scientific reports

Check for updates

OPEN

Data free knowledge distillation with feature synthesis and spatial consistency for image analysis

Pengchen Liang[1,2,7], Jianguo Chen[3,7], Yan Wu[4,7], Bin Pu[5], Haishan Huang[3], Qing Chang[6]✉ & Guo Ran[1]✉

Privacy and security concerns restrict access to original training datasets, posing significant challenges for model compression. Data-Free Knowledge Distillation (DFKD) emerges as a solution, aiming to transfer knowledge from teacher to student networks without accessing original data. Existing DFKD methods struggle to generate high-quality synthetic samples that capture the complexities of real-world data, leading to suboptimal knowledge transfer. Moreover, these approaches often fail to preserve the spatial attributes of the teacher network, resulting in shortcut learning and limited generalization.To address these issues, a novel DFKD strategy is proposed with three innovations: (1) an enhanced DCGAN generator with an attention module for synthesizing samples with improved micro-discriminative features; (2) a Multi-Scale Spatial Activation Region Consistency (MSARC) mechanism to accurately replicate the teacher's spatial attributes; and (3) an adversarial learning framework that creates a dynamic competitive environment between the generative and distillation phases. Rigorous evaluation of the method on several benchmark datasets, including CIFAR-10, CIFAR-100, Tiny-ImageNet, and medical imaging datasets such as PathMNIST, BloodMNIST, and PneumoniaMNIST, demonstrates superior performance compared to existing DFKD methods. Specifically, on CIFAR-100, the student network attains an accuracy of 77.85%, surpassing previous methods like CMI and SpaceshipNet. On BloodMNIST, the method achieves an accuracy of 80.50%, outperforming the next best method by over 5%.

Keywords Data-free knowledge distillation, Enhanced DCGAN with attention, Multi-scale spatial activation consistency, Adversarial learning

The paradigm shift towards innovative model compression techniques has been significantly catalyzed by breakthroughs in knowledge distillation, which has witnessed a flurry of activity and advancement[1,2]. Knowledge distillation transcends the traditional boundaries of model compression by facilitating knowledge transfer from a larger, complex teacher network to a smaller, more efficient student network. This process not only reduces the model size but also retains or even enhances the performance of the compressed model, making it a highly sought-after technique for model compression. However, the conventional approach to knowledge distillation often requires direct access to the original, often large and privacy-sensitive, datasets used to train the teacher model. This prerequisite poses significant challenges in sectors where data privacy and security are paramount, such as healthcare, finance, and personal services, where regulatory and ethical considerations demand strict confidentiality of data[3–5]. The sensitivity of data in these domains has propelled the exploration of Data-Free Knowledge Distillation (DFKD), an innovative subset of knowledge distillation that seeks to mitigate privacy concerns by obviating the need for original training data in the knowledge transfer process. DFKD represents a paradigm shift, offering a promising pathway to model compression that harmonizes with the rigorous data privacy standards of sensitive sectors, thereby unlocking new possibilities for the deployment of advanced neural networks in privacy-centric applications[6–11].

[1]The Department of Anesthesiology, Eye & ENT Hospital, Fudan University, Shanghai 200031, China. [2]School of Microelectronics, Shanghai University, Shanghai 201800, China. [3]School of Software Engineering, Sun Yat-sen University, Zhuhai 519000, Guangdong, China. [4]Huangdu Community Health Service Center, Shanghai 201800, China. [5]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China. [6]The Department Shanghai Key Laboratory of Gastric Neoplasms, Department of Surgery, Shanghai Institute of Digestive Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 201800, China. [7]These authors contributed equally: Pengchen Liang, Jianguo Chen and Yan Wu. ✉email: robie0510@hotmail.com; ranguo@eentanesthesia.com

However, this shift towards DFKD introduces its own set of challenges, particularly in synthesizing high-quality training data and effectively transferring complex knowledge from the teacher to the student network[12–16]. Conventional DFKD techniques typically utilize generators that produce synthetic images lacking critical micro-discriminative features, leading to subpar student network training, particularly in tasks requiring high-level feature discrimination[17,18]. Additionally, these methods often fall short in enabling the student network to accurately replicate the complex spatial features of the teacher network[19]. This shortfall will result in shortcut learning, where the student network fails to fully comprehend the data, thus hindering its ability to generalize to new, unseen data[19].

To address these shortcomings, this research addresses these gaps by introducing a refined DFKD approach that harnesses an improved DCGAN generator with an attention module, a Multi-Scale Spatial Activation Region Consistency (MSARC) mechanism, and an adversarial learning framework to overcome these challenges. The principal contributions of this investigation are delineated herein:

- A DCGAN generator with an attention module is employed to produce high-quality synthetic samples. These samples exhibit micro-discriminative features that mimic real images, facilitating effective knowledge transfer.
- The Multi-Scale Spatial Activation Region Consistency (MSARC) mechanism is proposed to ensure accurate capture of the teacher network's spatial features, addressing shortcut learning and enhancing generalization to unseen data.
- An adversarial learning framework is integrated, creating a dynamic interplay between synthetic data generation and knowledge distillation. This produces challenging samples that encourage deeper learning by the student network.
- The method is validated on diverse natural and medical image datasets, demonstrating versatility and effectiveness across various classification tasks.

Related work

Data-free knowledge distillation

DFKD has emerged as a pivotal approach to address the unavailability of original training datasets, especially in scenarios where data privacy and security are of paramount concern. Early DFKD techniques mainly focused on approximating the original dataset distribution using various statistical methods[19]. Notable among these are approaches that utilize the activation patterns of the teacher network to generate synthetic samples. To address the dependency on original data during the training process[20], proposed a strategy of storing some metadata during training and reconstructing training samples in the distillation phase. On the other hand[21], proposed a method of using Data Impressions (DI), created from random noise images, as substitute training data. In their study, the softmax space is modeled as a Dirichlet distribution and random noise images are optimized to generate data for training. Yin et al.[22] developed DeepInversion, a method to synthesize realistic images from a network's training distribution without needing original data, using a fixed teacher model and optimizing inputs with batch normalization.

More innovations in DFKD have explored the use of generative models, such as Generative Adversarial Networks (GANs), to create more realistic synthetic datasets[6,23–25]. FastDFKD[15], a method that accelerates DFKD by up to 100 times through a novel meta-synthesizer for efficient data synthesis, maintaining performance on CIFAR, NYUv2, and ImageNet. Chen et al.[24] introduced Data-Free Multi-Student Coevolved Distillation (DF-MSCD), an approach to improve Data-Free Knowledge Distillation by simultaneously distilling knowledge to multiple heterogeneous student models. This method addresses class imbalance, enhances interactions between teacher-student pairs and peer students, and employs multiple generators for diverse sample synthesis. CDFKD-MFS[26], a collaborative Data-Free Knowledge Distillation framework that compresses multiple teacher models into a compact student model without original data, demonstrated superior accuracy across various datasets. Despite advancements, producing synthetic data that mirrors real-world complexity remains a challenge for effectively training student networks in data-free environments. Inspired by the fine-grained visual classification[27], we integrated an attention module into the generator, aiming to capture the diversity and intricacy of real datasets more effectively.

Feature-level knowledge distillation

Feature-level knowledge distillation emerged as an area of significant interest, focusing on the transfer of rich, intermediate representations from the teacher to the student network[28,29]. This approach extended beyond traditional output-level knowledge transfer, enabling a deeper mimicry of the teacher network's behavior[30]. Techniques such as attention transfer and feature map matching proved effective in capturing the nuanced knowledge embedded in the intermediate layers of deep neural networks[5]. An efficient knowledge distillation[29] method that leveraged an attention-based meta-network to autonomously evaluate and utilize the relative similarities across all feature levels between teacher and student networks, facilitating optimal feature distillation without manual link selection. Chen et al.

[31] proposed a novel approach to knowledge distillation that introduced cross-stage connection paths between teacher and student networks, emphasizing the importance of the connection path across levels. This knowledge review mechanism was both effective and structurally efficient, leading to superior performance across diverse tasks. FAKD[32], which employed feature-level augmentations and novel surrogate loss functions for knowledge distillation in semantic segmentation, leading to improved performance on various benchmarks without substantial overhead. However, optimizing these feature-level distillation techniques to work efficiently in data-free environments presented unique challenges, particularly in maintaining the balance between transfer efficiency and model complexity. In light of these developments, this research aimed to contribute to the field of

DFKD by addressing the challenges in synthetic data generation and feature-level knowledge transfer. Inspired by SpaceshipNet[33], the proposed method combined an advanced generative model with sophisticated feature-level distillation techniques to enhance the effectiveness of knowledge transfer in data-free settings.

DFKD applications in medical imaging

The application of Data-Free Knowledge Distillation (DFKD) within the realm of medical imaging represents a critical frontier in the advancement of healthcare technologies[34]. Medical images, such as X-rays, MRIs, and CT scans, are pivotal in diagnosing and monitoring a wide array of conditions, yet they pose significant challenges related to data privacy, accessibility, and the need for high-fidelity analysis[35,36]. The sensitivity of medical data and the ethical implications of its use necessitate innovative approaches that can circumvent these challenges while maintaining or enhancing the diagnostic capabilities of deep learning models[37–42]. The primary challenge in applying DFKD to medical imaging is the generation of synthetic images that faithfully replicate the complex, varied, and highly nuanced features of real medical images. These images must not only preserve the pathological features critical for diagnosis but also adhere to the diversity found across different patients, conditions, and imaging modalities. Recent advancements have showcased the potential of DFKD in generating synthetic images that can be used for training deep learning models without compromising patient privacy or data security[23]. For instance, generative adversarial networks (GANs) have been tailored to produce high-quality synthetic medical images that mimic the distribution of real datasets[43,44]. These synthetic datasets enable the training of robust diagnostic models without direct access to sensitive or proprietary medical data. The implications of DFKD in medical imaging are profound. It enables the development and refinement of diagnostic models in environments where access to large-scale medical datasets is impractical or impossible. This is particularly relevant for rare diseases, where available data is scarce, or in low-resource settings where data collection faces logistical and ethical hurdles.

Proposed method

Problem formulation

In the domain of DFKD, a critical question arises: How can we effectively transfer intricate knowledge from a complex teacher network to a student network without access to the original training data? This transfer is challenging due to the absence of real data that typically guides the student network's learning process.

To formalize this, let's consider a teacher network T that has been trained on a dataset D with samples $x \in X$ and labels $y \in Y$. The goal of traditional knowledge distillation is to train a student network S to approximate the function learned by T. The process involves minimizing a loss function that measures the difference between the outputs of S and T. This is commonly represented as the minimization of the distillation loss, as defined as:

$$\mathcal{L}_{KD} = \mathcal{L}(S(x), T(x)), \tag{1}$$

where $\mathcal{L}$ typically denotes a loss function, such as the Kullback-Leibler divergence or Mean Squared Error, measuring the discrepancy between the outputs of the student network $S(x)$ and the teacher network $T(x)$.

However, in DFKD, D is unavailable. In this study, one instead has a generator G that synthesizes data $\hat{x} = G(z)$, where z is a random noise vector. The challenge now is to optimize G alongside S such that the synthetic data $\hat{x}$ is effective for distilling knowledge from T to S. This optimization problem can be represented as:

$$\min_{G,S} \mathcal{L}_{KD}(S(G(z)), T(G(z))) + \mathcal{L}_G, \tag{2}$$

where $\mathcal{L}_{KD}$ is a knowledge distillation loss, and $\mathcal{L}_G$ is a loss term that ensures the fidelity of the generated samples $G(z)$ to real data characteristics.

The proposed framework addresses this question by enhancing the capability of G through an advanced DCGAN generator with an integrated attention module and optimizing S using a novel Multi-Scale Spatial Activation Region Consistency (MSARC) mechanism. This approach aims to produce high-quality synthetic samples that capture the necessary features for effective knowledge transfer, thereby facilitating efficient and robust learning in the student network, even in the absence of real training data. The overall architecture of the proposed method is illustrated in Fig. 1.

Enhanced generator with attention

The approach innovates upon the standard Deep Convolutional Generative Adversarial Network (DCGAN) architecture, focusing on the synthesis of discriminative features essential for effective Data-Free Knowledge Distillation. The integration of an attention module within the DCGAN framework plays a pivotal role in this enhancement.

Architecture and feature enhancement

The architecture of DCGAN is divided into distinct blocks, each equipped with a specially designed attention module. This configuration allows for a targeted approach to feature synthesis at different layers of the generator. The primary objective is to generate synthetic samples $\hat{x}$ from a noise vector z, which can be described as:

$$\hat{x} = G\left(z; \Theta_G\right), \tag{3}$$

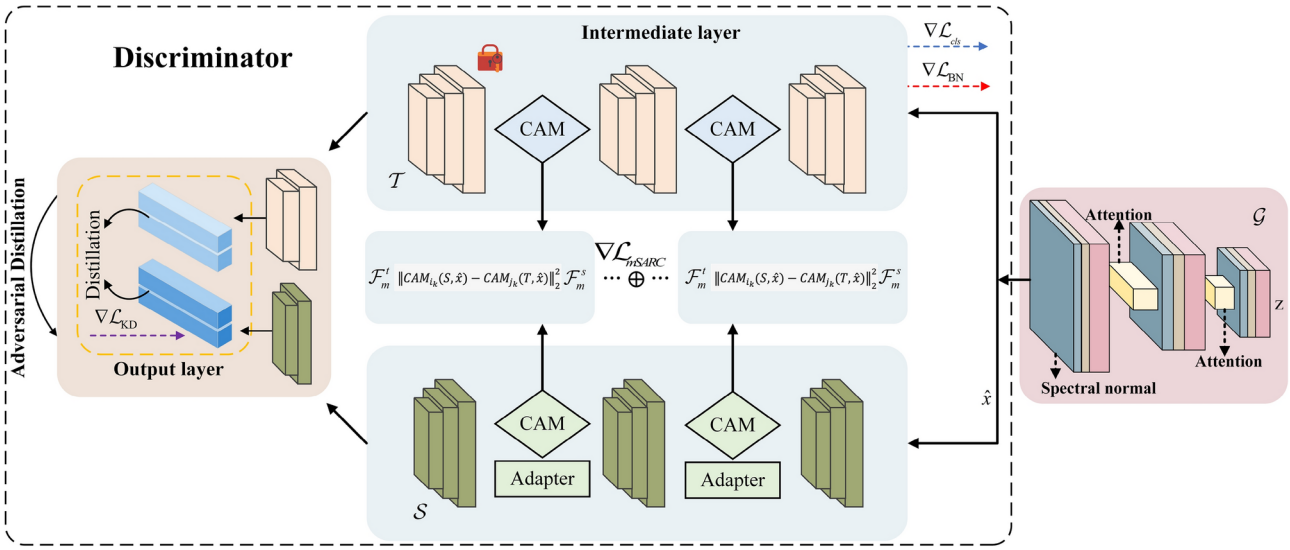

Fig. 1. Schematic of the proposed data-free knowledge distillation (DFKD) framework. The architecture integrates adversarial distillation at the output layer with feature-level knowledge transfer through Class Activation Maps (CAM) at intermediate layers. The discriminator differentiates between the teacher (T) and student (S) network outputs, while CAMs guide the distillation process by aligning intermediate representations. The generator (G) employs attention mechanisms and spectral normalization to synthesize high-quality data for training the student network. The MSARC gradient, $\nabla \mathcal{L}_{mSARC}$, optimizes the alignment of spatial activation regions, enhancing the fidelity of knowledge transfer.

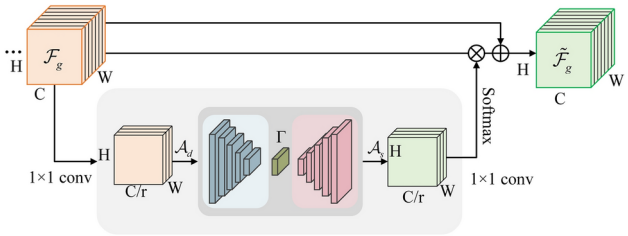

Fig. 2. Spatial attention module of the enhanced DCGAN generator. In this diagram, the symbol $\otimes$ denotes element-wise multiplication, and $\oplus$ represents element-wise addition.

where G represents the generator with parameters Θ_G. Spatial attention mechanism (as shown in Fig. 2): the attention mechanism employed in this model is akin to an encoder-decoder structure, focusing on the contextual interplay within the features. Each attention module processes the input feature map $F_g \in \mathbb{R}^{C\times H\times W}$ to produce an attention map A_s. The process is defined as:

$$\begin{aligned} A_d &= \mathrm{Conv}^{1\times1}(F_g) \\ \Psi &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}^{3\times3}(A_d))) \\ \Gamma &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}^{3\times3}(\mathrm{MP}(\Psi)))), \end{aligned} \tag{4}$$

where $\mathrm{Conv}^{1\times1}$ and $\mathrm{Conv}^{3\times3}$ are convolution operations with kernel sizes 1×1 and 3×3, respectively, MP is the max-pooling operation, and BN represents batch normalization.

Attention map generation and fusion

The attention map A_s is obtained through up-sampling and convolutional operations applied to Γ:

$$\begin{aligned} \Psi' &= \mathrm{MUP}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{DC}^{3\times3}(\Gamma)))) \\ A_s &= \mathrm{Conv}^{1\times1}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{DC}^{3\times3}(\Psi')))), \end{aligned} \tag{5}$$

where $\mathrm{DC}^{3\times3}$ represents a 3×3 kernel deconvolution, and MUP denotes max unpooling. The synthetic features are then refined by fusing A_s with F_g :

$$F'_g = \lambda \cdot (\text{Softmax}(A_s) \otimes F_g) + F_g, \tag{6}$$

where λ is a hyperparameter that balances the influence of the attention map. Through this advanced architecture and attention-based feature enhancement, the DCGAN generator produces synthetic samples that effectively embody the intricate characteristics necessary for distilling knowledge from a teacher network in data-free scenarios.

To enhance the capability of the generator within the DCGAN architecture, the proposed method has integrated a meta-learning optimization strategy[15]. This innovative approach enables the generator to rapidly adapt to new tasks or data distributions, significantly improving the synthesis of discriminative features essential for effective Data-Free Knowledge Distillation.

Meta-learning strategy
The essence of the meta-learning implementation lies in optimizing the generator's parameters Θ_G in a way that facilitates quick adaptation. Specifically, the proposed method employs a model-agnostic meta-learning (MAML) approach, which is designed to prepare the generator for fast learning with a minimal number of gradient updates. This is achieved by training the generator not only to perform well on a given task but also to ensure that its parameters are positioned in a part of the parameter space where the model has high learning adaptability. The optimization can be formalized as follows:

$$\Theta^*_G = \arg\min_{\Theta_G} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}\left(f_{\Theta_G}\right) + \gamma \cdot \Omega(\Theta_G), \tag{7}$$

where $\mathcal{T}_i$ represents tasks sampled from a distribution of tasks $p(\mathcal{T})$, $\mathcal{L}_{\mathcal{T}_i}$ is the loss function associated with task $\mathcal{T}_i$, $\Omega(\Theta_G)$ denotes a regularization term, and γ is a weighting coefficient.

Adaptation phase
During the adaptation phase, the generator undergoes a rapid fine-tuning process, leveraging a small set of examples from a new task. This process consists of one or more gradient updates to the parameters, starting from the meta-learned initial parameters Θ^*_G. The update rule is given by:

$$\Theta'_G = \Theta^*_G - \eta \nabla_{\Theta^*_G} \mathcal{L}_{\mathcal{T}_{\text{new}}}\left(f_{\Theta^*_G}\right), \tag{8}$$

where $\mathcal{T}_{\text{new}}$ is the new task, $\mathcal{L}_{\mathcal{T}_{\text{new}}}$ is the loss on this new task, and η is the learning rate for adaptation.

Through the integration of meta-learning, the generator not only becomes capable of producing high-quality synthetic samples but also acquires the flexibility to quickly adjust to new and unseen tasks. This dual capability is crucial for deploying the DCGAN framework in Data-Free Knowledge Distillation scenarios, where the ability to generate task-specific synthetic data can significantly enhance the distillation process.

Multi-scale spatial activation region consistency

The Multi-Scale Spatial Activation Region Consistency (MSARC) mechanism is a key element of the proposed framework, designed to enhance the student network's replication of the teacher network's spatial features. This is particularly crucial when both networks are trained on synthetic data, ensuring the preservation of spatial hierarchies and contextual details in the distillation process.

Mechanism overview
The MSARC mechanism aligns Class Activation Maps (CAMs) from the teacher network *T* with those from the student network *S*, using synthetically generated data as input. This alignment ensures that the student network learns and mimics the spatial feature distributions of the teacher across multiple scales. The alignment is mathematically expressed as:

$$L_{MSARC} = \sum_{l=1}^{L} \alpha_l \cdot \text{MSE}(\text{CAM}^l_S(\hat{x}), \text{CAM}^l_T(\hat{x})), \tag{9}$$

where *L* is the total number of layers selected for alignment, CAM^l_S and CAM^l_T are the CAMs at layer *l* for the student and teacher networks, respectively, and α_l are weighted for each layer's contribution to the loss.

CAM extraction
For both the teacher (T) and student (S) networks, CAMs are computed from synthetic input samples $\hat{x}$. As shown in Fig. 3, the method varies based on the layer depth[33]:

Deep layers (final convolutional layer before pooling). CAM for class *c* at layer *l* is:

$$\text{CAM}^c_l(\hat{x}) = \sum_p w^c_p \cdot f^l_p(\hat{x}), \tag{10}$$

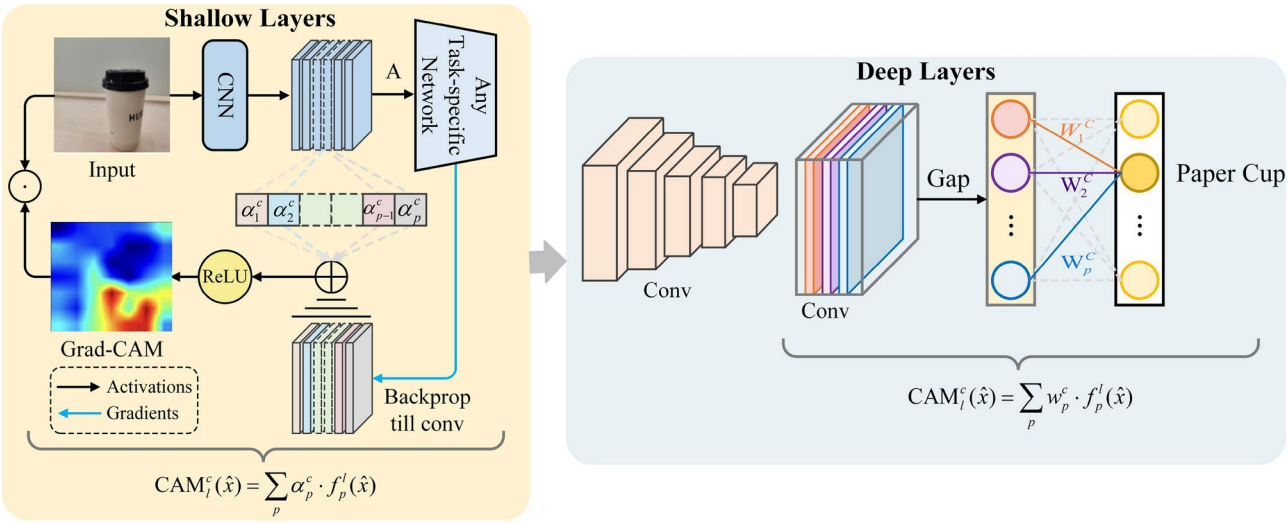

Fig. 3. Comparative visualization of Class Activation Mapping (CAM) techniques for deep and shallow network layers, as illustrated in the schematic diagram. For shallow layers, Grad-CAM is utilized where gradients and activations produce CAMs through backpropagation, as shown on the left. In contrast, for deep layers, CAMs are derived from the final convolutional layer's feature maps weighted by class-specific parameters from the fully-connected layer, depicted on the right.

where $f_p^l(\hat{x})$ is the feature map of channel p at layer l, and w_p^c is the class-specific weight from the fully-connected layer.

Shallow layers (earlier convolutional layers). CAM for class c at layer l uses Grad-CAM inspired method:

$$\mathrm{CAM}_l^c(\hat{x}) = \sum_p \alpha_p^c \cdot f_p^l(\hat{x}), \tag{11}$$

where α_p^c is the gradient-based importance weight for class c with respect to feature maps f_p^l.

To provide a more detailed understanding of the Grad-CAM process used in shallow layers, we present the step-by-step computation and visualization procedure:

1. For a given class c and layer l, we first calculate the gradient of the score for class c with respect to the feature maps of layer l:

$$\alpha_p^c = \frac{\partial y^c}{\partial f_p^l} \tag{12}$$

where y^c is the score for class c, and f_p^l is the feature map of channel p in layer l.

2. These gradients are then global average pooled to obtain the importance weights for each channel:

$$\alpha_p^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{,} \partial f_{ij}^l \tag{13}$$

where Z is the total number of pixels in the feature map.

3. Finally, we compute the CAM for class c by taking a weighted sum of the feature maps:

$$\mathrm{CAM}_l^c(\hat{x}) = \mathrm{ReLU}\left(\sum_p \alpha_p^c \cdot f_p^l(\hat{x})\right) \tag{14}$$

4. For visualization, the CAM is upsampled to the size of the original image and overlaid on it, typically displayed as a heatmap. The highlighted areas in the heatmap indicate regions of the image that contribute significantly to the prediction of the given class.This method allows us to effectively visualize class-specific activation regions even in the shallow layers of the network, thereby achieving spatial activation region consistency across multiple scales.

This approach ensures accurate CAM computation across different network layers, reflecting the feature importance for classification tasks.

Scale-wise consistency
The Mean Squared Error (MSE) is used to measure the similarity between the CAMs of the teacher and student networks. The consistency loss at each layer *l* is calculated as:

$$\mathrm{MSE}(\mathrm{CAM}_S^l, \mathrm{CAM}_T^l) = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{CAM}_{S_i}^l - \mathrm{CAM}_{T_i}^l)^2, \tag{15}$$

where *N* is the number of spatial elements in the CAMs. By implementing the MSARC mechanism, the approach ensures that the student network learns a spatially consistent representation from the teacher network, thereby improving performance in complex classification tasks and enhancing generalization capability, particularly with synthetic data training.

Adversarial distillation and training objectives

The optimization process involves two primary components: the generator *G* and the student network *S*, each with its distinct objective function. Their training is conducted in an adversarial distillation manner.

Generator optimization
The initial phase is concentrated on optimizing the generator *G* to synthesize more realistic and diverse samples. The objective function for the generator combines two essential loss terms: L_{BN}, to ensure the quality and diversity of the synthetic data, and L_{cls}, a cross-entropy loss that encourages the generated images $\hat{x} = G(z)$ to be classified into specific categories by the teacher network *T*. In formulating this optimization goal, the proposed method also considered the potential impact of a meta-learning strategy, which aims to enhance the generator's adaptability to new tasks and data distributions through subtle adjustments in its training process. The optimization objective for the generator is expressed as:

$$\min_G \kappa\left(L_{BN} + L_{cls}\right) - L_{KD}, \tag{16}$$

where L_{KD} is the knowledge distillation loss, and κ is a hyperparameter that balances these components.

Student network optimization
After the generation of synthetic data, the next phase involves training *S* using these samples. The total objective for the student network is:

$$\min_S L_{KD} + \beta L_{mSARC}, \tag{17}$$

where L_{KD} is the knowledge distillation loss, L_{mSARC} is the Multi-Scale Spatial Activation Region Consistency loss, and β is a hyperparameter for balancing these losses.

Adversarial distillation approach
The training utilizes an adversarial distillation strategy, where both the generator and the student network are iteratively optimized. This approach enables a dynamic adaptation of the synthetic data and the student network's learning process, ensuring an effective transfer of knowledge from the teacher network.

By alternating between optimizing *G* for improved data synthesis and *S* for enhanced knowledge absorption, the algorithm strikes a balance between data realism and learning efficacy. This approach effectively addresses the challenges of Data-Free Knowledge Distillation, leading to a student network that is both robust and capable of generalizing well.

Theoretical analysis of synthetic data impact

The distillation process involves a loss function, quantifying the difference between the student and teacher models. This loss is expressed as:

$$\mathcal{L}_{KD}(S, T, G) = \int_{\mathcal{Z}} \|S(G(z)) - T(G(z))\|^2 \rho(z) dz, \tag{18}$$

where Z is the domain of the noise vector z and $\rho(z)$ is the probability density function over z. between Data Quality and Distillation Loss

To establish the relationship between the quality of synthetic data and the distillation loss, a metric $\delta(G, G^*)$ is introduced, which quantifies the discrepancy between the actual generator G and the ideal generator G^*. This discrepancy is quantified as:

$$\delta(G, G^*) = \sup_{z \in Z} \|G(z) - G^*(z)\|, \tag{19}$$

representing the maximum deviation between the outputs of G and G^* across all possible noise vectors z. This assumption indicates that the difference between S and T on ideal synthetic data $G^*(z)$ is bounded by a constant C, i.e., for all $z \in Z$, it holds that:

$$\|S(G^*(z)) - T(G^*(z))\| \leq C. \tag{20}$$

For the actual generator G,the following inequality holds:

$$\begin{aligned}
&\|S(G(z)) - T(G(z))\| \\
&\leq \|S(G(z)) - S(G^*(z))\| + \|S(G^*(z)) - T(G^*(z))\| \\
&+ \|T(G^*(z)) - T(G(z))\| \\
&\leq \|S(G(z)) - S(G^*(z))\| + C + \|T(G^*(z)) - T(G(z))\| \\
&\leq \|S\|_{Lip} \cdot \|G(z) - G^*(z)\| + C + \|T\|_{Lip} \cdot \|G(z) - G^*(z)\| \\
&= (C + (\|S\|_{Lip} + \|T\|_{Lip}) \cdot \delta(G, G^*)),
\end{aligned} \tag{21}$$

where $\|S\|_{Lip}$ and $\|T\|_{Lip}$ are the Lipschitz constants of the student and teacher models, respectively, indicating the maximum sensitivity of these models to changes in their inputs. This demonstrates that the upper bound on the discrepancy between the outputs of the student and teacher models is determined not only by their inherent performance gap C but also by the disparity between the generator G and the ideal generator G^*. When incorporating this inequality into the distillation loss expression, the following result is obtained:

$$\mathcal{L}_{KD}(S, T, G) \leq \int_{\mathcal{Z}} (C + (\|S\|_{Lip} + \|T\|_{Lip}) \cdot \delta(G, G^*))^2 \rho(z) dz. \tag{22}$$

This reveals a key insight: as the discrepancy $\delta(G, G^*)$ diminishes (i.e., as G approaches the ideal), the upper bound of the distillation loss $\mathcal{L}_{KD}(S, T, G)$ decreases accordingly. Enhancing the quality of synthetic data by reducing the difference between G and G^* effectively lowers the output discrepancy between the student and teacher models, thereby potentially improving the performance of the student model.

Experiments

Experimental setting

In the comprehensive evaluation, a diverse range of backbone networks was employed to rigorously assess the effectiveness of the proposed method. Specifically, ResNet[45], known for its deep residual learning framework, VGG[46], recognized for its simplicity and depth, and Wide ResNet[47], which widens ResNet's layers to provide a different architectural perspective, were utilized. The key hyperparameters in the experiment are shown in Table 1.

The experiments were methodically conducted across three benchmark classification datasets: CIFAR-10[48], CIFAR-100[48], and Tiny-ImageNet[49]. CIFAR-10 and CIFAR-100 are staple datasets in machine learning, comprising low-resolution images (32×32 pixels) across 10 and 100 classes, respectively. They provide a controlled environment to evaluate the method's effectiveness in handling a wide range of classes. Tiny-ImageNet,

Hyperparameter	Value	Description
Learning Rate for Generator	0.005	Learning rate for optimizing the generator network parameters
Learning Rate for Noise Vector	0.015	Learning rate for optimizing the noise vector input to the generator
Temperature for Distillation	20.0	Temperature parameter to soften the teacher's output probabilities
Batch Size	128	Number of samples per batch during training
Attention Module Weight	3×10^{-2}	Balances the influence of the attention map in the generator
Generator Optimization Steps	2	Number of optimization steps for the generator per iteration
Use of Meta-Learning	Enabled	Meta-learning is used in the generator's optimization
Momentum	0.9	Momentum factor used in the optimizer for training
Weight Decay	0.0001	Weight decay (L2 regularization) coefficient used in the optimizer

Table 1. Key hyperparameters used in the proposed methods.

a subset of the larger ImageNet dataset, contains 200 classes, offering a more challenging classification task with its increased number of classes and slightly higher resolution images (64 × 64 pixels). To further evaluate the performance of the proposed method on datasets with higher image resolutions, experiments were extended to Imagenette and ImageNet100. Imagenette, a subset of ImageNet, includes a selection of easily classifiable classes from ImageNet[50], and ImageNet100 is a more compact version of ImageNet with 100 classes. Both these datasets present images at a resolution of 224*times*224, posing a more challenging scenario and allowing us to assess the scalability and robustness of the approach in handling high-resolution image data.

The experimental framework leveraged three medical datasets[51,52]-PathMNIST, BloodMNIST, and PneumoniaMNIST-used at their original resolutions (64×64 for PathMNIST and BloodMNIST; 224×224 for PneumoniaMNIST), tailored for distinct classification tasks in healthcare: PathMNIST focuses on classifying 9 types of tissues from colorectal cancer histology images, challenging the model with the complexity of histopathological analysis. BloodMNIST comprises images of normal blood cells categorized into 8 classes, testing the model's accuracy in hematological cell classification. PneumoniaMNIST includes pediatric chest X-rays for binary classification of pneumonia, assessing the model's capability in diagnosing conditions from radiographic imagery. These datasets underscore the method's adaptability and potential impact on medical diagnostics, showcasing effectiveness across varied medical imaging tasks.

To evaluate the performance of the proposed DFKD framework, with a specific focus on classification tasks, accuracy is adopted as the primary metric. Accuracy, in the context of classification, is defined in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which are derived from the confusion matrix. Specifically, accuracy is calculated as the ratio of correctly predicted observations (both true positives and true negatives) to the total observations in the dataset. The formula for accuracy is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{23}$$

where: *TP* (True Positives) are the correctly predicted positive values, which means the predictions that are true and classified as true. *TN* (True Negatives) are the correctly predicted negative values, indicating the predictions that are false and classified as false. *FP* (False Positives) represents the incorrect positive predictions, where the predictions are false but classified as true. *FN* (False Negatives) are the incorrect negative predictions, where the predictions are true but classified as false.

This accuracy metric serves as a straightforward measure of the student network's performance in correctly classifying instances as compared to the ground truth. It provides a comprehensive view of how effectively the student network, trained through the DFKD framework, has managed to replicate the classification capabilities of the teacher network without access to the original training data.

Performance comparison

Table 2 and Fig. 4 show the superior performance of the DFKD method when compared with other contemporary strategies such as DAFL[25], ZSKT[53], ADI[22], DFQ[14], LS-GDFD[54] and CMI[17]. Conducted under uniform conditions with identical teacher networks to ensure fairness, the method showed exemplary performance across all datasets tested, including CIFAR-10, CIFAR-100, and Tiny-ImageNet, demonstrating its effectiveness in extracting and transferring knowledge from a range of teacher networks.

The proposed technique exhibited strong adaptability and proficiency in CIFAR-10 evaluations, achieving competitive results across various teacher-student network configurations, which underscores its versatility in handling complex network architectures. In CIFAR-100, the method effectively distilled subtle and intricate information necessary for detailed classification tasks. Rigorous testing on the challenging Tiny-ImageNet dataset, characterized by higher-resolution images, demonstrated its capability to manage and learn from high-resolution data, essential for applications requiring precise recognition.

Our Data-Free Knowledge Distillation technique also showcased impressive results on high-resolution datasets, particularly Imagenette and ImageNet100, as detailed in Table 3. This performance is especially significant given the increased complexity and higher resolution of these datasets. In the Imagenette experiments, the method achieved a test accuracy of 79.73%, closely trailing the teacher network's accuracy of 80.41%. This outcome is indicative of the efficacy of the approach, as it nearly matches the performance of the teacher network, despite the absence of real training data. The results are even more compelling when compared with other methods like CMI and DFQ, where the approach surpasses them by a notable margin. Similarly, on ImageNet100, the method attained a test accuracy of 62.58%, outperforming the other DFKD techniques.

The effectiveness of the proposed Data-Free Knowledge Distillation approach was rigorously evaluated across three medical datasets: PathMNIST, BloodMNIST, and PneumoniaMNIST. The performance metrics, specifically test accuracy percentages, are summarized in Table 4, showcasing a comparison between this method and other state-of-the-art Data-Free Knowledge Distillation methods, including CMI, DFQ, and SpaceshipNet. The teacher model for these experiments was WRN-40-2, and the student model was WRN-40-1. In the PathMNIST dataset, the method achieved a test accuracy of 77.09%, closely competing with SpaceshipNet's 78.91% and surpassing both CMI's 70.81% and DFQ's 75.60%. This demonstrates strong capability in capturing the nuanced features required for effective histological image classification. For BloodMNIST, the approach outperformed all compared methods with a test accuracy of 80.50%, significantly higher than SpaceshipNet's 75.11%, DFQ's 73.34%, and CMI's 67.08%. This highlights superior performance in classifying blood cell images, attesting to robustness and efficiency in handling diverse medical imaging tasks. PneumoniaMNIST results further validate the effectiveness of this method, achieving a test accuracy of 78.53%. This score is slightly higher than SpaceshipNet's 73.78% and comfortably exceeds the performances of CMI (69.25%). Overall, the experimental

Dataset	Teacher	Student	Test accuracy (%)									
			T.	S.	DAFL	ZSKT	ADI	DFQ	LS-GDFD	CMI	SpaceshipNet	Ours
CIFAR-10	ResNet-34	ResNet-18	95.70	95.20	92.22	93.32	93.26	94.61	95.02	94.84	95.39 (95.27*)	94.91
	VGG-11	ResNet-18	92.25	95.20	81.10	89.46	90.36	90.84	N/A	91.13	92.27 (92.19*)	92.34
	WRN-40-2	WRN-16-1	94.87	91.12	65.71	83.74	83.04	86.14	N/A	90.01	90.38 (90.42*)	90.60
	WRN-40-2	WRN-40-1	94.87	93.94	81.33	86.07	86.85	91.69	N/A	92.78	93.56	92.97
	WRN-40-2	WRN-16-2	94.87	93.95	81.55	89.66	89.72	92.01	N/A	92.52	93.25	93.31
CIFAR-100	ResNet-34	ResNet-18	78.05	77.10	74.47	67.74	61.32	77.01	77.02	77.04	77.41 (77.10*)	77.85
	VGG-11	ResNet-18	71.32	77.10	57.29	34.72	54.13	68.32	N/A	70.56	71.41 (71.56*)	71.43
	WRN-40-2	WRN-16-1	75.83	65.31	22.50	30.15	53.77	54.77	N/A	57.91	58.06 (57.42*)	57.92
	WRN-40-2	WRN-40-1	75.83	72.19	34.66	29.73	61.33	61.92	N/A	68.88	68.78	67.94
	WRN-40-2	WRN-16-2	75.83	73.56	40.00	28.44	61.34	59.01	N/A	68.75	69.95	69.71
Tiny-ImageNet	ResNet-34	ResNet-18	66.44	64.87	N/A	N/A	N/A	63.73	N/A	64.01	64.04 (64.14*)	65.21

Table 2. Comparative results of DFKD on CIFAR-10, CIFAR-100, and Tiny-ImageNet. In this table, T. represents the teacher network and S. denotes the student network, both of which were trained using labeled data from their respective training sets. This notation is consistently applied in the subsequent tables. Benchmark results for DAFL, ZSKT, ADI, DFQ, and LS-GDFD are obtained from[17]. Results marked with an asterisk * are based on executing the code provided in[33], conducted under the experimental settings outlined in[33]. This approach ensures a standardized comparison across various DFKD methodologies.

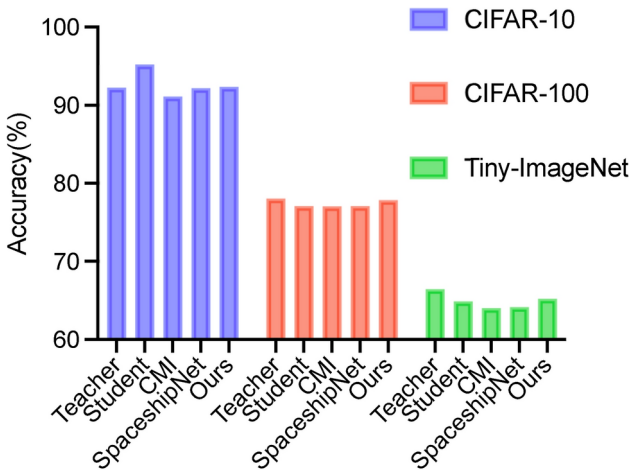

Fig. 4. Comparison of classification accuracy across CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets. The chart illustrates the performance of the teacher network, student network, and other Data-free distillation methods including CMI, SpaceshipNet, and the proposed method.

Dataset	Test accuracy (%)				
	T.	CMI	DFQ	SpaceshipNet	Ours
Imagenette	80.41	74.80	75.31	78.86	79.73
ImageNet100	71.62	56.60	54.73	61.57	62.58

Table 3. The results of the Data-Free Knowledge Distillation technique on higher resolution datasets, specifically Imagenette and ImageNet100. These experiments used ResNet-34 as the teacher network and ResNet-18 as the student network. T. represents the teacher network.

Dataset	Test accuracy (%)				
	T.	CMI	DFQ	SpaceshipNet	Ours
PathMNIST	92.15	70.81	75.60	78.91	77.09
BloodMNIST	96.89	67.08	73.34	75.11	80.50
PneumoniaMNIST	89.25	69.25	76.75	73.78	78.53

Table 4. Comparison of test accuracy (%) on medical datasets using different data-free knowledge distillation methods. The teacher model employed is WRN-40-2, and the student model is WRN-40-1. Our approach exhibits competitive performance across PathMNIST, BloodMNIST, and PneumoniaMNIST, highlighting its efficacy in data-free distillation within the medical imaging domain.

outcomes underscore the competitive performance of the Data-Free Knowledge Distillation method across a variety of medical imaging tasks. Figure 5 visually illustrates the synthetic images generated by this approach, further demonstrating its capability to effectively replicate and synthesize complex medical images for diverse datasets.

Ablation study

Impact of meta-learning optimization in the enhanced DCGAN generator

The ablation study (as shown in Table 5) meticulously contrasts the efficacy of the full method, which incorporates an Enhanced DCGAN Generator with Meta-Learning Optimization, against variations that omit this pivotal enhancement. Notably, when the full method's performance is juxtaposed with that of employing a Standard DCGAN Generator equipped with Meta-Learning Optimization, a discernible decrement in accuracy becomes apparent. Specifically, on the CIFAR-10 dataset, accuracy experiences a downturn from 94.91% to 91.67%, marking a substantial decrease of 3.24 percentage points. This pattern of decline extends across the CIFAR-100 and Tiny-ImageNet datasets as well, with reductions of 4.37% and 3.11%, respectively. However, a more nuanced examination reveals that the variation featuring the Enhanced DCGAN without Meta-Learning Optimization yields accuracies of 92.31%, 75.09%, and 62.96% across CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. This indicates a less pronounced but still significant performance drop compared to the full method, underscoring the integral role of Meta-Learning Optimization in conjunction with the

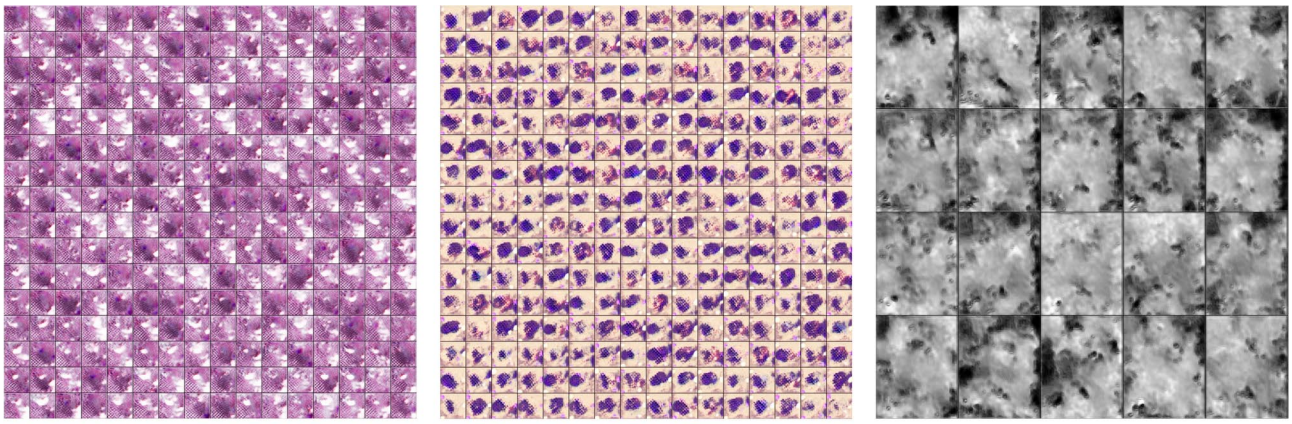

Fig. 5. Visualization of synthetic images generated by the proposed approach on three medical datasets. This Figure showcases the capability of the Data-Free Knowledge Distillation method to produce high-quality synthetic images across diverse medical imaging tasks, including histological images from PathMNIST, blood cell images from BloodMNIST, and pediatric chest X-rays from PneumoniaMNIST.

Component	CIFAR-10 test accuracy (%)	CIFAR-100 test accuracy (%)	Tiny-ImageNet test accuracy (%)
Full methods (enhanced DCGAN with meta-learning optimization + MSARC + adversarial strategy)	94.91	77.85	65.21
Standard DCGAN with meta-learning optimization	91.67	73.48	62.10
Enhanced DCGAN without meta-learning optimization	92.31	75.09	62.96
Without MSARC mechanism	89.82	70.59	57.35
Without adversarial distillation strategy	92.12	75.90	61.75

Table 5. Ablation study results demonstrating the impact of key components on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets with ResNet-34 as the Teacher Network and ResNet-18 as the Student Network. This table compares the performance of the full proposed method with variations where specific components are excluded, highlighting the contribution of each component to the overall effectiveness of the method.

Enhanced DCGAN's architectural modifications. These findings elucidate the paramount importance of Meta-Learning Optimization within the Enhanced DCGAN framework. While the architectural enhancements alone (without Meta-Learning Optimization) contribute positively to the distillation process-evidenced by a smaller performance reduction compared to the Standard DCGAN Generator-the incorporation of Meta-Learning Optimization synergistically amplifies the generator's capability. This optimization enables the generator to produce synthetic data that more effectively encapsulates the complex feature distributions necessary for an efficient and robust knowledge transfer from the teacher to the student network. In essence, the juxtaposition of the Enhanced DCGAN's performance, with and without Meta-Learning Optimization, alongside the Standard DCGAN Generator, provides compelling evidence of the critical role played by Enhanced DCGAN. It not only enhances the quality of synthetic data generation but also significantly bolsters the overall efficacy of the knowledge distillation process, as substantiated by the observed improvements in accuracy across all evaluated datasets.

Role of MSARC mechanism

The exclusion of the MSARC mechanism from the proposed framework starkly highlights its indispensability, as evidenced by significant accuracy declines across CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets. Without MSARC, a marked decrease in accuracy was observed: a decline of 5.09 percentage points for CIFAR-10, 7.26% for CIFAR-100, and 7.86% for Tiny-ImageNet, underscoring the mechanism's pivotal role in enhancing the knowledge distillation process. These results underscore MSARC's critical function in aligning Class Activation Maps (CAMs) between the teacher and student networks, facilitating a nuanced and effective transfer of spatial knowledge. This alignment is paramount, especially in handling datasets characterized by complex and diverse data representations, where grasping the underlying feature distributions is essential for achieving high accuracy. The MSARC mechanism acts as a bridge, narrowing the representational gap between teacher and student models. It enriches the student model's learning process, ensuring not only the acquisition of correct classifications but also the development of a comprehensive understanding of the feature hierarchies that

underpin those classifications. This deeper insight allows the student model to more closely mirror the teacher model's decision-making process, enhancing its generalization capability. In essence, MSARC's contribution to the DFKD framework is transformative, significantly boosting the student model's learning efficiency and accuracy. It ensures that the distillation process meticulously captures and transfers the spatial and contextual nuances intrinsic to the teacher model's representations. The performance degradation observed in its absence reaffirms MSARC's essential role in achieving effective and robust knowledge distillation across challenging datasets.

Effectiveness of adversarial distillation strategy
The adversarial distillation strategy's importance is also evident, with a noticeable performance decline observed in its absence. For example, without this strategy, the accuracy of CIFAR-10 falls to 92.12%, indicating a decrease of 2.79 percentage points. Similarly, on CIFAR-100 and Tiny-ImageNet, the accuracies reduce by 1.95% and 3.46%, respectively. This highlights how the adversarial distillation strategy enhances the learning process by dynamically adapting the synthetic data and the student network's learning, making it an integral component of the knowledge distillation process. By fostering a competitive yet constructive interaction between the generator (producing synthetic data) and the student network, this strategy ensures that the synthetic data evolves in a manner that optimally benefits the student's learning trajectory. This adaptive process is crucial for maintaining a high fidelity of knowledge transfer from the teacher to the student network, especially in the absence of real training data.

Overall, the ablation study validates the necessity and effectiveness of each component in the method. The combined use of the Enhanced DCGAN generator, the MSARC mechanism, and the adversarial distillation strategy leads to notable improvements in performance, demonstrating the synergy of these components in our Data-Free Knowledge Distillation framework.

Ablation study results for λ
The effect of varying the hyperparameter λ within the Enhanced DCGAN Generator with Attention Module was explored. This hyperparameter λ is crucial for balancing the influence of the attention map on the synthetic features generated by the model. The attention mechanism, as outlined in the previous section, is vital for synthesizing samples that closely resemble real data characteristics, a key requirement for successful Data-Free Knowledge Distillation. The results, presented in Table 6, show that λ significantly impacts the accuracy of the student network across the CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets.

The optimal setting for λ is found to be $3e^{-2}$, achieving the highest accuracy rates of 94.91% on CIFAR-10, 77.85% on CIFAR-100, and 65.21% on TinyImageNet. Deviations from this optimal value resulted in a noticeable decrease in model performance. This demonstrates the delicate balance required in the attention mechanism, where λ must be fine-tuned to enhance the quality of synthetic samples without overwhelming the original features of the generated data. These findings underscore the importance of the attention module in the Enhanced DCGAN architecture.

Discussions and limitations

This work presents a comprehensive exploration into the realm of Data-Free Knowledge Distillation (DFKD), demonstrating the efficacy of the proposed method across a spectrum of datasets and network architectures. Rigorous experiments, including an ablation study, validate the significant contributions of the enhanced DCGAN generator, the Multi-Scale Spatial Activation Region Consistency (MSARC) mechanism, and an adversarial distillation strategy. The method's adaptability is further highlighted by its performance on high-resolution datasets and challenging medical imaging tasks, showcasing its potential for wide-ranging applications in scenarios where access to original training data is restricted or impossible.

However, the exploration of DFKD also surfaces inherent limitations and challenges. Firstly, while our approach excels in generating high-quality synthetic data that closely mimics real data distributions, the complexity, and computational overhead associated with the enhanced DCGAN generator and attention mechanism cannot be overlooked. These components, though crucial for capturing nuanced data characteristics, necessitate substantial computational resources, potentially limiting their applicability in resource-constrained environments. Furthermore, the balancing act between the fidelity of generated data and the preservation of informative features for distillation, as controlled by the hyperparameter (λ), underscores the sensitivity of the method to hyperparameter settings. This necessitates careful tuning to achieve optimal performance,

λ Value	CIFAR-10 Test Accuracy (%)	CIFAR-100 Test Accuracy (%)	Tiny-ImageNet Test Accuracy (%)
$1e^{-2}$	94.68	77.35	64.85
$3e^{-2}$	94.91	77.85	65.21
$5e^{-2}$	94.87	77.75	65.15
$7e^{-2}$	94.82	77.65	65.05
$9e^{-2}$	94.75	77.50	64.95

Table 6. Impact of varying hyperparameter (λ) on model accuracy. This table illustrates how different settings of λ, a critical parameter in the attention mechanism of the Enhanced DCGAN Generator, affect the accuracy of CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets.

introducing challenges in scenarios where such fine-tuning is not feasible due to time or computational constraints. Moreover, while the method has shown promising results across diverse datasets, the question of its scalability and efficiency in the face of extremely large and complex datasets remains. The ability to generate synthetic data that accurately reflects the vast diversity and intricate details of such datasets is critical for the success of DFKD methods.

Conclusion

This study introduces a Data-Free Knowledge Distillation methodology that integrates an advanced DCGAN generator with an attention mechanism and the Multi-Scale Spatial Activation Region Consistency (MSARC) mechanism, demonstrating effective high-quality synthetic sample generation and knowledge transfer across a variety of datasets. Through ablation studies, the critical role of each component is underscored, showing significant promise for applications requiring high-detail image classification, such as medical imaging, in environments where data privacy is paramount. Future efforts will aim to adapt and refine this method for even more complex scenarios, addressing the growing challenge of knowledge distillation in the absence of accessible data.

Data availability

The datasets used in this study are publicly available and can be accessed as follows: CIFAR-10 and CIFAR-100 are available at https://www.cs.toronto.edu/~kriz/cifar.html; Tiny-ImageNet can be downloaded from https://github.com/tjmoon0104/pytorch-tiny-imagenet; Imagenette is accessible via https://github.com/fastai/imagenette; ImageNet100 is available at https://github.com /danielchyeh/ImageNet-100-Pytorch; and the PathMNIST, BloodMNIST, and PneumoniaMNIST datasets can be found at https://medmnist.com/.

Code availability

The code that supports the findings of this study is available from the corresponding authors upon reasonable request.

Received: 24 August 2024; Accepted: 4 November 2024
Published online: 11 November 2024

References

1. Rao, J. *et al.* Parameter-efficient and student-friendly knowledge distillation. In *IEEE Transactions on Multimedia* (2023).
2. Rao, J. *et al.* Dynamic contrastive distillation for image-text retrieval. In *IEEE Transactions on Multimedia* (2023).
3. Mahajan, A. & Bhat, A. A survey on application of knowledge distillation in healthcare domain. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 762–768 (IEEE, 2023).
4. Yu, R., Liu, S. & Wang, X. Dataset distillation: A comprehensive review. *arXiv preprint*[SPACE]arXiv:2301.07014 (2023).
5. Wang, L. & Yoon, K.-J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 3048–3068 (2021).
6. Fang, G. *et al.* Data-free adversarial distillation. *arXiv preprint*[SPACE]arXiv:1912.11006 (2019).
7. Li, X., Wang, S., Sun, J. & Xu, Z. Memory efficient data-free distillation for continual learning. *Pattern Recognit.* **144**, 109875 (2023).
8. Liu, B. *et al.* Privacy-preserving student learning with differentially private data-free distillation. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. 01–06 (IEEE, 2022).
9. Zhuang, Y., Lyu, L., Shi, C., Yang, C. & Sun, L. Data-free adversarial knowledge distillation for graph neural networks. *arXiv preprint*[SPACE]arXiv:2205.03811 (2022).
10. Li, J. et al. Dynamic data-free knowledge distillation by easy-to-hard learning strategy. *Inf. Sci.* **642**, 119202 (2023).
11. Li, X. *et al.*d^3k : Dynastic data-free knowledge distillation. In *IEEE Transactions on Multimedia* (2023).
12. Chawla, A., Yin, H., Molchanov, P. & Alvarez, J. Data-free knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3289–3298 (2021).
13. Zhang, Y. *et al.* Data-free knowledge distillation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7852–7861 (2021).
14. Choi, Y., Choi, J., El-Khamy, M. & Lee, J. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 710–711 (2020).
15. Fang, G. et al. Up to 100x faster data-free knowledge distillation. *In Proceedings of the AAAI Conference on Artificial Intelligence* **36**, 6597–6604 (2022).
16. Kang, M. & Kang, S. Data-free knowledge distillation in neural networks for regression. *Expert Systems with Applications*. Vol. 175. 114813 (2021).
17. Fang, G. *et al.* Contrastive model inversion for data-free knowledge distillation. *arXiv preprint*[SPACE]arXiv:2105.08584 (2021).
18. Nayak, G. K., Mopuri, K. R. & Chakraborty, A. Effectiveness of arbitrary transfer sets for data-free knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1430–1438 (2021).
19. Liu, Y., Zhang, W., Wang, J. & Wang, J. Data-free knowledge transfer: A survey. *arXiv preprint*[SPACE]arXiv:2112.15278 (2021).
20. Lopes, R. G., Fenu, S. & Starner, T. Data-free knowledge distillation for deep neural networks. *arXiv preprint*[SPACE]arXiv:1710.07535 (2017).
21. Nayak, G. K., Mopuri, K. R., Shaj, V., Radhakrishnan, V. B. & Chakraborty, A. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*. 4743–4751 (PMLR, 2019).
22. Yin, H. *et al.* Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8715–8724 (2020).
23. Yu, X., Yan, L., Yang, Y., Zhou, L. & Ou, L. Conditional generative data-free knowledge distillation. *Image Vis. Comput.* **131**, 104627 (2023).
24. Chen, W. et al. Better together: Data-free multi-student coevolved distillation. *Knowledge-Based Syst.* **283**, 111146 (2024).
25. Chen, H. *et al.* Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3514–3522 (2019).
26. Hao, Z., Luo, Y., Wang, Z., Hu, H. & An, J. Cdfkd-mfs: Collaborative data-free knowledge distillation via multi-level feature sharing. *IEEE Trans. Multimed.* **24**, 4262–4274 (2022).
27. Shao, R., Zhang, W., Yin, J. & Wang, J. Data-free knowledge distillation for fine-grained visual categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1515–1525 (2023).

28. Park, S. & Kwak, N. Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In *ECAI 2020*. 1411–1418 (IOS Press, 2020).
29. Ji, M., Heo, B. & Park, S. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 7945–7952 (2021).
30. Qi, L. *et al.* Multi-scale aligned distillation for low-resolution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14443–14453 (2021).
31. Chen, P., Liu, S., Zhao, H. & Jia, J. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5008–5017 (2021).
32. Yuan, J., Phan, M. H., Liu, L. & Liu, Y. Fakd: Feature augmented knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 595–605 (2024).
33. Yu, S., Chen, J., Han, H. & Jiang, S. Data-free knowledge distillation via feature exchange and activation region constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24266–24275 (2023).
34. Li, M. & Yang, G. Data-free distillation improves efficiency and privacy in federated thorax disease analysis. In *2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology*. 131–132 (IEEE, 2023).
35. Kang, M. *et al.* One-shot federated learning on medical data using knowledge distillation with image synthesis and client model adaptation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 521–531 (Springer, 2023).
36. Li, N., Wang, N., Ou, W. & Han, W. Fedtd: Efficiently share telemedicine data with federated distillation learning. In *International Conference on Machine Learning for Cyber Security*. 501–515 (Springer, 2022).
37. Pesapane, F., Volonté, C., Codari, M. & Sardanelli, F. Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. *Insights Imaging* **9**, 745–753 (2018).
38. Coyner, A. S. et al. Synthetic medical images for robust, privacy-preserving training of artificial intelligence: application to retinopathy of prematurity diagnosis. *Ophthalmol. Sci.* **2**, 100126 (2022).
39. Altaheri, H., Muhammad, G. & Alsulaiman, M. Dynamic convolution with multilevel attention for EEG-based motor imagery decoding. *IEEE Internet Things J.* **10**, 18579–18588 (2023).
40. Sun, J., Li, C., Wang, Z. & Wang, Y. A memristive fully connect neural network and application of medical image encryption based on central diffusion algorithm. In *IEEE Transactions on Industrial Informatics* (2023).
41. Sun, J., Zhai, Y., Liu, P. & Wang, Y. Memristor-based neural network circuit of associative memory with overshadowing and emotion congruent effect. In *IEEE Transactions on Neural Networks and Learning Systems* (2024).
42. Zhang, R., Zong, Q., Dou, L. & Zhao, X. A novel hybrid deep learning scheme for four-class motor imagery classification. *J. Neural Eng.* **16**, 066004 (2019).
43. Xu, I. R. et al. Generative adversarial networks can create high quality artificial prostate cancer magnetic resonance images. *J. Pers. Med.* **13**, 547 (2023).
44. Dimitriadis, A., Trivizakis, E., Papanikolaou, N., Tsiknakis, M. & Marias, K. Enhancing cancer differentiation with synthetic MRI examinations via generative models: A systematic review. *Insights Imaging* **13**, 188 (2022).
45. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778 (2016).
46. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*[SPACE]arXiv:1409.1556 (2014).
47. Zagoruyko, S. & Komodakis, N. Wide residual networks. *arXiv preprint*[SPACE]arXiv:1605.07146 (2016).
48. Krizhevsky, A., Hinton, G. *et al.* Learning Multiple Layers of Features from Tiny Images. (2009).
49. Le, Y. & Yang, X. Tiny imagenet visual recognition challenge. *CS* **231N**(7), 3 (2015).
50. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255 (IEEE, 2009).
51. Yang, J., Shi, R. & Ni, B. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 191–195 (2021).
52. Yang, J. et al. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Sci. Data* **10**, 41 (2023).
53. Micaelli, P. & Storkey, A. J. Zero-shot knowledge transfer via adversarial belief matching. *Adv. Neural Inf. Process. Syst.* **32** (2019).
54. Luo, L., Sandler, M., Lin, Z., Zhmoginov, A. & Howard, A. Large-scale generative data-free distillation. *arXiv preprint*[SPACE]arXiv:2012.05578 (2020).

Acknowledgements

Thanks to the following projects for funding this research: the Natural Science Foundation of Guangdong Province under Grant 2023A1515011179, and the Shanghai University of Engineering Science Medical-Engineering Interdisciplinary Project (2023LXY-RUIJINO1Z).

Author contributions

P.L., J.C., and B.P. designed the algorithm and wrote the main manuscript text. H.H. conducted the literature review. P.L. assisted with data analysis. Y.W. translated the manuscript, refined the English language, and organized the data. G.R. and Q.C. reviewed the manuscript. All authors reviewed and approved the final version of the manuscript for publication.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.C. or G.R.