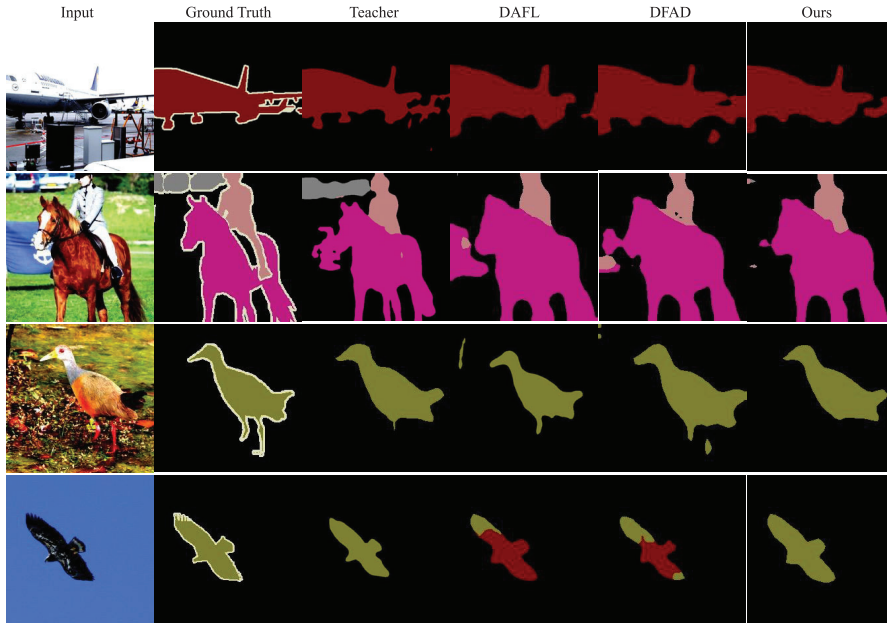


Fig. 7 Qualitative segmentation results on VOC 2012 produced from MobileNet-V2. Given only the ImageNet pre-trained ResNet-50, which is adopted as backbone. The segmentation results are obtained on generated samples. Our method achieves the best performance when the original training data is not available. Moreover, our method is comparable with that of the teacher network.

4.4 Ablation Study

Below we perform several extended experiments on CIFAR dataset to verify the performance of our DDAD method. To investigate the robustness of DDAD, we consider discussing the effectiveness of each discriminator in our framework and distill student network which has the same architecture as the teacher network using our approach.

The effectiveness of each discriminator. We look into the effect of each discriminator in our dual discriminator adversarial distillation framework. And we conduct the following experiments on CIFAR-100. And we employ the ResNet-34 as the teacher model, ResNet-18 as the student model. As can be seen from Eq. 7, δ and γ are hyperparameters that adjust the balance between two discriminators. Thus we range the hyperparameters δ and γ from 0.005 to 0.025, 0.01 to 0.2 respectively. The defaults of δ and γ are 0.01 and 0.1 respectively. When we change one of the hyperparameters, the other keeps the default.

Table 3 presents that the classification results of DDAD approach are affected by setting different hyper-parameters. As can be seen from Table 3, the accuracy performance is insensitive to the choice of γ over a large range from 0.05 to 0.15, 0.4 to 0.6 on CIFAR100/Caltech101 and CamVid datasets,

Table 3 Effects of hyper-parameters δ , γ in DDAD on CIFAR-100, Caltech101 and CamVid datasets. For classification task, we use the ResNet-34 for the teacher model, the ResNet-18 for the student model. For segmentation task, we adopt DeepLabV3 model as teacher with a pre-trained ResNet-50 as backbone, and MobileNet-V2 as student.

CIFAR100	δ						γ					
	0	0.005	0.01	0.015	0.02	0.025	0	0.01	0.05	0.1	0.15	0.2
Acc (%)	67.63	69.52	75.04	69.92	68.43	68.07	67.92	69.87	74.42	75.04	74.90	68.13
Caltech101	δ						γ					
	0	0.005	0.01	0.015	0.02	0.025	0	0.01	0.05	0.1	0.15	0.2
Acc (%)	68.01	72.39	75.01	73.17	70.68	68.71	68.63	70.37	73.91	75.01	73.70	69.92
CamVid	δ						γ					
	0	0.005	0.01	0.015	0.02	0.025	0	0.3	0.4	0.5	0.6	0.7
mIoU (%)	46.7	48.6	51.8	47.3	45.8	44.9	43.9	44.7	46.9	51.8	47.5	45.4

Table 4 The classification on several public datasets. Given only the pre-trained ResNet-34 teacher networks. Note that the student has the same network architecture as the teacher network.

Dataset	Network	Accuracy	
		Teacher	Student
CIFAR-10	ResNet-34	95.54	93.82
CIFAR-100	ResNet-34	77.50	75.79
Caltech101	ResNet-34	76.60	74.96

respectively. However, a carefully adjusted hyperparameters do bring the best performance. When we set one of the hyperparameters to 0, it means that we use the single discriminator in our framework. Experimental results show that both of the discriminators in our framework bring performance improvement.

To be specific, when we set the hyperparameter δ to 0 as shown in Table 3, it means only the discriminator D_2 is employed in our framework. In such a situation, the student network only achieves the accuracy of 67.63%, since the generator is trained only for synthesizing samples. It fools the student but ignores the intrinsic statistics of the pre-trained teacher. On the contrary, when we set γ to 0 in Table 3, it means that we only employ the discriminator D_1 in our framework. Note that, the discriminator D_1 is fixed during training in this situation. Thus, we only train and update the generator G in the whole training process. And the distillation process of the student is followed in the next stage. As presented in Table 3, the student only achieves the accuracy of 67.92 % in such a situation. In addition, it also can be time-consuming to distill the student network.

Difference between generated samples and real data. To further explore the effectiveness of our method, we compare the classification results between the generated samples and original training data. Thus we adopt the same network architecture of teacher as the student network in our framework. The performance is presented in Table 4.

As can be seen from Table 4, the ResNet-34 networks obtain a 95.54% accuracy and a 77.50% accuracy on CIFAR-10 and CIFAR-100 datasets, respectively. However, the student networks, which also use the same ResNet-34

networks as the teacher network, achieve the accuracy of 93.82% and 75.79%, respectively. Thus, we conclude that the performances of students are very close to those of the teacher models. For Caltech101 dataset the student also achieves very comparable performance to those of teacher networks. These results imply that the generated samples from our framework could well mimic the original training data and improve the performance of student network in knowledge distillation.

5 Conclusion

In this paper, we propose a novel data-free knowledge distillation framework to obtain a compact but efficient student model without the need of the original training data. This is achieved by designing a dual discriminator adversarial distillation approach in two stages. In the first stage, we employ two different discriminators to train the generator which could produce images that instead of the original training data. Moreover, because of the dual discriminator, the generated samples could mimic the original training data distribution. Here, we not only use the teacher model's intrinsic statistics but also get the maximum discrepancy on the student model to mimic the data distribution. In the second stage, we distill knowledge from the teacher to the student model with the generated samples from the first stage. Experimental results on several classification and segmentation benchmarks demonstrate that the proposed approach yields state-of-the-art data-free distillation performance, outperforming existing data-free knowledge distillation methods. In the future, we will explore the prior knowledge to further improve the diversity of synthetic images on the proposed dual discriminator adversarial distillation framework.

Acknowledgment

We thank supports of the National Natural Science Foundation of China (No. 61971388, U1706218, L1824025), and Key Natural Science Foundation of Shandong Province (No. ZR2018ZB0852).

Data availability statement

The datasets generated during and/or analysed during the current study are available in <https://github.com/ouc-ocean-group>.

References

1. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778

2. Li X, Wu J, Sun Z, Ma Z, Cao J, Xue JH (2021) Bsnet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing* 30:1318–1331
3. Li X, Li S, Omar B, Wu F, Li X (2021) Reskd: Residual-guided knowledge distillation. *IEEE Transactions on Image Processing* 30:4735–4746
4. Sun X, Xv H, Dong J, Zhou H, Chen C, Li Q (2021) Few-shot learning for domain-specific fine-grained image classification. *IEEE Transactions on Industrial Electronics* 68(4):3588–3598
5. Ren S, He K, Girshick RB, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*, Quebec, Canada, December 7-12, pp 91–99
6. Lou J, Wang Y, Nduka C, Hamed M, Mavridou I, Wang F, Yu H (2020) Realistic facial expression reconstruction for VR HMD users. *IEEE Trans Multim* 22(3):730–743
7. Feng M, Gilani SZ, Wang Y, Zhang L, Mian A (2021) Relation graph network for 3d object detection in point clouds. *IEEE Transactions on Image Processing* 30:92–107
8. Chen L, Jiang Z, Tong L, Liu Z, Zhao A, Zhang Q, Dong J, Zhou H (2020) Perceptual underwater image enhancement with deep learning and physical priors. *IEEE Transactions on Circuits and Systems for Video Technology* 31(8):3078–3092
9. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12*, pp 3431–3440
10. Ming Y, Meng X, Fan C, Yu H (2021) Deep learning for monocular depth estimation: A review. *Neurocomputing* 438:14–33
11. Chen C, Sun X, Hua Y, Dong J, Xv H (2020) Learning deep relations to promote saliency detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(07):10510–10517
12. Zhang Y, Sun X, Dong J, Chen C, Lv Q (2021) Gpnet: Gated pyramid network for semantic segmentation. *Pattern Recognition* 115:107940
13. Lv Q, Sun X, Chen C, Dong J, Zhou H (2021) Parallel complement network for real-time semantic segmentation of road scenes. *IEEE Transactions on Intelligent Transportation Systems* <https://doi.org/10.1109/TITS.2020.3044672>
14. Wu T, Tang S, Zhang R, Cao J, Zhang Y (2021) Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing* 30:1169–1179
15. Han X, Song X, Yao Y, Xu XS, Nie L (2020) Neural compatibility modeling with probabilistic knowledge distillation. *IEEE Transactions on Image Processing* 29:871–882
16. Bian C, Feng W, Wan L, Wang S (2021) Structural knowledge distillation for efficient skeleton-based action recognition. *IEEE Transactions on Image Processing* 30:2963–2976

17. Martínez-González A, Villamizar M, Canévet O, Odobez JM (2020) Efficient convolutional neural networks for depth-based multi-person pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology* 30(11):4207–4221
18. Zhao Y, Wang R, Chen Y, Jia W, Liu X, Gao W (2021) Lighter but efficient bit-depth expansion network. *IEEE Transactions on Circuits and Systems for Video Technology* 32(5):2063–2069
19. Tan DS, Lin Y, Hua K (2021) Incremental learning of multi-domain image-to-image translations. *IEEE Transactions on Circuits and Systems for Video Technology* 31(4):1526–1539
20. Gupta S, Agrawal A, Gopalakrishnan K, Narayanan P (2015) Deep learning with limited numerical precision. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML2015, Lille, France, 6-11 July*, pp 1737–1746
21. Zhai H, Lai S, Jin H, Qian X, Mei T (2021) Deep transfer hashing for image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 31(2):742–753
22. Han S, Pool J, Tran J, Dally WJ (2015) Learning both weights and connections for efficient neural network. In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, Quebec, Canada, December 7-12.*, pp 1135–1143
23. Guo J, Zhang W, Ouyang W, Xu D (2021) Model compression using progressive channel pruning. *IEEE Transactions on Circuits and Systems for Video Technology* 31(3):1114–1124
24. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. *Computer Science* 14(7):38–39
25. Kang HJ (2020) Accelerator-aware pruning for convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology* 30(7):2093–2103
26. Liu T, Lam KM, Zhao R, Qiu G (2021) Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology* <https://doi.org/10.1109/TCSVT.2021.3060162>
27. Romero A, Ballas N, Kahou SE, Chassang A, Bengio Y (2015) Fitnets: Hints for thin deep nets. *3rd International Conference on Learning Representations, ICLR*
28. Zagoruyko S, Komodakis N (2017) Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *5th International Conference on Learning Representations, ICLR*
29. Yim J, Joo D, Bae J, Kim J (2017) A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 7130–7138
30. Park W, Kim D, Lu Y, Cho M (2019) Relational knowledge distillation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 3967–3976

31. Tung F, Mori G (2019) Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1365–1374
32. Chen H, Wang Y, Xu C, Yang Z, Liu C, Shi B, Xu C, Xu C, Tian Q (2019) Data-free learning of student networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3514–3522
33. Micaelli P, Storkey AJ (2019) Zero-shot knowledge transfer via adversarial belief matching. In: Advances in Neural Information Processing Systems, pp 9551–9561
34. Nayak GK, Mopuri KR, Shaj V, Radhakrishnan VB, Chakraborty A (2019) Zero-shot knowledge distillation in deep networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, 9-15 June, pp 4743–4751
35. Lopes RG, Fenu S, Starner T (2017) Data-free knowledge distillation for deep neural networks. arXiv preprint: arXiv:1710.07535
36. Yoo J, Cho M, Kim T, Kang U (2019) Knowledge extraction with no observable data. In: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, December 8-14, pp 2701–2710
37. Yin H, Molchanov P, Alvarez JM, Li Z, Mallya A, Hoiem D, Jha NK, Kautz J (2020) Dreaming to distill: Data-free knowledge transfer via deepinversion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, pp 8712–8721
38. Fang G, Song J, Shen C, Wang X, Chen D, Song M (2019) Data-free adversarial distillation. arXiv preprint: arXiv:1912.11006
39. Haroush M, Hubara I, Hoffer E, Soudry D (2020) The knowledge within: Methods for data-free model compression. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, pp 8491–8499
40. Mordvintsev A, Olah C, Tyka M (2015) Inceptionism: Going deeper into neural networks. Tech. rep., Google AI, URL <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>
41. Bucila C, Caruana R, Niculescu-Mizil A (2006) Model compression. In: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, pp 535–541
42. Ba J, Caruana R (2014) Do deep nets really need to be deep? In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, December 8-13, pp 2654–2662
43. Zhao H, Sun X, Dong J, Chen C, Dong Z (2020) Highlight every step: Knowledge distillation via collaborative teaching. IEEE Transactions on Cybernetics <https://doi.org/10.1109/TCYB.2020.3007506>
44. Deng J, Pan Y, Yao T, Zhou W, Li H, Mei T (2019) Relation distillation networks for video object detection. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), Oc-

- tober 27 - November 2, pp 7022–7031
45. Liu Y, Chen K, Liu C, Qin Z, Luo Z, Wang J (2019) Structured knowledge distillation for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, pp 2604–2613
 46. Hou Y, Ma Z, Liu C, Loy CC (2019) Learning lightweight lane detection cnns by self attention distillation. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, pp 1013–1021
 47. Liu H, Zhu X, Lei Z, Cao D, Li SZ (2021) Fast adapting without forgetting for face recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 31(8):3093–3104
 48. Ge S, Zhao S, Li C, Zhang Y, Li J (2020) Efficient low-resolution face recognition via bridge distillation. *IEEE Transactions on Image Processing* 29:6898–6908
 49. Ge S, Zhao S, Li C, Li J (2019) Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing* 28(4):2051–2062
 50. Tang Y, Wei Y, Yu X, Lu J, Zhou J (2020) Graph interaction networks for relation transfer in human activity videos. *IEEE Transactions on Circuits and Systems for Video Technology* 30(9):2872–2886
 51. Li T, Li J, Liu Z, Zhang C (2020) Few sample knowledge distillation for efficient network compression. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, pp 14627–14635
 52. Bai H, Wu J, King I, Lyu MR (2020) Few shot network compression via cross distillation. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, pp 3203–3210
 53. Wang D, Li Y, Wang L, Gong B (2020) Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, pp 1495–1504
 54. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y (2014) Generative adversarial nets. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, Montreal, Quebec, Canada, December 8-13, pp 2672–2680
 55. Liu R, Fusi N, Mackey L (2018) Model compression with generative adversarial networks. arXiv preprint: arXiv:1812.02271
 56. Chung I, Park S, Kim J, Kwak N (2020) Feature-map-level online adversarial knowledge distillation. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, 13-18 July 2020, Virtual Event, pp 2006–2015

57. Nguyen TD, Le T, Vu H, Phung DQ (2017) Dual discriminator generative adversarial nets. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, December 4-9, 2017, Long Beach, CA, USA, pp 2670–2680
58. Choi Y, Choi JP, El-Khamy M, Lee J (2020) Data-free network quantization with adversarial knowledge distillation. arXiv preprint: arXiv:2005.04136
59. Fang G, Song J, Wang X, Shen C, Wang X, Song M (2021) Contrastive model inversion for data-free knowledge distillation. arXiv preprint: arXiv:2105.08584
60. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, June 27-30, pp 770–778
61. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein MS, Berg AC, Li F (2015) Imagenet large scale visual recognition challenge. *Int J Comput Vis* 115(3):211–252
62. Brostow GJ, Shotton J, Fauqueur J, Cipolla R (2008) Segmentation and recognition using structure from motion point clouds. In: *European conference on computer vision*, Springer, pp 44–57
63. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgb-d images. In: *European conference on computer vision*, Springer, pp 746–760
64. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, June 27-30, 2016, pp 3213–3223
65. Everingham M, Gool LV, Williams CKI, Winn JM, Zisserman A (2010) The pascal visual object classes (VOC) challenge. *Int J Comput Vis* 88(2):303–338
66. Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. In: *4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings