D^3K: Dynastic Data-Free Knowledge Distillation

Xiufang Li, Qigong Sun, Licheng Jiao, *Fellow, IEEE,* Fang Liu, *Senior Member, IEEE,* Xu Liu, *Member, IEEE,* Lingling Li, *Senior Member, IEEE,* Puhua Chen, *Member, IEEE,* Yi Zuo,

***Abstract*—Data-free knowledge distillation further broadens the applications of the distillation model. Nevertheless, the problem of providing diverse data with rich expression patterns needs to be further explored. In this paper, a novel dynastic data-free knowledge distillation (D^3K) model is proposed to alleviate this problem. In this model, a dynastic supernet generator (D-SG) with a flexible network structure is proposed to generate diverse data. The D-SG can adaptively alter architectural configurations and activate different subnet generators in different sequential iteration spaces. The variable network structure increases the complexity and capacity of the generator, and strengthens its ability to generate diversified data. In addition, a novel additive constraint based on the differentiable dhash (D-Dhash) is designed to guide the structure parameter selection of the D-SG. This constraint forces the D-SG to constantly jump out of the fixed generation mode and generate diverse data in semantics and instance. The effectiveness of the proposed model is verified on the experimental benchmark datasets (MNIST, CIFAR-10, CIFAR-100, and SVHN).**

***Index Terms*—generated data diversity, knowledge distillation, dynastic network, image classification, Dhash.**

I. Introduction

THE feature representation and fitting ability of deep neural networks (DNNs) has prompted their wide use in different multimedia data interpretation tasks, such as image recognition and classification [1], [2], object dectction [3], [4], video target tracking [5], [6] and natural language processing [7]. With the increased use of task requirements, the depth and complexity of DNNs are also increasing. Thousands of network layers have been designed and applied to the research on practical problems [8]. In the big data era, complex and powerful deep neural network models have made outstanding contributions to the construction of smart cities [9]–[11].

It is undeniable that the deeper neural networks have a significantly improved reasoning ability [8], [12], [13]. Simultaneously, more computing and storage resources are needed. However, some network layers in DNNs do not contribute to the results in the cognition and inference process. Therefore, effectively utilizing the inference calculation of DNNs is a preferred alternative to transplantation to portable devices. To implement these applications, a series of strategies dedicated to DNN model compression have been proposed, such as model quantization [14], [15], network pruning [16], [17], neural network search [18] and knowledge distillation [19], [20].

Knowledge distillation is a network training method that transfers knowledge from a trained network (called teacher network) to another new network (called student network) [20]. This method can not only realize the model compression by migrating the knowledge from a complex large network to a lightweight network, but also realize the network performance integration [21], [22]. Knowledge distillation has achieved satisfactory results in the above tasks when the training data are available, that is, data-driven knowledge distillation.

As a specific training method of DNNs, knowledge distillation also relies on a large amount of training data [19]. Due to privacy, transmission restrictions, and moral imperatives, the original training data in many practical applications are not available. Especially in some practical applications where there is a stronger demand for lightweight neural networks, such as medical image segmentation and video understanding [23]–[26]. Thus, finding ways to reconstruct training data has become a research focus for data-free knowledge distillation model. Currently, the data reconstruction methods applied to data-free knowledge distillation include selecting other alternative data, using metadata, mining data impressions, and data generation. The method of selecting other relevant data to replace the original training data has obvious shortcomings, such as data collection difficulties, and distribution difference between relevant data and the original data. The acquisition of metadata limits the application flexibility of the distillation models. The method of mining data impressions focuses on the inter-category discriminations and the lack of intra-category differences. Therefore, reconstructing data through data generation becomes a priority [27]–[30].

Ideally, the generated data should capture the distribution of the original training data (as shown in Fig.1(a)). In detail,

This work was supported in part by the Key Scientific Technological Innovation Research Project by Ministry of Education, the National Natural Science Foundation of China Innovation Research Group Fund(61621005), the State Key Program and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (61836009), the Major Research Plan of the National Natural Science Foundation of China (91438201, 91438103, and 91838303), the National Natural Science Foundation of China (U1701267, 62076192, 62006177, 61902298, 61573267, 61906150, and 62276199), the 111 Project, the Program for Cheung Kong Scholars and Innovative Research Team in University (IRT 15R53), the ST Innovation Project from the Chinese Ministry of Education, the Key Research and Development Program in Shaanxi Province of China(2019ZDLGY03-06), the National Science Basic Research Plan in Shaanxi Province of China(2019JQ-659, 2022JQ-607), China Postdoctoral fund(2022T150506) the Scientific Research Project of Education Department In Shaanxi Province of China (No.20JY023), the fundamental research funds for the central universities (XJS201901, XJS201903, JBF201905, JB211908), and the CAAI-Huawei MindSpore Open Fund, the Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

(Corresponding author: Licheng Jiao.)
E-mail: lchjiao@mail.xidian.edu.cn

Xiufang Li, Licheng Jiao, Fang Liu, Xu Liu, Lingling Li, Puhua Chen and Yi Zuo are the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an Shaanxi Province 710071, China.

Qigong Sun is with SenseTime Research & Shanghai AI Lab.
E-mail: sunqigong@sensetime.com. .

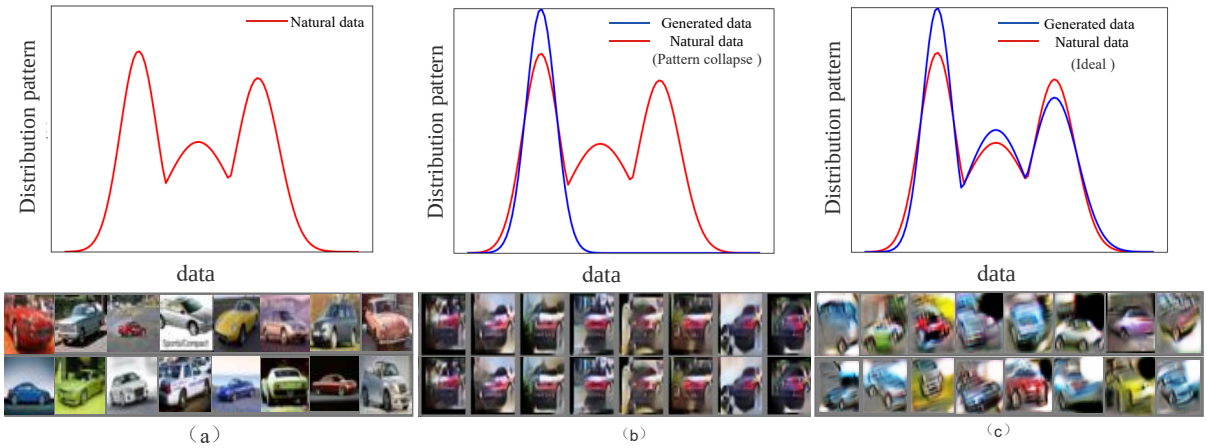


Fig. 1: Simplified diagram of the data distribution pattern and visualization. The upper part of the image is the simplified representation of the data distribution pattern, and the lower part is the corresponding data visualization. In the upper part, the red curve represents a simplified representation of the natural data distribution pattern, and the blue curve represents a simplified representation of the generated data distribution pattern. Corresponding to the distribution above, in the lower part, (a) is the natural data, (b) is the generated data with pattern collapse and (c) is the idea generated data.

a trained teacher network learns the features of the original training data. Only by using the generated data with a similar or the same distribution as the distillation input, can the teacher network provide effective guidance information for the student network in knowledge distillation. In principle, the conventional data generation model is more inclined to generate a single or limited data pattern, that is, cause a pattern collapse [31], [32]. The generated image data with a pattern collapse are shown in Fig.1(b). The generated data are monotonous. The experimental results also verify this problem [28], [33], [34]. When these generated data with a single expression are used for knowledge distillation, the feature extraction and learning ability of the trained teacher network are restricted, which has a negative impact on the distillation. As an alternative, generating diverse data with different expression patterns is a superior choice for data-free knowledge distillation.

As a research focus, some strategies to improve the diversity of generated data have been proposed. Fang *et al.* [35] and micaelli *et al.* [36] mine the generation diversity by the adversarial training. This method easily generates samples far away from the original training data distribution, which is contrary to the initial purpose. Subsequently, DFQ [28] further improved this ideology by utilizing the guidance of the BN layer information of a trained teacher network, which stores the statistical distribution of the original training data [37]. Different from the above methods, the CMI model designs new constraints based on constrastive learning for generators, so that the generator generates diverse data in high-level semantic features [30]. Given the independence between different generators, Luo *et.al* proposed a generation mode using multiple generators. Generators with the same number of data categories are designed to expand the generated data diversity [29]. Although this method has achieved ideal results, it still has obvious shortcomings. Multiple generators increases the volume and computation of the model, especially for datasets with a large number of classes, such as CIFAR-100.

In this paper, we propose a new data-free distillation model. Specifically, a dynamic data-free knowledge distillation model called D^3K is proposed. Explicitly, the generator is designed as a dynamic supernet, and its complexity and capacity increase accordingly. The dynamic supernet generator can adaptively change the network structure parameters, so that the generation configuration of the generator can continuously transform in different iterative spaces. The variable network structure expands the generation space of the generator and gives it the ability to mine the data information learned by the teacher network as much as possible. The rich data information mined from the teacher network is conducive to making the generator generate more authentic and diversified data. Furthermore, to guide the parameter configuration selection of the dynamic supernet generator, a novel similarity constraint based on differentiable Dhash is designed for the generator. This constraint makes the dynastic supernet generator adaptively select credible structural parameters and optimization directions, which are conducive to generating diverse data in semantics and instance.

The main research contributions of the proposed method are as follows:

- A novel data-free knowledge distillation model is proposed, which provides an innovative idea to further mine the data information learned by the trained teacher network. This model can generate data that are similar to the original training data in both diversity and authenticity. The generated data are beneficial to improving the distillation performance.
- A dynastic supernet generator is proposed to replace the fixed generator. The architectural configurations of

the dynastic supernet generator can adaptively change with the network iteration. The variable network structure increases the capacity of the generator, which allows the generator to generate diverse data.
- A novel similarity constraint based on differentiable Dhash is designed as an additional constraint for the generator. This new constraint function with a simple calculation provides guidance for the dynamic generator. It enables the dynamic generator to adaptively select structural parameters that can generate diverse data.
- Experiments verify the effectiveness of this novel distillation model. We applied this distillation method to several classical datasets, and achieved a better performance than the existing knowledge distillation model.

The rest of the paper is organized as follows. In Section II, the works related to the proposed model are briefly summarized. In Section III, the proposed D^3K method and execution process are introduced in detail. In Section IV, the effectiveness of our proposed method is verified through comparison and ablation experiments is verified. In Section V, a brief summary of this paper is presented.

II. Related Work

A. Data-Free Knowledge Distillation

Facing the problem that the original training data are not available, the data-free knowledge distillation methods hope to obtain reconstruction data for the distillation process from these trained teacher networks.

1) Data Reconstruction: The data reconstruction methods applied to data-free knowledge distillation mainly include selecting other alternative data [38], using the data prior stored in the trained teacher network [39], [40], and training a new generator under the guidance of the teacher network to generate data similar to the original training data [29], [41]. The method of selecting other alternative data is time-consuming and difficult. The methods that use the data prior stored by the teacher only consider the local information of the teacher network. Yin *et al.* first incorporated model reverse training theory into the data-free knowledge distillation task [37]. In addition, GAN is often used for data-free knowledge distillation [27], [29].

2) Data Diversity: The expression diversity of the generated data is also the decisive factor in data-free knowledge distillation. Lopes *et.al* modeled the data category information in the softmax layer as a Dirichlet distribution, and enriched the generated data diversity by adjusting the concentration parameters of the Dirichlet distribution [39]. To avoid duplication of redundant data, a method of generating diverse data by minimizing the JS divergence of the teacher network and the student network in the reverse training process of the teacher network is proposed [37]. At the same time, the idea of using the competition mechanism to generate diverse data is also applied to the data-free knowledge distillation model based on the generative adversarial network.

DAFL [27] adds constraints to the feature representation and information entropy of the generated data, so that the generated data have rich category information and assist in the class balance. DFAD [35] generates harder samples through adversarial training. It constructs an optimal upper bound for the generator, and constantly forces the generator to generate diverse data in the optimization process. This method is also used by the later data-free knowledge distillation method [30]. Subsequently, DFQ [28] further optimizes the above ideology. Additionally, LS-GDFD [29] takes multiple generators to increase the diversity of the generated images. The independence of the different generators makes them generate different images. CMI uses the contrastive learning to add constraints for the generator, so that the generator can continuously generate data that have different feature representations from the existing generated data [30].

B. Generative Adversarial Networks

Generative adversarial networks (GANs) are one of the classical data generation models [42]. It is widely used in different data generation tasks, such as style transfer [43], image super-resolution [44] and image classification [45]. The GAN is composed of a generation network G and a discrimination network D. The purpose of G is to generate the required data, and the purpose of D is to judge the authenticity of the generated data. Through this game process, the performances of G and D improve. With the emergence of GANs, DNNs have made unprecedented achievements in the data interpretation task [46]–[48].

Originally, to implement reliable applications, a series of derivative versions were proposed. It includes improving the data authenticity [49]–[52], preventing mode collapse [10]. Subsequently, the method of using an integrated generator or discriminator to enrich the generated data was also proposed [53]–[55].

C. Dynastic Neural Network

The dynamic neural network has become a new research topic in DNNs. It can adaptively adjust the network structure, so as to obtain significant advantages in the adaptive ability, accuracy, and computational efficiency. Common dynamic network implementation methods include the following: designed elastic network depth [56], width [57] and dynastic connection route [58].

At present, the application of dynamic neural networks primarily focuses on improving the efficiency of DNNs, such as designing a network with an adjustable structure to meet the needs of different computing resources [59]–[61], designing different network structure parameters to provide more optional optimization schemes for the model [62]. In addition, a dynamic network is designed to adaptively select the interpretation scheme for different inputs [63]. In addition, the dynamic neural network is also used for knowledge mining [64].

III. Dynamic Data-Free Knowledge Distillation

A. Preliminary

Suppose $F_t(\cdot)$ is a trained image classification network (teacher network). The process of transferring the inference ability of the trained model $F_t(\cdot)$ to another network $F_s(\cdot)$

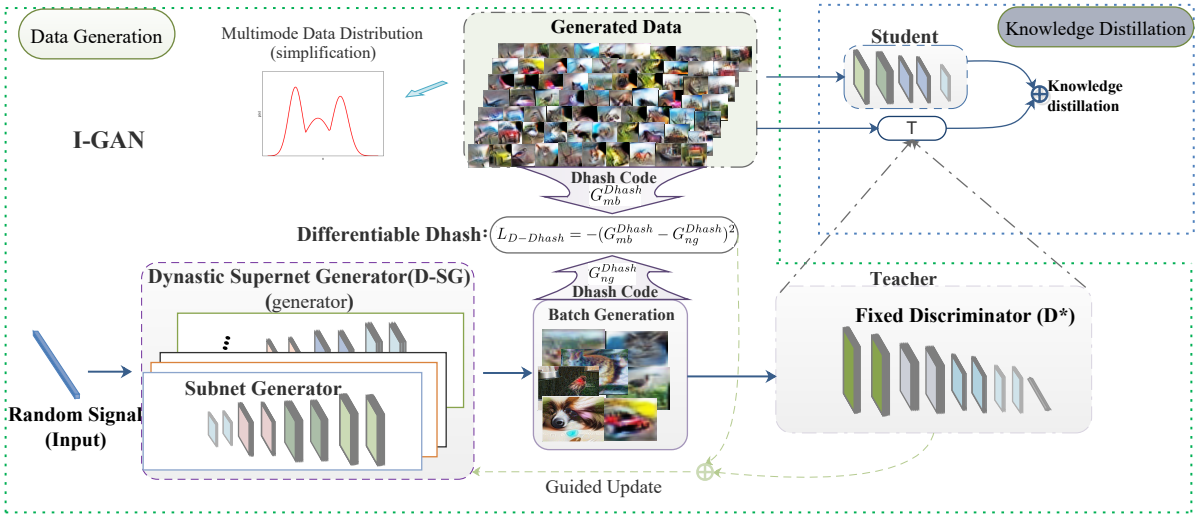


Fig. 2: Schematic diagram of the D^3K. The green dot box is the data generation module, and the blue dot box is the distillation process. The generation process is achieved by I-GAN, which consists of a Dynamic Supernet Generator (D-SG) and a fixed discriminator. A new constraint-based on differentiable dhash (D-Dhash) is applied to the dynamic Supernet generator and enables it to continuously update the subnet generators. Variable subnet generators generate diversified expression data that are different from the existing data, so that the generated data of distillation data have richer expression patterns, as shown in "Multimode data distribution (simplified)".

(student network) is called knowledge distillation. The training goal of a knowledge distillation between networks is to minimize the output difference of $F_t(\cdot)$ and $F_s(\cdot)$ for the same input data x, as shown in the following formula:

$$L = \sum_{(x)\in D} L_{kd}(F_t(x,\tau), F_s(x,\tau)), \tag{1}$$

where D is the original training dataset. τ is the temperature coefficient and L_{kd} is the objective distillation function.

In data-free knowledge distillation, the original training data x are unavailable. To achieve knowledge distillation, the original data must be reconstructed as shown:

$$L = \sum_{(\hat{x})\in \hat{D}} L_{kd}(F_t(\hat{x},\tau), F_s(\hat{x},\tau)), \tag{2}$$

where $\hat{D}$ is the reconstructed data dataset. Therefore, how to use the trained $F_t(\cdot)$ network to construct training data $\hat{D}$ similar to the original data D is a challenging problem.

B. Overview of the D^3K

In this part, the proposed dynamic data-free knowledge distillation model called D^3K is introduced. The core of this method is to provide authentic and diverse training data for the distillation process. As illustrated in Fig.2, the model consists of two components: the data generation module and the knowledge distillation module.

The data generation module is an improved GAN (I-GAN), which is composed of a dynastic supernet generator (D-SG) with variable architectural configurations and a fixed discriminator (D). The D-SG can adaptively change the generator network structure parameters and select different subnet generators in the generation iteration process. A new constraint based on differentiable dhash (D-Dhash) is designed to guide the structural configuration selection of the D-SG. The variable subnet in the D-SG enhances the complexity of the generator and enables it to generate data with rich expression. Through the adversarial learning between the D-SG and the discriminator, I-GAN generates the data closer to the real data (including distribution and diversity), and its objective function is:

$$L_{I-GAN} = \mathbb{E}_{y\sim p_{data}(y)}[log(D(y))] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(D\text{-}SG(z)))]. \tag{3}$$

D-$SG(\cdot)$ adaptively activates unique subnet generators for data generation in each generation iteration. z is the random input. Instead of the simultaneous training and updating of the generator and discriminator in I-GAN, the discriminator is replaced by a trained teacher network (called D^*). The discrimination ability of the fixed discriminator is used to guide and adjust the dynastic training of the generator. Based on this guidance, the dynastic generator continuously generates data with different expression patterns. The objective function of the I-GAN is expressed as:

$$D\text{-}SG^*(z) = arg \min_{D\text{-}SG} \mathbb{E}_{z\sim p_z(z)}[log(1-D^*(D\text{-}SG(z)))]. \tag{4}$$

Thus, the generative adversarial process is transformed into a process of continuously optimizing the D-$SG(\cdot)$ under the guidance of the fixed discrimination network D^*. The D-$SG(\cdot)$ continuously activates different subnet architectures for generators during the generation iteration. The variable network structure of the generator is in favor of improving its ability to mine the data information learned by the teacher.

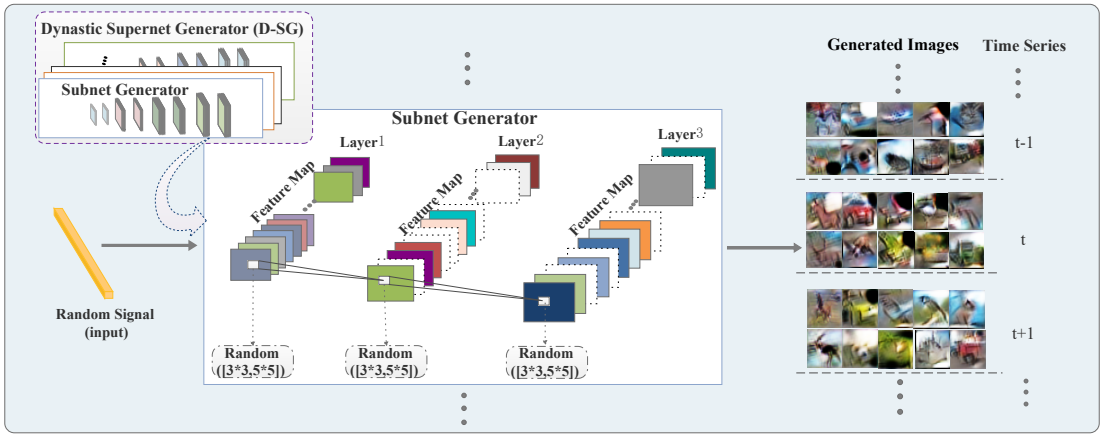


Fig. 3: Schematic diagram of the D-SG. The left shows the simplified figure of the D-SG, and the middle is the detailed version of one of the subnet generators. The right is the data generated by different subnet generators in the different temporal spaces. In the detailed version of one of the subnet generators. The solid line color block in the figure represents the feature channels participating in the training, and the number of channels is the network width of the current layer. The dotted white block indicates the feature channel that does not participate in the training (it does not exist in the actual training). On the right side of the image is the variation of generated data with the variation of generator structure in different generation iterations.

The generator has the ability to generate diverse data that accords with the original training data distribution during the update process. In addition to this optimization goal, a new constraint based on differentiable dhash for the D-$SG(\cdot)$ is designed and forces the generator to generate data with rich instance and semantic features. Then, the generated data with different modes are evenly distributed in the generated data obeying the "Multimode Data Distribution" in Fig.2, so that it can better cover the distribution of the original data. These generated data close to the original training data are taken as the knowledge distillation training data and can implement the effective migration of the knowledge learned from the teacher network to the student network. Next, the dynastic supernet generator and the new constraints in I-GAN will be introduced in detail from their principle and design perspectives.

C. Dynastic Supernet Generator

As a generation model based on implicit statistical mechanisms, the GAN with explicit reward and punishment mechanisms easily captures the single or limited data expression modes [42], [65]. This mechanism limits the expression diversity of the generated data, as shown by the blue line in Fig.1(b). To give the generator the ability to generate diverse data that obey the complex multimodal function as a natural data distribution (simplified as the red line in Fig.1(a)), we model the generator as a dynastic supernet, called the dynastic supernet generator (D-SG).

1) D-SG Design: Different from the conventional static networks, the D-SG is a dynamic network that can continuously change the network structure by flexibly supporting different network structure parameters. In detail, this design captures diverse architecture spaces such as different network widths and kernel sizes in D-SG, so that the generation network takes a different $subnet(\cdot)$ as generators in each generation process. The variable network structure causes the generator to have unique feature construction and selection abilities, which improves the ability of the model to generate diversified data. The objective function of this D-SG is:

$$\begin{aligned} &\min L_{D\text{-}SG}(subnet(W,K),z),\\ &s.t. \quad subnet(W,K)\in\{D\text{-}SG\}, \end{aligned} \tag{5}$$

where L_{D-SG} is the loss of the D-SG, W and K are the sets of the variable network widths and convolutional kernel sizes of the subnet generator, and z is a random input. The $subnet$ is the subnet generator of D-SG. The simplified diagram of D-SG is shown in Fig.3. In Fig.3, the simplified figure on the left shows the D-SG, and the middle is the detailed version of one of the subnet generators. On the right is the data generated by different subnet generators in the different temporal spaces. In the detailed version of the subnet generators, the solid line color rectangles represent the feature channels that actually participate in the calculation. The dotted line blank rectangles are some randomly screened channels, and they do not participate in the network calculation. In the convolution operation of each layer, the value between $\{16 \backsim 256\}$ is randomly selected as the feature channel number, that is, the network width. When performing the convolution operation, the convolution kernel size is randomly selected in $\{3,5\}$ to extract the features of this layer.

The feature map F_{im} at position (x,y) in the ith layer and mth channel is calculated in the convolution process as shown in Eq.6.

$$\begin{aligned} F_{im}^{xy} &= f\Big(\sum_{n}^{N_i-1}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} W_{inm}^{pq} F_{(i-1)n}^{(x+p)(y+q)}\Big),\\ s.t. \quad & m\in M_i,\\ & N_i, M_i \in random\{16\sim 256\},\\ & P_i = Q_i \in random\{3,5\}, \end{aligned} \tag{6}$$

where N_i and M_i are the dimensions of the input channel and output channel respectively. P_i and Q_i are the height and width of the kernel. W_{inm}^{pq} is the weight value at the position (p, q) of the kernel connected to the nth feature channel in the previous layer and mth feature channel in the current layer. In this D-SG, the structural parameters N_i, M_i, P_i, and Q_i are variable in the generation iteration.

This modeling method expands the capacity of the generator and makes it easier to capture and generate diverse data. Next, the design principle of this D-SG is introduced.

2) Principle Analysis: In deep convolution neural networks (CNNs), the convolution operation obtains the abstract data features through layer by layer feature extraction, and realizes the cognition and understanding of data. In detail, one or a group of convolution kernels achieves global understanding by gradually extracting local features in different parts. A convolution kernel achieving convolution in different feature regions remains unchanged, which is a weight-sharing process. The weight-sharing operation can greatly reduce the number of parameters and simplify the calculation. In addition, the weight-sharing operation enables a set of convolution kernels to extract a specific data feature and store it in a specific feature channel [66]. Therefore, different channels contain basically different features, such as texture features in different directions and frequencies. Increasing the width in the deep convolution neural network makes each layer learn richer features (color, edge, direction, etc.). Therefore, it can be seen that the network width is a sensitive structural parameter.

In addition to the width, the size of the convolution kernel also affects the network feature extraction and data interpretation. In deep CNNs, the size of the convolution kernel determines how many local features are used to realize the overall understanding of the input data. The varying convolution kernel makes the network use different local information to understand the input features.

Based on the above analysis, the network width and kernel size are taken as variable parameters to adjust the dynamic variation of the D-SG. In detail, the training objectives of G and D in GANs are different, and D is concerned with the correlation between the extracted features and the labels. The G is concerned with how to obtain data, that is, how to restore the data features. The training of D-SG is similar to the conventional G in I-GAN. It can continuously train the dynamic generator to adaptively predict the data features of a given object at different angles. Capturing the data features of different angles and directions, the dynastic generator has the ability to generate data with different expression modes.

D. Differentiable Dhash

The variable network structure increases the complexity of the generator and gives it have the ability to generate diverse data. To further strengthen this advantage, a novel constraint for the D-SG is designed. We think that the data with different patterns are discrepant in their semantic and instance representation. Additionally, the difference hash simulates human perception by focusing on the features that drive human vision, such as color and frequency [67], [68]. The difference hash can sensitively detect the semantic and instance differences between different images. Explicitly, the hash algorithm can generate distinguishable hash codes for different images. These hash codes can be regarded as the "fingerprint" of the image. The smaller the difference in the "fingerprint" is, the higher the similarity.

Consequently, the D-SG continuously generate diverse data. A new differentiable difference hashing (D-Dhash) constraint is designed for calculating and minimizing the similarity between the generated image and the memory bank which stores the generated data in the last iteration. Compared with a method that only focuses on the similarity of the data features, the differential hashing algorithm is more in line with the human perception when comparing image similarities. This method can enhance the diversity of the generated image at both the instance and semantic levels. The following is a detailed description of the constraint.

Define an image x. First, sparsing data representation through average pooling with the kernel (4,4):

$$H_x = avepool2d(x, 4, 4), \tag{7}$$

where H_x is the pooled image. Then, calculate the grayscale of the image:

$$G_x = Gray(H_x), \tag{8}$$

where H_x^i is the i-th channel feature in H_x. Then, the image is reshaped and changed into a 1-dimensional vector:

$$G_x^{reshape} = Reshape(G_x). \tag{9}$$

We calculate the mean g_{mean} of $G_x^{reshape}$, and normalize $G_x^{reshape}$ to obtain G_x^{mean}:

$$G_x^{mean} = \sum_{l=1}^{L} G_x^{reshape}(i) - g_{mean}, \tag{10}$$

where L is the size of $G_x^{reshape}$. Hashing code G_x^{Dhash} generation:

$$G_x^{Dhash} = \sum_{l=1}^{L} G_x^{Dhash}(i) = \begin{cases} 0 & if \quad G_x^{mean}(i) \le G_x^{mean}(i+1) \\ 1 & if \quad G_x^{mean}(i) > G_x^{mean}(i+1) \end{cases}, \tag{11}$$

Dhash encodes every image to its binary expression. When calculating the difference hash value, we convert the non-differentiable Hamming distance into an equivalent differentiable computation. In short, the mean square error instead of the XOR bit operation is used to calculate the Hamming distance between the different encodings. The loss function is as follows:

$$L_{D-Dhash} = -(G_{mb}^{Dhash} - G_{ng}^{Dhash})^2, \tag{12}$$

where G_{mb}^{Dhash} is the DHash encoded value in the memory bank and G_{ng}^{Dhash} is the DHash encoded value of the generated data. This transformation enables the model to be optimized through gradient descent. We define the gradient propagation

for loss based on D-DHash encoding as:

$$\begin{aligned}\frac{\partial L_{D-Dhash}}{\partial x} &= \frac{\partial L_{D-Dhash}}{\partial G_{ng}^{Dhash}} * \frac{\partial G_{ng}^{Dhash}}{\partial x} \\ &= \frac{\partial L_{D-Dhash}}{\partial G_{ng}^{Dhash}} * \frac{\partial G_{ng}^{Dhash}}{\partial G_{ng}^{mean}} * \frac{\partial G_{ng}^{mean}}{\partial x} \\ &= \frac{\partial L_{D-Dhash}}{\partial G_{ng}^{Dhash}} * \lambda * \frac{\partial G_{ng}^{mean}}{\partial x},\end{aligned} \tag{13}$$

where λ is a superparameter. According to the encoding method of Eq.(11), λ needs to be greater than 0. Since the encoding process of Eq.(11) maps the data to the fingerprint space and it does not involve addition, subtraction, multiplication, and division. Therefore, the value of λ only affects the update speed and fineness of the generated data, and will not have a fatal impact on the results. We compared 0.3, 0.6, 1.0, and 1.2 in the experiment, and finally chose 1.0 which has the better performance.

E. Generator Loss

To improve the validity of the generated data, a series of data generation techniques that provide guidance for generating the data distribution and constraint boundaries have been applied.

Distribution Constraint. Under the isotropic Gaussian assumption, the generated data features are statistically close to the original training data [37], and its loss function is:

$$L_{bn}(\hat{x}) = \sum_{l} D_{kl}(\mathcal{N}(\mu_l(\hat{x}), \sigma_l^2(\hat{x}))\|\mathcal{N}(\mu_l, \sigma_l^2)), \tag{14}$$

where μ_l and σ_l are the mean and variance of the l-th BN layer in the trained teacher network respectively, and $\hat{x}$ is the generated data. $\mathcal{N}(\cdot)$ is the statistical information. $\mu_l(\hat{x})$ and $\sigma_l(\hat{x})$ are the mean and variance activated by the generated data in the l-th BN layer. D_{kl} is the KL divergence used to measure the distribution difference.

Category Information. The class discrimination ability of the teacher network as a discriminator is further mined [27], and under its guidance, the generator generates diversified data with accurate class information. Its loss function is:

$$L_{cls}(\hat{x}) = L_{CE}(F_t(\hat{x}), c), \tag{15}$$

where c is the category information. L_{CE} is the cross-entropy loss.

Adversarial Distillation. Inspired by the robustness optimization principle, a data generation constraint based on the differential cognition of teachers and students is proposed to provide an excellent optimization upper bound for the model [35], [36]. The model loss function based on this constraint is:

$$L_{adv}(\hat{x}) = -D_{kl}(F_t(\hat{x})\|F_s(\hat{x})) \tag{16}$$

To further make the generated data close to the original training data, we incorporate the above strategies and design a new model to further enrich the expression diversity of the generated data. Finally, the objective function of the $D\text{-}SG$ is expressed as:

$$L_{D\text{-}SG} = \alpha L_{bn} + \beta L_{cls} + \gamma L_{adv} + \theta L_{D\text{-}Dhash} \tag{17}$$

F. D^3K Implementation

Algorithm 1 systematically summarizes the scheme of the D^3K. In Algorithm 1, $subneti$ represents the activated subnet generator architecture in the $D\text{-}SG$. During the generation process, the structure of the $subneti$ is changeable.

Algorithm 1: D^3K

Input: A given teacher network F_t
Output: A portable student network F_s
Structuring data pool;
Designing $\{D\text{-}SG(\cdot)\}$;
step 1: Data generation;
for $number\ of\ batches$ **do**
  Designing $D\text{-}SG(\cdot)$ and initialize the weights;
  Initializing input noise: z = $\mathcal{N}(0, 1)$;
  Dynamically select a specific $subnet$ generator:
  $D\text{-}SG(\cdot) \longleftarrow$
  $random(subnet1, subnet2, subnet3, \cdots)$;
  for $number\ of\ iterations$ **do**
    Generate data $\hat{x}$, $\hat{x} \longleftarrow D\text{-}SG(z)$;
    Construct $L_{D\text{-}Dhash}$;
    Update weight: $L_{D\text{-}SG} \longleftarrow$
    $\alpha L_{bn} + \beta L_{cls} + \gamma L_{adv} + \theta L_{D\text{-}Dhash}$;
    $\theta_{D\text{-}SG} \longleftarrow \theta_{D\text{-}SG}$ - $\eta \nabla_{\theta_{D\text{-}SG}} L$;
  end
  Saving best generated data $\hat{x}$ to data pool
end
step 2: Knowledge distillation:;
for $number\ of\ batches$ **do**
  input generated data $\hat{x}$;
  Computing loss: $L \longleftarrow L_{KD}(F_s(\hat{x}, \theta_s), F_t(\hat{x}, \theta_t))$;
  Update weight:
  $\theta_s \longleftarrow \theta_s - \eta \nabla_{\theta_s} L_{KD}(F_s(\hat{x}, \theta_s), F_t(\hat{x}, \theta_t))$
end

IV. EXPERIMENT

In this section, a series of experiments are designed to verify the effectiveness of the proposed method on a series of image classification benchmarks. First, the implementation details of the experiment are described. Then, the performance of the proposed method is verified by comparing it with the experimental results of the state-of-the-art methods. Third, the advantages of the different optimization strategies in the novel model are verified by ablation experiments. Finally, the effectiveness of the D^3K is analyzed.

A. Experimental Setup

In this part, the effectiveness of the proposed data-free knowledge distillation method has been verified on a set of benchmark datasets: MNIST [69], CIFAR-10, CIFAR-100 [70], and SVHN [71]. According to the existing knowledge distillation, we perform classification tasks on these datasets.

MNIST[1]: MNIST is a handwritten digital image dataset that was initiated by the National Institute of Standards and

[1]http://yann.lecun.com/exdb/mnist/

TABLE I: Structural parameters of the dynastic supernet generator (D-SG)

Net Layer	Operation(Input)	Structural Parameters
Layer1	BN1 UpSample($\times$ 2) conv1(k_s$\times$ k_s $\times$ conv1_O $\times$conv1_I)	kernel_size(k_s) = {3,5}, conv1_I = Random[16 $\sim$ 256], conv1_O = Random[16 $\sim$ 256]
Layer2	BN2 LeakyRelu UpSample($\times$ 2) conv2(k_s$\times$ k_s $\times$ conv1_I $\times$conv2_I)	kernel_size(k_s) = {3,5}, conv1_I = Random[16 $\sim$ 256], conv2_I = Random[16 $\sim$ 256]
Layer3	BN3 UpSample($\times$ 2) conv3(k_s$\times$ k_s $\times$ conv2_I $\times$ 3)	kernel_size(k_s) = {3,5}

Technology (NIST). There are a total of 250 handwritten digital images.

CIFAR[2]: CIFAR-10 and CIFAR-100 are the natural images datasets with a size of 32×32. CIFAR-10 contains 60000 images and 10 categories. Each category has 6000 images. There are 50000 images in the training set and 10000 images in the testing set.

CIFAR-100 contains 60000 images and has 100 categories. Each category contains 600 images, of which 500 images are training images and 100 images are test images.

SVHN[3]: The Street View House Numbers (SVHN) Dataset is an image digit recognition dataset of over 600,000 digit images coming from real-world data. The images are cropped to 32×32. Each image contains a set of Arabic numerals '0-9'.

The Distillation packages including ResNet-34 $\rightarrow$ ResNet-18, WRN-40-2 $\rightarrow$ WRN-40-1, WRN-40-2 $\rightarrow$ WRN-16-1, and WRN-40-2 $\rightarrow$ WRN-16-2, VGG11 $\rightarrow$ ResNet-18 are applied to our experiments. which is consistent with the network selected by the existing data-free model for knowledge distillation. In addition, a series of data-free self distillation experiments are conducted on WRN-40-2 and WRN-16-1. In our experiments, all the teacher networks are trained with the original training data, and the student networks are trained with data generated from the teacher networks. Similar to the existing data-free knowledge distillation models, we use the Adam Optimizer with a 1e3 learning rate to update the generator and an SGD optimizer with a 0.9 momentum to optimize the student training. The initial learning rate of the student is 0.1. Referring to the existing research, we set α, β and γ in Eq.(17) as 1, 1 and 0.5 respectively. The temperature coefficient τ of the distillation process in Eq.(2) is set to 0.5.

B. Comparative Experiments

To reflect the performance advantages of our proposed method, we select the existing classical data-free knowledge distillation models for comparison. such as DAFL [27], DFAD [35], ZSKT [36], DFQ [28], CMI [30], LS-GDFD [29], MB-DFKD [72], PRE-DFKD [73], and FastDFKD [74], and so on. In order to verify the stability of the model, six repeated experiments are performed and the experimental results are expressed in the form of mean (μ)$\pm$ variance(σ^2).

[2]https://www.cs.toronto.edu/ kriz/cifar.html

[3]http://ufldl.stanford.edu/housenumbers/

TABLE II: Accuracy comparisons of the different knowledge distillation methods on ResNet-18 based on pre-trained ResNet-34 and VGG-11, and the original training data are the CIFAR-10 and CIFAR-100 datasets. The results marked by "*" come from our re-implementations. ✓ represents the original data that participated in the experiment and ✗represents that there are no training data.

Model	Method	Original	CIFAR-10	CIFAR-100
T_1(ResNet-34)	only training	✓	95.29	77.49
S_1(ResNet-18)	only training	✓	95.20	77.10
	PKD [20]	✓	94.34*	76.87*
	ADI [37]	✗	93.26	61.32*
	DAFL [27]	✗	92.22	74.47
	DFAD [35]	✗	93.30	67.70
	ZSKT [36]	✗	93.32*	67.74*
	MB-DFKD [72]	✗	92.40	75.35
	PRE-DFKD [73]	ding55	87.40	70.20
	FastDFKD [74]	✗	94.05	74.34
	DFQ [28]	✗	94.61	77.01
	CMI [30]	✗	94.84	77.04
	NaturalInversion [75]	✗	94.87	73.25*
	Ours	✗	**95.07$\pm$0.06**	**77.06$\pm$0.03**
T_2(VGG-11)	only training	✓	92.25	71.32
S_1(ResNet-18)	only training	✓	95.20	77.10
	ADI [37]	✗	90.37	54.12*
	DAFL [27]	✗	81.10*	57.29*
	ZSKT [36]	✗	89.46*	34.72*
	FastDFKD [74]	✗	90.53	67.44
	DFQ [28]	✗	90.84	66.72*
	CMI [30]	✗	91.13	70.56
	Ours	✗	**91.77$\pm$0.21**	**71.05$\pm$0.15**

TABLE I shows the structural parameters of the D-SG. The D-SG is a simple three-layer convolutional neural network. It can be seen from the table that when performing the convolution operation of each layer, the number of feature channels is a random number in the range of $\{16 \sim 256\}$, and the size of the convolution kernel is also randomly selected in $\{3 \times 3\}$ or $\{5 \times 5\}$. The structural variety of the D-SG strengthens the expression ability of the network without increasing the storage and computing requirements.

TABLE II reports the distillation results of ResNet-34 $\rightarrow$ ResNet-18 and VGG-11 $\rightarrow$ ResNet-18 on the CIFAR dataset. Better than the existing data-free knowledge distillation model, our proposed model achieves the optimal distillation results. From the experimental results, the variance of six repeated experiments is less than 0.5. Compared with other methods, the advantages of this model are stable on CIFAR-10 and CIFAR-100. On the CIFAR-10 dataset, under the guidance of ResNet-34, the performance of ResNet-18 is basically close to the training results using the original training data. In addition,

TABLE III: Accuracy comparisons of the different knowledge distillation methods on WRN-16-1, WRN-40-1 and WRN-16-2 based on pre-trained WRN-40-2, and the original training data are CIFAR-10 and CIFAR-100 dataset. The results marked by "*" come from our re-implementations. ✓ represents the original data that participated in the experiment and ✗represents that there are no training data.

Model	Method	Original	CIFAR-10	CIFAR-100
T_1(WRN-40-2)	only training	✓	94.87	75.83
S_1(WRN-16-1)	only training	✓	91.12	65.31
	ADI [37]	✗	83.04	53.77*
	DAFL [27]	✗	65.71	22.50
	DFAD [35]	✗	85.48*	52.35*
	ZSKT [36]	✗	84.24	30.15*
	DFQ [28]	✗	86.14	55.00*
	FastDFKD [74]	✗	89.29	54.02
	CMI [30]	✗	90.01	57.91
	Ours	✗	**90.90±0.20**	**60.39±0.39**
S_2(WRN-40-1)	only training	✓	93.94	72.19
	ADI [37]	✗	86.85*	61.32*
	DAFL [27]	✗	81.33*	34.66*
	DFAD [35]	✗	87.81*	57.23*
	ZSKT [36]	✗	88.39	29.73*
	DFQ [28]	✗	91.69	61.92*
	FastDFKD [74]	✗	92.51	63.91
	CMI [30]	✗	92.78	68.88
	Ours	✗	**93.27±0.36**	**69.95±0.17**
S_3(WRN-16-2)	only training	✓	93.95	73.56
	ADI [37]	✗	89.72*	61.34*
	DAFL [27]	✗	81.55*	40.00*
	DFAD [35]	✗	87.74*	54.65*
	ZSKT [36]	✗	89.66	28.44*
	DFQ [28]	✗	92.01	59.01*
	FastDFKD [74]	✗	92.45	65.12
	CMI [30]	✗	92.52	68.75
	Ours	✗	**93.50±0.39**	**69.92±0.11**

under the same experimental conditions, our proposed method (95.07%) exceeds LS-GDFD (95.02%). Similar to ResNet-34 → ResNet-18, the distillation results of VGG-11 → ResNet-18 also have obvious advantages over the existing methods.

On the CIFAR-100 dataset, the distillation accuracy of ResNet-18 from VGG-11 is 0.27% less than that of VGG-11 trained by original training data. The distillation accuracy of ResNet-18 from ResNet-34 is 0.05% less than that of ResNet-18 trained by the original training data. The two distillation results are better than those of other existing methods. The best precision of four repeated tests of PRE-DFKD on the CIFAR-100 dataset is the same as our best precision, however, its best accuracy on CIFAR-10 is 1.0% lower than ours.

TABLE III shows the distillation results of WRN-40-2 → WRN-40-1, WRN-40-2 → WRN-16-1, and WRN-40-2 → WRN-16-2 on the CIFAR dataset. From the comparison results in this table, we can see that our proposed method is superior to the existing data-free knowledge distillation models. On the CIFAR-10 dataset, compared with the network trained directly with the original training data, the accuracy loss of our proposed method is less than 0.7%. Experiments on the CIFAR-100 dataset further demonstrate the advantages of our proposed method. The distillation accuracy of WRN-40-2 to WRN-16-1 is nearly 2.5% higher than that of the best existing models.

To further verify the effectiveness of the proposed model on different datasets, TABLE IV compares its performance and existing methods on the SVHN dataset. Similar to the CIFAR

TABLE IV: Accuracy comparisons of the different knowledge distillation methods on the SVHN dataset based on pre-trained WRN-40-2. The results marked by "*" come from our re-implementations. ✓ represents the original data that participated in the experiment and ✗represents that there are no training data.

Model	Method	Original	SVHN
T_1(WRN-40-2)	only training	✓	97.45
S_1(WRN-40-1)	only training	✓	97.28
	ZSKT [36]	✗	89.74*
	DFQ [28]	✗	95.01*
	CMI [30]	✗	95.96*
	Ours	✗	**96.78±0.25**
T_1(WRN-40-2)	only training	✓	97.45
S_2(WRN-16-1)	only training	✓	96.74
	ZSKT [36]	✗	89.06*
	DFQ [28]	✗	94.73*
	CMI [30]	✗	95.84*
	Ours	✗	**96.44±0.31**
T_2(Resnet34)	only training	✓	96.49
S_2(Resnet18)	only training	✓	96.08
	ZSKT [36]	✗	93.72*
	DFQ [28]	✗	94.90*
	CMI [30]	✗	95.21*
	MB-DFKD [72]	✗	95.40
	Ours	✗	textbf95.96±0.17
T_3(VGG11)	only training	✓	96.07
S_2(Resnet18)	only training	✓	96.08
	ZSKT [36]	✗	92.35*
	DFQ [28]	✗	93.97*
	CMI [30]	✗	94.86*
	Ours	✗	**95.55±0.41**

TABLE V: Accuracy comparisons of the different knowledge distillation methods on the MNIST dataset based on pre-trained ResNet-18. The results marked by "*" come from our re-implementations. ✓ represents the original data that participated in the experiment and ✗represents that there are no training data.

Model	Method	Original	MNIST
T(resnet18)	only training	✓	99.68
S(resnet8)	only training	✓	99.32
	student-KD [20]	✓	99.37
	CMI [30]	✗	99.04*
	Ours	✗	**99.22±0.05**

TABLE VI: Self distillation accuracy comparison of WRN-40-2, WRN-16-1 on SVHN, CIFAR-10, and CIFAR-100 datasets. In the results, A/B: A represents distillation accuracy, and B is the accuracy obtained by training with the original training data.

Model	Method	SVHN	CIFAR-10	CIFAR-100
WRN-40-2	ZSKT [36]	85.63/97.45	87.26/94.87	40.39/75.83
	DFAD [35]	87.71/97.45	88.19/94.87	60.25/75.83
	DFQ [28]	95.03/97.45	92.35/94.87	70.84/75.83
	CMI [30]	95.12/97.45	94.09/94.87	72.74/75.83
	Ours	**96.73**/97.45	**94.72**/94.87	**75.43**/75.83
WRN-16-1	ZSKT [36]	88.71/96.74	77.63/91.12	30.04/65.31
	DFAD [35]	90.09/96.74	81.43/91.12	51.81/65.31
	DFQ [28]	94.04/96.74	87.57/91.12	54.98/65.31
	CMI [30]	94.53/96.74	90.21/91.12	57.03/65.31
	Ours	**96.29**/96.74	**90.64**/91.12	**60.06**/65.31

dataset, the WRN-40-2 → WRN-40-1, WRN-40-2 → WRN-16-1, and ResNet-34 → ResNet-18, and VGG-11 → ResNet-18 are selected as distillation packages in these experiments. The results show that the method proposed in this paper has a better distillation effect on the SVHN dataset.

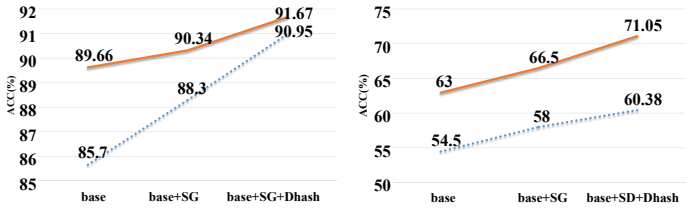


Fig. 4: Ablation experiments on the effects of D-SG and diversity constraints on model performance. The olid orange line indicates the distillation accuracy of VGG-11 → ResNet-18, and the blue dotted line indicates the distillation accuracy of WRN-40-2 → WRN-16-1. (a) is based on CIFAR-10, and (b) is based on CIFAR-100.

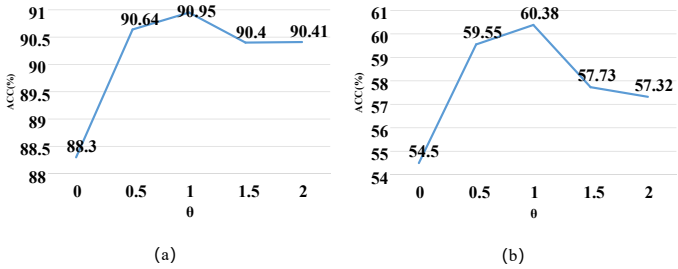


Fig. 5: Distillation accuracy variation curves of WRN-40-2 → WRN-16-1 under different $L_{D-Dhash}$ loss weight coefficient θ on CIFAR-10 and CIFAR-100 dataset. Ablation experiments on the effects of $L_{D-Dhash}$ weighting factor θ on model performance. (a) is based on CIFAR-10, and (b) is based on CIFAR-100.

Because we need the statistical information of the BN layer, we choose ResNet-18 as the teacher network and ResNet-8 as the student network in the experiment based on MNIST data. It can be seen from the results in TABLE II, TABLE III and TABLE IV that CMI can better represent the most advanced distillation results. Therefore, it is taken as a representative model in the comparative experiment. In the experiment, the student-KD is the knowledge distillation method using the original training data [20]. It and the CMI model are selected for comparison. As seen from the results in the TABLE V, the student accuracy obtained by this method is close to the accuracy obtained by the training data. The accuracy losses do not exceed 0.2%.

The above experiments show that the model proposed in this paper has obvious advantages in the model compression task. To further verify the effect of the model, a series of data-free self distillation experiments are implemented. In detail, the teacher and student have the same network framework, and the trained teacher networks are used to train the student networks. The self distillation experiments are based on the SVHN and CIFAR datasets, and the experimental results are shown in TABLE VI.

The experimental results of the data-free self distillation task further verify the advantages of the proposed D^3K. From the

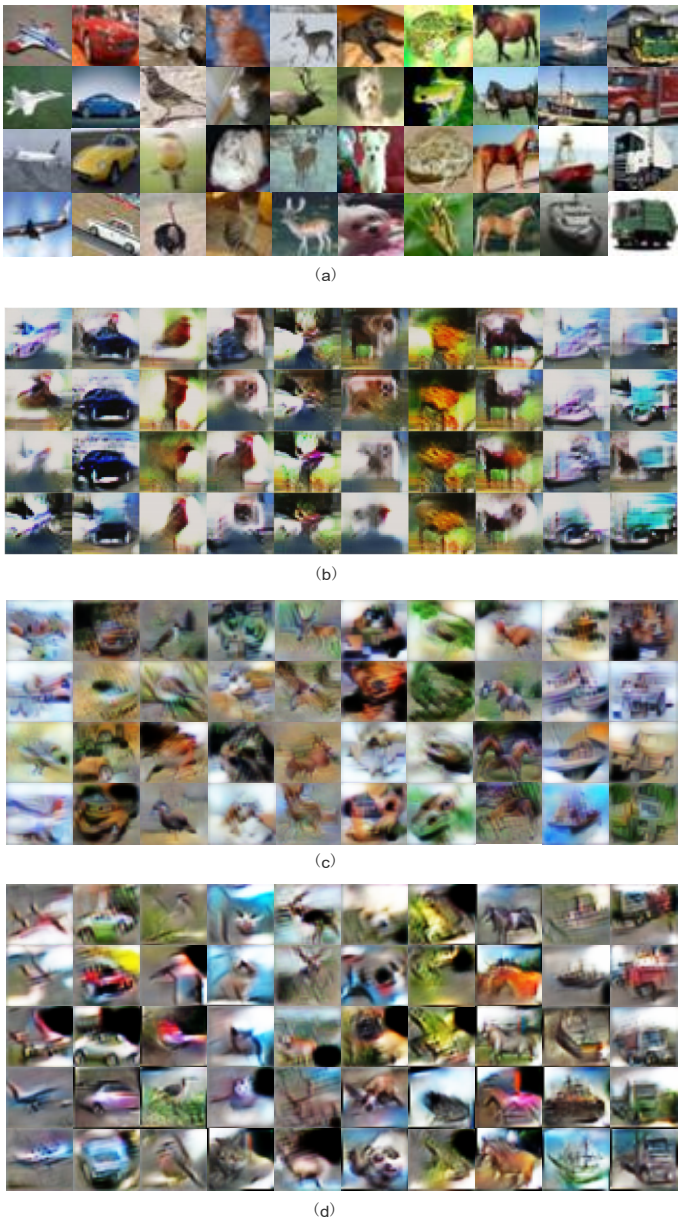
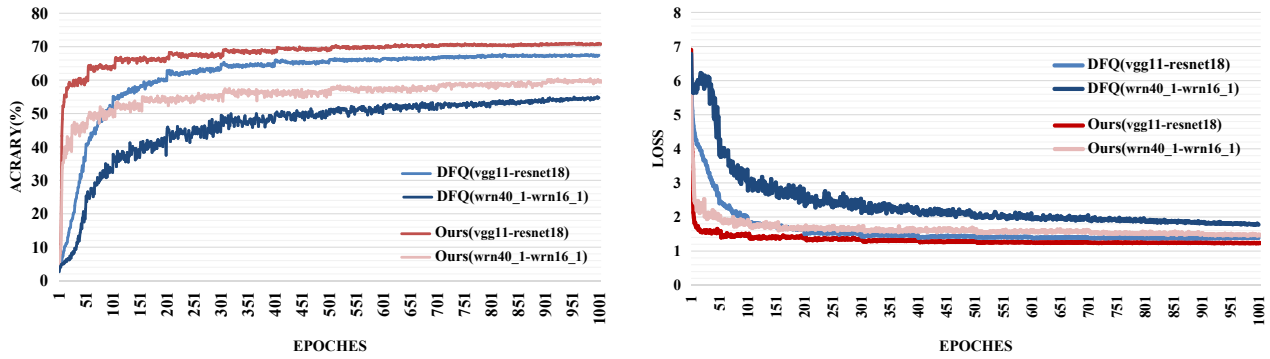


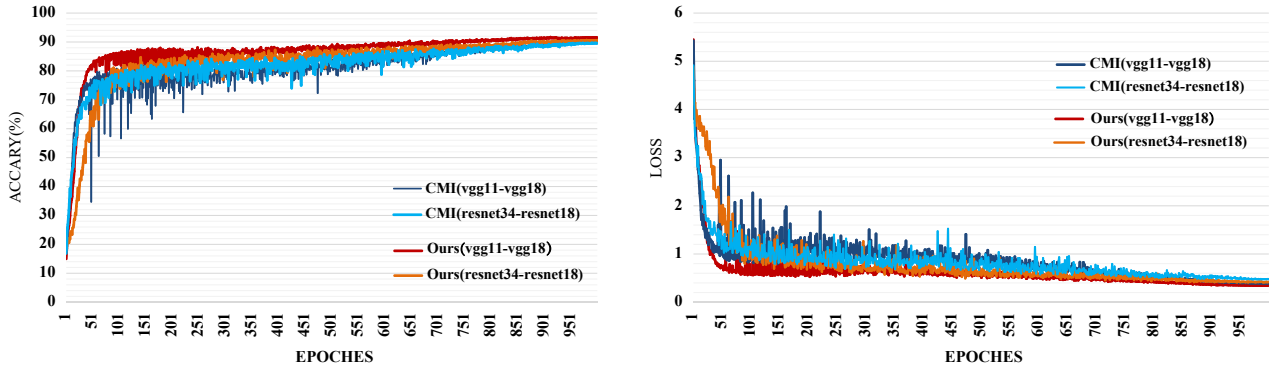
Fig. 6: The generated data in data-free KD from WRN40-2 to WRN16-1 for CIFAR-10. (a) are randomly selected original training data, (b) shows the randomly selected generated images in DFQ [28], (c) shows the randomly selected generated images in CMI [30] and (d) shows the randomly selected generated images by the proposed D^3K.

Fig. 7: Generated SVHN images in D^3K.



(a) (b)

Fig. 8: Loss change comparison results of DFQ and D^3K on the CIFAR-100 dataset. (a) is the accuracy variation curve, (b) is the loss variation curve.



(a) (b)

Fig. 9: Accuracy change comparison results of CMI and D^3K on the CIFAR-10 dataset. (a) is the accuracy variation curve, (b) is the loss variation curve.

experimental results on the CIFAR-10 and SVHN, it can be found that without the participation of the original training data, the data-free distillation model proposed in this paper can obtain almost the same classification accuracy as the model trained by the original training data. The accuracy decline is concentrated at approximately 0.5% and dose no exceed 1%. Although this advantage decreases slightly in the CIFAR-100 dataset, the self distillation experiment further verifies the effectiveness and advantages of our proposed D^3K.

C. Ablation Experiments

In the ablation studies, the influence of the proposed strategies such as the design of D-SG and the diversity constraint loss $L_{D-Dhash}$ for the generator is verified. In order to quantitatively verify the effectiveness of our proposed strategy, we selected two distillation models (WRN-40-2 $\rightarrow$ WRN-16-1 and VGG-11 $\rightarrow$ ResNet-18) for ablation analysis on the CIFAR dataset.

Fig.4 presents the impact of the D-SG and $L_{D-Dhash}$ constraints. (a) shows the ablation result on CIFAR-10, and (b) is the ablation result on CIFAR-100. The solid red line indicates the distillation accuracy of VGG-11 $\rightarrow$ ResNet-18, and the blue dotted line indicates the distillation accuracy of WRN-40-2 $\rightarrow$ WRN-16-1. The "base" is the distillation model without D-SG and the diversity constraint loss $L_{D-Dhash}$. "base+SG" is the distillation model with D-SG and without the $L_{D-Dhash}$ constraint. "base+SG+Dhash" is the proposed model in this paper. According to Fig.4, we can see that both D-SG and the diversity constraints effectively improve the distillation effect.

Fig.5 shows the impact of the $L_{D-Dhash}$ weighting factor θ on the distillation performance. The ablation experiment is based on the distillation process from WRN-40-2 to WRN-16-1 on the CIFAR-10 and CIFAR-100 datasets. It can be seen from the figure that on the CIFAR-10 and CIFAR-100 datasets, a better distillation accuracy is obtained when the $L_{D-Dhash}$ weighting factor θ is 1.

D. Effectiveness Analysis

The above experimental results verify the advantages of the proposed model in the perspective of quantitatively. To qualitatively verify the effectiveness of the method proposed in this paper, we compare the data generated by this method with the data generated by CMI and DFQ.

Fig.6 shows the image comparison results, (a) are original training data selected randomly in the training dataset, (b) are the generated images with DFQ, (c) are the generated images with CMI and (d) are the generated images with the proposed D^3K in this paper. Each vertical column in the figure is a category, and the categories from left to right are "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", and "truck". As can be seen from the figure, CMI and our proposed method generate more diversified data than DFQ. Moreover, compared with CMI, our proposed method improves the overall diversity of data expression(including the semantic and instance diversity). This advantage is clearly reflected in the different categories of generated data, especially "automobile", "bird", "cat", "horse", "dog" and "ship".

Through comparison, it can be found that the data generated by the proposed method have richer expression patterns. The generated images have diverse angles and colors, and they are closer to the original training data. The generated "automobile" is shown in Fig.1(c). The variance of six experiments is less than 0.5. It can be found that it is closer to the "automobile" in the original training data (shown in Fig.1(a)) in terms of distribution patterns and visualization.

Fig.7 shows the SVHN data generated by our proposed method in the data-free self distillation process based on WRN16-1. The digits from left to right are $(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)$. It can be found that this method enriches the expression diversity of the generated data.

In addition to the visualization of the generated image, we also verify the effectiveness of the proposed method from the perspective of convergence. To clearly show the convergence of the model proposed in this paper, we select CMI and DFQ as comparative models to analyze the convergence on the CIFAR dataset. Fig.8(a) and Fig.8(b) show the comparison results with the DFQ model on the CIFAR-100 dataset. Fig.8(a) shows the accuracy variation curves and Fig.8(b) shows the loss variation curves. It is obvious from the different variation curves that our method has a faster convergence speed and higher accuracy. Fig.9(a) and Fig.9(b) show the comparison results with the CMI model on the CIFAR-10 dataset. Fig.9(a) shows the accuracy variation curve and Fig.9(b) shows the loss variation curve. Through comparison, it can be seen that our method and CMI are competitive in convergence. Moreover, our method achieves a higher classification accuracy. A dynamic generator with an adaptive structure variation can quickly generate diversity data in semantics and instance, so that the student network can learn more abundant data features in a shorter iteration space. This advantage is conducive to the faster convergence of the model.

This section verifies the effectiveness of the proposed model by generating image visualization and model convergence. The validity analysis further verifies the advantages of D^3K in the data-free knowledge distillation task.

V. Conclusion

This paper introduces a novel data-free knowledge distillation model called D^3K. The D^3K proposes an improved GAN and models the fixed generator as a dynastic supernetwork called a dynastic supernet generator (D-SG). The dynamic changes of the D-SG enhance its ability to mine data information in the trained teacher network. Thus, the design of the D-SG improves the complexity of the generation patterns, and then improves the expression diversity of the generated data. To strengthen this effect, a new constraint based on the differentiable dhash is designed. This constraint provides the guidance for both the adaptive selection of the dynamic generator structural parameters and the optimization direction of the generator. Variable generative structures force the D-SG to constantly update the generation pattern. Without any original training data, the model achieves comparable results with the original training data in some classic network structures. In addition, we also found the optimal hyperparametric setting for the model through ablation experiments. The effectiveness of the proposed model framework is verified by a quantitative and qualitative analysis on several benchmark datasets.

References

[1] S. Qiu, S. Anwar, and N. Barnes, "Geometric back-projection network for point cloud classification," *IEEE Trans. Multimedia*, vol. 24, pp. 1943–1955, 2022.

[2] Q. Huang, Y. Liang, J. Wei, Y. Cai, H. Liang, H.-f. Leung, and Q. Li, "Image difference captioning with instance-level fine-grained feature representation," *IEEE Trans. Multimedia*, vol. 24, pp. 2004–2017, 2022.

[3] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1968–1979, 2022.

[4] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang, "New generation deep learning for video object detection: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 3195 – 3215.

[5] Z. Chen, J. Li, J. Wu, J. Chang, Y. Xiao, and X. Wang, "Drift-proof tracking with deep reinforcement learning," *IEEE Trans. Multimedia*, vol. 24, pp. 609–624, 2022.

[6] K. Yang, Z. He, W. Pei, Z. Zhou, X. Li, D. Yuan, and H. Zhang, "Siamcorners: Siamese corner networks for visual tracking," *IEEE Trans. Multimedia*, vol. 24, pp. 1956–1967, 2022.

[7] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

[9] D. A. Tedjopurnomo, Z. Bao, B. Zheng, F. M. Choudhury, and A. K. Qin, "A survey on modern deep neural network for traffic prediction: Trends, methods and challenges," *IEEE Trans. Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1544–1561, 2022.

[10] F. Trenta, S. Conoci, F. Rundo, and S. Battiato, "Advanced motion-tracking system with multi-layers deep learning framework for innovative car-driver drowsiness monitoring," in *Proc. IEEE Int. Conf. Automatic Face & Gesture Recognit. (FG 2019)*, pp. 1–5, IEEE, 2019.

[11] F. Ahmad, A. Abbasi, B. Kitchens, D. Adjeroh, and D. Zeng, "Deep learning for adverse event detection from web search," *IEEE Trans. Knowledge and Data Engineering*, vol. 34, no. 6, pp. 2681–2695, 2022.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4700–4708, 2017.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–9, 2015.

[14] Y. Zhou, S.-M. Moosavi-Dezfooli, N.-M. Cheung, and P. Frossard, "Adaptive quantization for deep neural network," in *Proc. Conf. Artif. Intell.*, 2018.

[15] Q. Sun, F. Shang, K. Yang, X. Li, Y. Ren, and L. Jiao, "Multi-precision quantized neural networks via encoding decomposition of {-1,+ 1}," in *Proc. Conf. Artif. Intell.*, vol. 33, pp. 5024–5032, 2019.

[16] Z. Wang, C. Li, and X. Wang, "Convolutional neural network pruning with structural redundancy reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 14913–14922, 2021.

[17] S. Srinivas and R. V. Babu, "Data-free parameter pruning for deep neural networks," *arXiv preprint arXiv:1507.06149*, 2015.

[18] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[19] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[21] H. Zhao, X. Sun, J. Dong, C. Chen, and Z. Dong, "Highlight every step: Knowledge distillation via collaborative teaching," *IEEE Trans. Cybernetics*, pp. 2070 – 2081, 2020.

[22] A. Malinin, B. Mlodozeniec, and M. Gales, "Ensemble distribution distillation," *arXiv preprint arXiv:1905.00076*, 2019.

[23] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 36–46, Springer, 2021.

[24] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, "Cross-modality deep feature learning for brain tumor segmentation," *Pattern Recognition*, vol. 110, p. 107562, 2021.

[25] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, "Scribble-supervised video object segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 2, pp. 339–353, 2021.

[26] W. Zhao, J. Zhang, L. Li, N. Barnes, N. Liu, and J. Han, "Weakly supervised video salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16826–16835, 2021.

[27] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, "Data-free learning of student networks," in *Proc. IEEE Inter. Conf. Comput. Vis.*, pp. 3514–3522, 2019.

[28] Y. Choi, J. Choi, M. El-Khamy, and J. Lee, "Data-free network quantization with adversarial knowledge distillation," in *IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, pp. 710–711, 2020.

[29] L. Luo, M. Sandler, Z. Lin, A. Zhmoginov, and A. Howard, "Large-scale generative data-free distillation," *arXiv preprint arXiv:2012.05578*, 2020.

[30] G. Fang, J. Song, X. Wang, C. Shen, X. Wang, and M. Song, "Contrastive model inversion for data-free knowledge distillation," *arXiv preprint arXiv:2105.08584*, 2021.

[31] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," *Proc. Adv. Neural Inf. process. syst.*, vol. 30, 2017.

[32] H. Thanh-Tung and T. Tran, "Catastrophic forgetting and mode collapse in gans," in *Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1–10, IEEE, 2020.

[33] K. Liu, W. Tang, F. Zhou, and G. Qiu, "Spectral regularization for combating mode collapse in gans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6382–6390, 2019.

[34] D. Bang and H. Shim, "Mggan: Solving mode collapse using manifold-guided training," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2347–2356, 2021.

[35] G. Fang, J. Song, C. Shen, X. Wang, D. Chen, and M. Song, "Data-free adversarial distillation," *arXiv preprint arXiv:1912.11006*, 2019.

[36] P. Micaelli and A. Storkey, "Zero-shot knowledge transfer via adversarial belief matching," *arXiv preprint arXiv:1905.09768*, 2019.

[37] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 8715–8724, 2020.

[38] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and data mining*, pp. 535–541, 2006.

[39] R. G. Lopes, S. Fenu, and T. Starner, "Data-free knowledge distillation for deep neural networks," *arXiv preprint arXiv:1710.07535*, 2017.

[40] G. K. Nayak, K. R. Mopuri, V. Shaj, V. B. Radhakrishnan, and A. Chakraborty, "Zero-shot knowledge distillation in deep networks," in *IEEE Int. Conf. Mach. Learn.*, pp. 4743–4751, PMLR, 2019.

[41] M. Phuong and C. H. Lampert, "Distillation-based training for multi-exit architectures," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1355–1364, 2019.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Pro. Adv. Neural Inf. process. syst.*, vol. 27, 2014.

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1125–1134, 2017.

[44] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4681–4690, 2017.

[45] F. Liu, L. Jiao, and X. Tang, "Task-oriented gan for polsar image classification and clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2707–2719, 2019.

[46] X. Zhang, Y. Zhu, W. Chen, W. Liu, and L. Shen, "Gated switchgan for multi-domain facial image translation," *IEEE Trans. Multimedia*, vol. 24, pp. 1990–2003, 2022.

[47] Z. Yu, Z. Zhang, W. Cao, C. Liu, J. Philip Chen, and H. S. Wong, "Gan-based enhanced deep subspace clustering networks," *IEEE Trans. Knowledge and Data Engineering*, pp. 1–1, 2020.

[48] H. Tan, X. Liu, B. Yin, and X. Li, "Cross-modal semantic matching generative adversarial networks for text-to-image synthesis," *IEEE Trans. Multimedia*, vol. 24, pp. 832–845, 2022.

[49] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2726–2737, 2019.

[50] Z. Zhang, M. Li, H. Xie, J. Yu, T. Liu, and C. W. Chen, "Twgan: Twin discriminator generative adversarial networks," *IEEE Trans. Multimedia*, 2021.

[51] Y. Yu, Z. Gong, P. Zhong, and J. Shan, "Unsupervised representation learning with deep convolutional neural network for remote sensing images," in *Proc. Int. Conf. Image Graphics*, pp. 97–108, Springer, 2017.

[52] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1359–1371, 2018.

[53] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks," in *Int. Conf. Artif. Neural Netw.*, pp. 703–716, Springer, 2019.

[54] A. Ghosh, V. Kulharia, V. P. Namboodiri, P. H. Torr, and P. K. Dokania, "Multi-agent diverse generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 8513–8521, 2018.

[55] M. Ermaliuc, D. Stamate, G. D. Magoulas, and I. Pu, "Creating ensembles of generative adversarial network discriminators for one-class classification," in *Int. Conf. Engineering Applications of Neural Networks*, pp. 13–23, Springer, 2021.

[56] T. Bolukbasi, J. Wang, O. Dekel, and V. Saligrama, "Adaptive neural networks for efficient inference," in *Proc. Int. Conf. Mach. Learn.*, pp. 527–536, PMLR, 2017.

[57] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proc. Int. Conf. Comput. Vis.*, pp. 2736–2744, 2017.

[58] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[59] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment," *arXiv preprint arXiv:1908.09791*, 2019.

[60] Z. Wu, T. Nagarajan, A. Kumar, S. Rennie, L. S. Davis, K. Grauman, and R. Feris, "Blockdrop: Dynamic inference paths in residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 8817–8826, 2018.

[61] J. Lin, Y. Rao, J. Lu, and J. Zhou, "Runtime neural pruning," *in Proc. Adv. Neural Inf. process. syst.*, vol. 30, 2017.

[62] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5325–5334, 2015.

[63] A. Almahairi, N. Ballas, T. Cooijmans, Y. Zheng, H. Larochelle, and A. Courville, "Dynamic capacity networks," in *Proc. Int. Conf. Mach. Learn.*, pp. 2549–2558, PMLR, 2016.

[64] Z. Zhang, Z. Li, H. Liu, and N. N. Xiong, "Multi-scale dynamic convolutional network for knowledge graph embedding," *IEEE Trans. Knowledge and Data Engineering*, vol. 34, no. 5, pp. 2335–2347, 2022.

[65] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[66] A. Shahroudnejad, "A survey on understanding, visualizations, and explanation of deep neural networks," *arXiv preprint arXiv:2102.01792*, 2021.
[67] H. Sun, Y. Chen, A. Aved, and E. Blasch, "Collaborative multi-object tracking as an edge service using transfer learning," in *Int. Conf. High Performance Computing and Communications*, IEEE, 2020.
[68] A. Chougule, H. Tupsamudre, and S. Lodha, "Revelio: A lightweight captcha solver using a dictionary based approach," in *Int. Conf. Information Systems Security*, pp. 97–116, Springer, 2020.
[69] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[70] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.
[71] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng, "The street view house numbers (svhn) dataset," tech. rep., Technical report, Accessed 2016-08-01.[Online], 2018.
[72] K. Binici, N. T. Pham, T. Mitra, and K. Leman, "Preventing catastrophic forgetting and distribution mismatch in knowledge distillation via synthetic data," in *Pro. Conf. Applicat. Comput. Vis.*, pp. 663–671, 2022.
[73] K. Binici, S. Aggarwal, N. T. Pham, K. Leman, and T. Mitra, "Robust and resource-efficient data-free knowledge distillation by generative pseudo replay," in *Proc. Conf. Artif. Intell.*, pp. 6089–6096, AAAI Press, 2022.
[74] G. Fang, K. Mo, X. Wang, J. Song, S. Bei, H. Zhang, and M. Song, "Up to 100x faster data-free knowledge distillation," in *Proc. Conf. Artif. Intell.*, vol. 36, pp. 6597–6604, 2022.
[75] Y. Kim, D. Park, D. Kim, and S. Kim, "Naturalinversion: Data-free image synthesis improving real-world consistency,"

Xiufang Li received the master's degree in safety science and engineering from Xian University of Science and Technology, Xi'an, China in 2017. She is currently pursuing the Ph.D. degree in circuit and system from Xidian University, Xi'an China.

Currently, she is a member of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, and international Research Center for Intelligent Perception and computation, Xidian University, Xi'an China. Her research interests include deep learning and images processing.

Qigong Sun received the B.S. degrees in intelligence science and technology from Xidian University, Xi'an, China in 2015. He received the Ph.D. degrees from circuit and system from Xidian University, Xi'an, China in 2021.

He is currently the director of Applied Research Laboratory of SenseTime. He is also with Shanghai AI Laboratory, Shanghai, China. His current research interests include machine learning and image processing.

Licheng Jiao (SM'89-F'17) received the B.S.degree from Shanghai Jiaotong University, Shanghai, China, in 1982 and the M.S. and Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

Since 1992, he has been a Professor with the school of Electronic Engineering, Xidian University, Xi'an, where he is currently the Director of Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. He is in charge of about 40 important scientific research projects and has published more than 20 monographs and a hundred papers in international journals and conferences. Her research interests include image processing, natural computation, machine learning, and intelligent information processing.

Dr. Jiao is a member of the IEEE Xi'an Section Execution Committee, the Chairman of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, a Councilor of the Chinese Institute of Electronics, a committee member of the Chinese Committee of Neural Networks, and an expert of the Academic Degrees Committee of the State Council.

Fang Liu (SM'07) received the B.S. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 1984 and the M.S. degree in computer science and technology from Xidian University, Xi'an, in 1995. She is currently a Professor with the School of Computer Science, Xidian University. She is the author or coauthor of 5 books and more than 80 papers in journals and conferences. Her research interests include signal and image processing, synthetic aperture radar image processing, multiscale geometry analysis, learning theory and algorithms, optimization problems, and data mining.

Xu Liu (Member, IEEE) received the B.S. degrees in Mathematics and applied mathematics from North University of China, Taiyuan, China in 2013. He received the Ph.D. degrees from Xidian University, Xi'an, China, in 2019. He is currently associate professor of Huashan elite and postdoctoral researcher of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, China. He is the chair of IEEE Xidian university student branch(2015-2019). His current research interests include machine learning and image processing.

Lingling Li (SM'22) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2011 and 2017, respectively. From 2013 to 2014, she was an exchange Ph.D. student with the Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Spain. She is an associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, School of Artificial Intelligence, Xidian University. Her research interests include image processing, deep learning, and pattern recognition.

Puhua Chen ((Member, IEEE) received the B.S. degree in environmental engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the Ph.D. degree in circuit and system from Xidian University, Xi'an, China, in 2016. She is currently an associate professor with the School of Artificial Intelligence, Xidian University. Her current research interests include machine learning, pattern recognition and remote sensing image interpretation.

Yi Zuo received the B.S. degree in Economics and Management from XiDian University, Xi'an, China, in 2022. He is currently pursuing a Master's degree in the School of Artificial Intelligence at XiDian University. His research interests include image processing, natural computation, machine learning, and intelligent information processing.