Article

Global and Local Knowledge Distillation Method for Few-shot Classification of Electrical Equipment

Bojun Zhou [1], Jiahao Zhao [1] , Chunkai Yan [1] , Xinsong Zhang [2] and Juping Gu [2,3,*]

[1] School of information Science and technology, Nantong University, Nantong, Jiangsu 226019, China; zhoubj@ntu.edu.cn
[2] School of electrical engineering, Nantong University, Nantong, Jiangsu 226019, China
[3] School of Electronic & Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu 215009, China
* Correspondence: gu.jp@ntu.edu.cn

Abstract: With the increasing utilization of intelligent mobile devices for online inspection of electrical equipment in smart grids, the limited computing power and storage capacity of these devices pose challenges for deploying large algorithm models and it's hard to obtain a substantial number of images of electrical equipment in public. In this paper, we propose a novel distillation method that compresses the knowledge of teacher networks into a compact few-shot classification network, employing a global and local knowledge distillation strategy. Central to our method is exploiting the global and local relationship between the features exacted by the backbone of the teacher network and student network. We compare our method with recent state-of-the-art (SOTA) methods on three public datasets and achieve superior performance. Additionally, we contribute a new dataset, namely EEI-100, which is specifically designed for classification of electrical equipment. We validate our method on this dataset and demonstrate its exceptional prediction accuracy of 94.12% when utilizing only 5-shot images.

Keywords: few-shot classification; electrical equipment images; knowledge distillation

1. Introduction

As an essential component of the power system, daily inspection of electrical equipment is imperative to ensure the secure and stable operation of the power system [1]. The conventional manual screening and analysis approach can no longer meet the escalating demand for image analysis of electrical equipment. With the advent of the smart grid, an increasing number of unmanned aerial vehicles (UAVs) are being deployed for online inspection. The application of artificial intelligence techniques for condition monitoring of electrical equipment can significantly enhance the efficiency of detection and maintenance. Image classification is a crucial prerequisite for equipment condition monitoring based on image information. For instance, to monitor the normalcy of electrical equipment such as transformers or insulators, their images must be initially distinguished.

With the advancement of deep learning in image recognition applications, various classification and recognition methods for electrical equipment images based on deep convolutional neural network (CNN) have been proposed. However, deep CNN training often relies on large-scale labeled data, which is challenging to obtain for all categories of electrical equipment due to their safety and sensitivity. Therefore, this paper adopts the few-shot learning (FSL) method for electrical equipment image classification. The proposed approach involves randomly partitioning the power image dataset into a base class set and a new class set. The model backbone is trained on the base class set, and subsequently combined with the classifier to accomplish the recognition training of the new class, utilizing only a limited number of image samples.

Furthermore, power inspection heavily relies on intelligent mobile devices, such as inspection robots and UAVs. However, due to the limited storage capacity of these devices, the classification

model's capacity must not be excessively large. Otherwise, it cannot be deployed on such mobile devices. To address these challenges, this paper presents a novel few-shot electrical image classification algorithm based on knowledge distillation.

Knowledge distillation [2,3] is an efficient model compression method that compresses the knowledge of teacher networks into very small student networks. Early knowledge distillation methods that minimize the KL (Kullback-Leibler) divergence of predicted class probability distributions between student and teacher networks rely on the output of the last layer of the model and learn a limited amount of information. Recent work has begun to study the features of the middle layer of the distillation network, focusing on the learning of local features of the image. However, there are both differences in the global appearance and local details between electrical equipment, and the algorithm model must fully mine this information in the learning process in order to comprehensively represent the electrical equipment images and achieve higher classification precision. Therefore, this paper proposes a global and local knowledge distillation method for few-shot classification of electrical equipment.

1.1. Few-shot Classification

In recent years, FSL has attracted researchers' widespread attention in the field of computer vision and machine learning, and a large number of few-shot image classification algorithms have been proposed. Depending on the learning paradigm used, these methods can be broadly divided into two categories: meta-learning-based methods and transfer-based learning methods.

Meta-learning is a promising approach that leverages episodic training to simulate the real test environment by randomly selecting several subtasks. This enables the acquisition of meta-knowledge that facilitates the rapid identification of new categories. Based on the type of meta-knowledge learned, meta-learning methods can be classified into optimization-based and metric-based methods. Optimization-based meta-learning methods employ a two-tier optimization process to learn the optimizer for quickly processing new tasks. A well-known example of such methods is Model-Agnostic Meta-Learning (MAML) [4]. MAML obtains the optimal initialization parameters of the model through meta-training, enabling the model to adapt to new tasks after a few gradient updates. In addition, the learning rate and gradient direction are also important factors for the optimizer [5,6]. However, these methods require storage and computation of higher-order derivatives, resulting in high memory and computational costs. On the other hand, metric-based methods use nonparametric classifiers as the basic learner, avoiding the aforementioned issues. The key factors of these methods are feature extraction and similarity measurement, which offer ample room for improvement. PARN [7] proposed a feature extractor which is learning an offset for each cell in the convolution kernel to extract more efficient features, building by deformable convolutional layers. CC+rot [8] improved the transfer ability of feature extractors by adopting auxiliary self-supervised tasks. Zhang et al. [9] used the pre-trained visual saliency detection model to segment the foreground and background of the image, and then extract the foreground and background features respectively. With the proven effectiveness of attention mechanisms in extracting discriminating features, several few-shot classification (FSC) tasks have adopted this method, including CAN [10], AWGIM [11], and CTM [12]. Additionally, in metric-based meta-learning methods, the measurement of similarity is also crucial. SEN [13] combines Euclidean distance and norm distance to improve the effectiveness of Euclidean distance measurement in high-dimensional spaces. FPN [14] calculated the reconstruction error between the support sample and the query sample as the similarity score. DN4 [15] and Deep EMD [16] obtain rich similarity measures directly on local features.

Recent studies have indicated that FSC of transfer-learning method can attain comparable performance to that of meta-learning method with complex episodic training. Such methods typically combine pre-trained feature extractors on all base class datasets with arbitrary traditional classifiers to make classification decisions for query samples of unknown classes. Reference [17] showed that pre-training the entire base class dataset using the cross-entropy loss function, followed by fine-tuning the pre-trained model using support samples of the visible class, can provide a powerful baseline for FSC tasks. Since then, several works have been proposed to improve the representation

performance of feature extractors. For example, Neg-Cosin [18] proposed to use the non-negative interval Cosine loss function to optimize the model, thereby increasing the distance between the training sample and its corresponding parametric prototype, which can effectively improve the generalization performance of the model. S2M2[19] used manifold mixing as an effective regularization method to improve the generalization performance of the model. Reference [20] and [21] used rotation prediction and mirror prediction as self-supervised tasks to add to the pre-training process, and experimental results show that self-supervised tasks are effective methods to improve feature representation performance.

In conclusion, many recent works have emphasized the importance of feature representation, both meta-learning-based and transfer-learning-based methods tended to employ highly complex networks to enhance feature representation. Therefore, deploying these methods to real-world applications usually occupies high computing resources (storage space, computing power, etc.) and introduces high time delays, which cannot meet the actual needs of the classification tasks of electrical equipment images. Hence, in this study, we employ the knowledge distillation-based model compression algorithm to accomplish the task of few-shot image classification to reduce the model parameters.

1.2. Knowledge Distillation Methods

Knowledge distillation is one of the most effective model compression methods, which has garnered significant research interest in both industry and academia due to its simple training strategy and effective performance. It leverages the knowledge acquired by a teacher network with a large scale to guide the training of a small-scale student network, enabling the latter to achieve comparable performance despite having fewer parameters.

Two key elements in current knowledge distillation methods can be summarized as: (1) the definition of effective knowledge types, (2) Effective transfer of knowledge from teacher networks to student networks. The classical knowledge distillation method minimizes the KL divergence of the predicted class probability distribution between the student and teacher networks. In order to make better use of the knowledge information contained in the teacher network, the follow-up work focuses more on how to better mine the feature knowledge hidden in the middle layer of the network. For example, AT [22] proposed to take the spatial attention of the hidden layer features of the teacher network as knowledge, and instructs the student network to imitate its attention feature map. Recently, the relationship between samples features has been proposed as a more effective knowledge. RKD [23] proposed a relational knowledge distillation method, which used distance and angle to measure the relationship between samples features, as a valid type of knowledge during distillation. Peng et al. [24] used the kernel function to obtain higher-order relationships between samples features as effective distillating knowledge.

Although the model compression methods based on feature relation knowledge mentioned above can effectively improve the performance of small-capacity student networks, the current work only focuses on the local relationship between individual sample features, ignoring the global relationship between samples features. Therefore, this paper proposes a method based on global and local knowledge distillation and applies it to the task of FSC of electrical equipment images.

1.3. Electrical Images Classification

Image-based equipment condition monitoring has been proven to be effective in enhancing the working life of equipment and providing early failure warning. In recent years, machine learning has made significant progress in the field of image classification for electrical equipment. Bogdann presented a machine learning method for determining the state of each switch by analyzing images of the switches in power distribution substations [25].Zhang implemented FINet based on improved YOLOv5 to inspect the insulators and their defects for ensuring the safety and stability of power system [26].To address few fault cases and deficient monitoring information in transformer diagnostic tasks, Xu provides an improved few-shot learning method based on approximation space

and belief functions [27]. Yi proposed a label distribution CNNs classifier to estimate the aging time of the conductor morphology of high-voltage transmission line [28].

It is noteworthy that the majority of the aforementioned investigations have concentrated on a restricted range of electrical apparatus. These models necessitate a substantial quantity of training data to guarantee optimal performance. Nevertheless, acquiring adequate electrical equipment images in a practical setting may prove to be challenging, and the proportion of labeled samples is minimal. To a certain extent, the classification of electrical equipment images does not truly belong to a big data problem. Rather, it belongs to FSL domains.

In this paper, we propose three main contributions:

1. We present a novel distillation approach that compresses the knowledge of teacher networks into a compact student network, enabling efficient few-shot classification. The incorporation of global and local relationship strategies during the distillation process effectively directs the student network towards achieving performance levels akin to those of the teacher network.
2. We contribute a new dataset that contains 100 classes of electrical equipment with 4000 images. The dataset contains a wide range of various electrical equipment, including power generation equipment, distribution equipment, industrial electrical equipment, and household electrical equipment.
3. We demonstrate the effectiveness of our proposed method by validating it on three public datasets and comparing it with the SOTA methods on the electrical image dataset we introduced. Our proposed method outperforms all other methods and achieves the best performance.

2. Methodology

2.1. Problem Definition

In few-shot image classification tasks, given a certain size of image dataset I, it is randomly divided into three subsets: I_{train}, I_{val} and I_{test} . I_{train} is used as the base dataset for pre-training the classification model. Assuming that the pre-training set has C_b categories, the mth image sample is denoted as x_m, and its corresponding label is y_m. I_{val} is used for validation, while I_{test} is used as the new class dataset for testing the trained model. For I_{val} and I_{test}, multiple N-way-K-shot subtasks are randomly sampled, with each task consisting of a support sample set (I_S) and a query sample set (I_Q). I_S is constructed by randomly selecting N categories from I_{val} or I_{test}, and then randomly selecting K samples from each category. The set of the nth category is denoted as $I_n = \{(I_k, y_k)\}_{k=1}^{K}$, and the kth image in the nth category is denoted as I_k. I_Q is composed of Q samples randomly selected from each residual sample category, denoted as $I_Q = \{I_q\}_{q=1}^{Q}$, where I_q denotes the qth query sample. Therefore, the problem of few-shot image classification can be described as using the model trained on the base class dataset and the support sample set to make classification decisions for query samples.

2.2. FSC Network based on Global and Local Knowledge Distillation

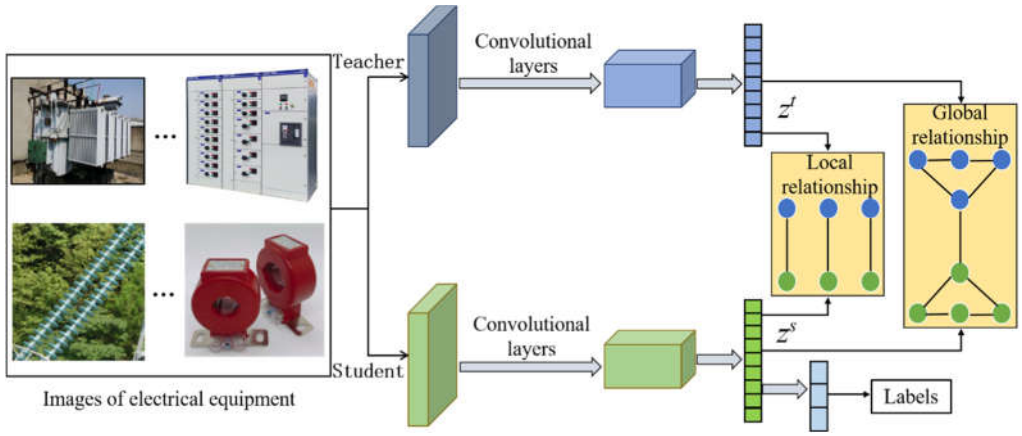


Figure 1. Network architecture of our proposed method. The input pairs that produce activations in the pre-trained teacher network produce similar activations in the student network. Global and local distillation bridges the gap between the feature representation of the student and teacher.

We propose a novel few-shot electrical image classification algorithm based on knowledge distillation. Figure 1 shows the overall architecture of our network. We first trained a high-performance teacher network through self-supervised learning, and then guided the training of the student network by the teacher network. To fully utilize the prior knowledge of the teacher network, we designed a knowledge distillation method based on global and local relationships. This method can transfer the global and local features of the images extracted by the teacher network to the student network, enabling the compact student network to learn more effective features about the images and achieve better image classification.

2.2.1. Pre-train of Teacher Network

The teacher model consists of a backbone convolutional neural network and two linear classifiers. The backbone network $f_\theta(\bullet)$ is used for feature extraction of images, one classifier $L_w(\bullet)$ is used for predicting the base class of image samples, and the other classifier $L_r(\bullet)$ is used for predicting the rotation category in self-supervised tasks. Additionally, each classifier is followed by a Softmax layer. M image samples are randomly selected from the base class dataset, and each image is rotated at 0°, 90°, 180°, and 270°, with its corresponding rotation label as $\hat{y}_m$ = [0,1,2,3].

When image x_m is fed into the teacher network, the d-dimensional feature representation $f_\theta(x_m)$ is extracted by the backbone network. The classification scores of the base class prediction classifier and the rotation prediction classifier for the features are expressed as S_b and S_r, as shown in Equation (1):

$$\begin{cases} S_b = L_w(f_\theta(x_m)) \\ S_r = L_r(f_\theta(x_m)). \end{cases} \tag{1}$$

Furthermore, the aforementioned classification scores are transformed into base class and rotation class prediction probabilities through a Softmax layer, as shown in Equation (2):

$$\begin{cases} p(y_m = c|x_m) = \dfrac{e^{S_{bc}}}{\sum_{c=1}^{C_b} e^{S_{bc}}} \\ p(\hat{y}_m = r|x_m) = \dfrac{e^{S_{rr}}}{\sum_{r=1}^{4} e^{S_{rr}}}, \end{cases} \tag{2}$$

where S_{bc} denotes the c[th] element of the score vector S_b, S_{rr} denotes the r[th] element of the score vector S_r, C_b denotes the number of base class labels, and $p(y_m = c|x_m)$ and $p(\hat{y}_m = r|x_m)$ are the probability output values of the base classifier and the rotation classifier, respectively. The cross-entropy loss function and the self-supervised loss function are calculated to obtain the training loss function, as shown in Equation (3):

$$L(\theta, w, r) = -\sum_{m=1}^{M}\sum_{c=1}^{C_b} y_{mc} \log p(y_m = c|x_m) - \sum_{m=1}^{M}\sum_{r=1}^{4} \hat{y}_{mr} \log p(\hat{y}_m = r|x_m), \quad (3)$$

where y_{mc} denotes the c[th] element of the one-hot encoded vector of y_m, and $\hat{y}_{mr}$ denotes the c[th] element of the one-hot encoded vector of $\hat{y}_m$. Based on the loss function in Equation (3), the parameters of the teacher network are optimized to complete the pre-training process.

2.2.2. Global and Local Knowledge Distillation

Firstly, a student network is constructed, which consists of a backbone neural network $B_\phi(\bullet)$ composed of a small number of convolutional layers and a linear classifier $C_H(\bullet)$. Next, a batch of M images randomly selected from the base dataset I_{train} is inputted into both the teacher network and the student network. The m[th] image is represented by feature maps $z_m^t = f_\theta(I_m)$ and $z_m^s = B_\phi(I_m)$ obtained from the backbone of the teacher network and the student network, respectively. Finally, the features are fed into the linear classifier to obtain the output value of the student network, as shown in Equation (4):

$$S_m = C_H(z_m^s). \quad (4)$$

Furthermore, the above output classification scores are transformed into classification prediction probabilities through the Softmax layer, as shown in Equation (5):

$$p_s(y_m = c|x_m) = \frac{e^{S_{mc}}}{\sum_{c=1}^{C_b} e^{S_{mc}}}, \quad (5)$$

where S_{mc} denotes the c[th] element of the score vector S_m.

The equation for calculating the cross-entropy loss function between the output values of a student network and the true labels is shown in Equation (6):

$$l_1(\phi, H) = -\sum_{m=1}^{M}\sum_{c=1}^{C_b} y_{mc} \log p_s(y_m = c|x_m). \quad (6)$$

In order to enable students to learn the representation of global features of images by the teacher network through online learning, we adopt the maximum mean discrepancy between the feature spaces of the two networks as the global loss function, which is calculated by Equation (7):

$$l_2(\phi, H) = \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} z_m^t z_{m'}^{t\,T} + \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} z_m^s z_{m'}^{s\,T} - \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} z_m^t z_{m'}^{s\,T}. \quad (7)$$

In addition, we calculate the Euclidean distance between each sample feature in the two networks as the local loss function, and its calculation formula is shown in Equation (8):

$$l_3(\phi,H)=\frac{1}{M}\sum_{m=1}^{M}\left\|z_m^t-z_m^s\right\|^2. \tag{8}$$

In summary, the total loss function for the student network is shown in Equation (9). Based on Equation (9), the student network is trained, and the parameters in the network are updated until optimal, thereby completing the knowledge distillation process from the teacher network to the student network.

$$L(\phi,H)=l_1(\phi,H)+\alpha_1 l_2(\phi,H)+\alpha_2 l_3(\phi,H). \tag{9}$$

2.2.3. Few-shot Evaluation

After completing the knowledge distillation task in Section 2.2.2, the base classifiers in the student network are first removed. Then, the parameters of the backbone neural network $B_\phi(\bullet)$ are fixed, and features are extracted from both the support and query samples. Finally, based on N-way-K-shot method, the query samples are classified using Equation (10), where the features of the k^{th} support sample and the q^{th} query sample are denoted as $B_\phi(I_k)$ and $B_\phi(I_q)$, respectively, and $g_\varphi\{\bullet\}$ is a classifier with parameters φ. Any traditional classifier can be used to complete the classification prediction task.

$$\hat{y}_q=g_\varphi\{B_\phi(I_q)\,|\,\sum_{k=1}^{NK}B_\phi(I_k)\}. \tag{10}$$

3. Experiments

Firstly, we invested a substantial amount of time in constructing a dataset, namely EEI-100 (electrical equipment image of 100 classes). Next, to assess the effectiveness of our proposed method, we performed ablation experiments on three public datasets and compared it with other few-shot image classification methods. Finally, we evaluated our method against SOTA approaches on EEI-100 dataset, showcasing the superior performance of our approach.

3.1. EEI-100 Dataset

EEI-100 contains 100 classes of electrical equipment with 4000 images. The majority of the images were obtained through on-site collection, with a small number of images sourced from online platforms. To the best of our knowledge, this is one of the first datasets specifically designed for classification of electrical equipment. This dataset is an extension of our previous EEI-40 [29]. It includes substation equipment, distribution station equipment and common electrical equipment, ranging from large-scale equipment such as heavy-duty transformers to small-scale equipment such as circuit breakers. A few images from the proposed dataset illustrated in Figure 2. More images illustrated in Appendix A.

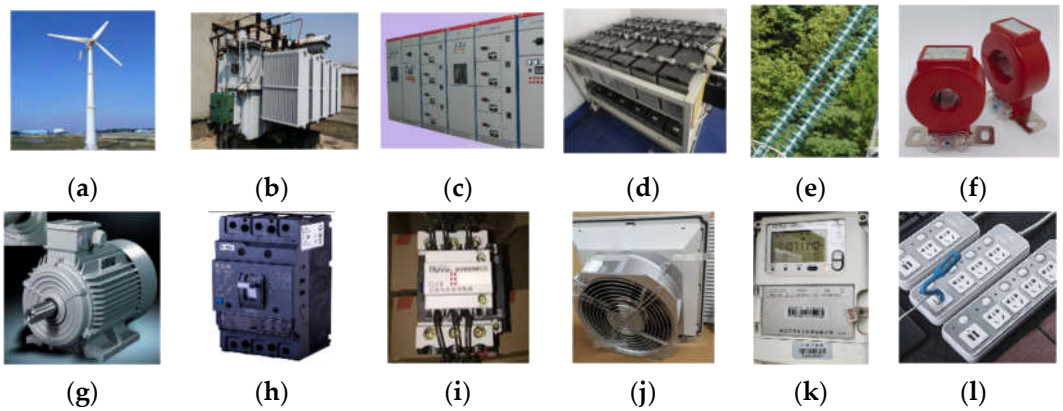


Figure 2. Some images of EEI-100 dataset. They represent different electrical equipment. (**a**)Wind power tower; (**b**)Heavy-duty transformer; (**c**)Heavy-duty distribution cabinet; (**d**)Energy storage battery pack; (**e**)Electrical insulator; (**f**) Split-core current transformer; (**g**) Three-phase moto; (**h**) Heavy-duty circuit breaker; (**i**) Contactor; (**j**) Cooling fan; (**k**) Electric energy meter; (**l**)Dragline board.

3.2. Experiments on Public Datasets

We evaluate our knowledge distillation method on three widely used public datasets, namely MiniImageNet, CIFAR-FS, and CUB. These datasets are commonly adopted for comparing distillation methods. We report the classification accuracy results of our method compared with well-known methods.

3.2.1. Experiment Setup

The experiments are conducted on a workstation equipped with NVIDIA 3090Ti GPU and implemented using Pytorch software. To ensure a fair comparison with current few-shot image classification methods, a commonly used 4-layer CNN and ResNet12 are adopted as the student network and teacher network, respectively. During the training phase, we use the SGD optimizer to optimize our models in all experiments, where the momentum is set to 0.9 and weight decay is set to 5×10-4. We train for 100 epochs, with an initial learning rate of 0.025, which is reduced by half after 60 epochs. In the testing phase, we conduct 5-way-1-shot and 5-way-5-shot tests. Specifically, we randomly perform 2000 classification subtasks on the testing dataset. In each subtask, 15 images are randomly selected from each class as query images for testing. The evaluation criterion for the algorithm's classification performance is the average accuracy of all subtasks, and the standard deviation of the accuracy under a 95% confidence interval should also be provided.

Please note that we also conducted a similar experiment on the 1080Ti GPU and achieved comparable performance. This significantly alleviates the economic burden associated with model training. Leveraging a 4-layer CNN architecture, the student model occupies a mere 2MB in size, which is approximately 50 times smaller than the teacher model. This lightweight model can be deployed on diverse edge processors, substantially lowering the hardware requirements for its implementation.

3.2.2. Parametric Analysis Experiment

It can be seen from Equation (9) that α_1 and α_2 are important hyperparameters in the process of distilling the student network.

Initially, we conducted experiments in which we temporarily ignored the influence of global knowledge by setting α_1 to 0. Through this analysis, we observed that α_2 near 1 yielded the best model performance.

Building upon this observation, we proceeded to fix α_2 at 1, and the value of parameter α_1 was varied with a step size of 0.1 within the range of [0,1]. The test accuracy of the student network under different values of α_1 is shown in Fig. 3(a) and (b). The results indicate that the model performance is

optimal when the value of α_1 is 0.5. Therefore, the value of α_1 was set to 0.5, and then α_2 was varied with a step size of 0.01 within the range of [0,0.1]. The test accuracy of the student network under different values of α_2 is shown in Fig. 3(c) and (d). The results reveal that the optimal value for α_2 is 0.1.

Additionally, after completing the search for α_1 and α_2 within their respective ranges, we extended our exploration beyond the boundaries of [0,1] and [0,0.1]. However, we found that no values outside of these ranges yielded superior results.

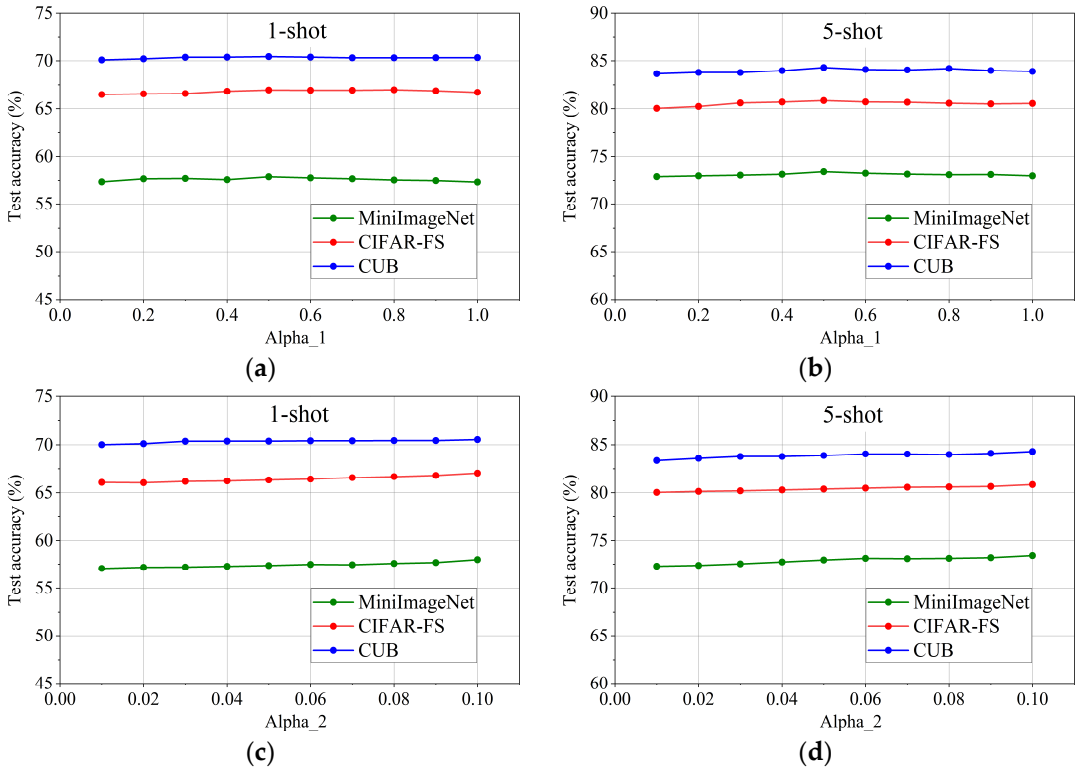


Figure 3. Test experiments under different values of α_1 and α_2 on three public datasets. (**a**) 1-shot test accuracy under different values of α_1; (**b**) 5-shot test accuracy under different values of α_1; (**c**) 1-shot test accuracy under different values of α_2; (**d**) 5-shot test accuracy under different values of α_2;.

3.2.3. Ablation Studies

The innovation of this work lies in proposing a knowledge distillation algorithm for global and local relationships. In order to verify the effectiveness of the proposed method, detailed ablation experiments are conducted on three public datasets. The knowledge distillation algorithms using only global and local relationships are denoted as Global and Local, respectively, and their fusion is denoted as Global-Local. The classification accuracies of these methods on 5-way-1-shot and 5-way-5-shot tasks are shown in Table 1.

Table 1. Results (%) of ablation experiment

Method	Backbone	MiniImageNet		CIFAR-FS		CUB	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Global	Conv4	57.32±0.84	72.90±0.64	66.40±0.93	80.44±0.67	70.20±0.93	83.88±0.57
Local	Conv4	57.65±0.83	73.06±0.64	66.63±0.93	80.64±0.67	70.12±0.93	83.66±0.57
Global-Local	Conv4	57.86±0.83	73.38±0.62	67.04±0.91	80.84±0.68	70.44±0.92	84.19±0.56

The results in the table indicate that for both 5-way-1-shot and 5-way-5-shot tasks on all datasets, the classification accuracy of Global-Local is consistently higher than that of Global and Local. The

experiments demonstrate that global and local relationships are complementary, and their fusion can extract richer image features. Therefore, the knowledge distillation algorithm based on global and local relationships can further improve the performance of knowledge distillation.

3.2.4. Classification Experiment Compared with Existing Methods

This paper compares our method with the SOTA methods in recent years, which are mainly divided into two categories: meta-learning-based methods and transfer learning-based methods. The comparison results with these methods are shown in Table 2.

Table 2. Comparison results (%) of the experiment on three public datasets

Method	Backbone	MiniImageNet		CIFAR-FS		CUB	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Meta-learning							
Relational	Conv4	50.44±0.82	65.32±0.70	55.00±1.00	69.30±0.80	62.45± 0.98	76.11± 0.69
MetaOpt SVM	Conv4	52.87±0.57	68.76±0.48	-*	-	-	-
PN+rot	Conv4	53.63±0.43	71.70±0.36	-	-	-	-
CovaMNet	Conv4	51.19±0.76	67.65± 0.63	-	-	52.42±0.76	63.76±0.64
DN4	Conv4	51.24±0.74	71.02±0.64	-	-	46.84±0.81	74.92±0.64
MeTAL	Conv4	52.63±0.37	70.52±0.29	-	-		
HGNN	Conv4	55.63±0.20	72.48±0.16	-	-	69.02±0.22	83.20±0.15
DSFN	Conv4	50.21±0.64	72.20±0.51	-	-	-	-
PSST	Conv4	-	-	64.37±0.33	80.42± 0.32	-	-
Transfer-learning							
Baseline++	Conv4	48.24±0.75	66.43±0.63	-	-	60.53±0.83	79.34±0.61
Neg-Cosine	Conv4	52.84±0.76	70.41±0.66	-	-	-	-
SKD	Conv4	48.14	66.36	-	-	-	-
CGCS	Conv4	55.53±0.20	72.12±0.16	-	-	-	-
Our method	Conv4	57.86±0.83	73.38±0.62	67.04±0.91	80.84±0.68	70.44±0.92	84.19±0.56

* - indicates that the method described in the literature was not evaluated on certain datasets.

According to the results in Table 2, the following observations can be made:

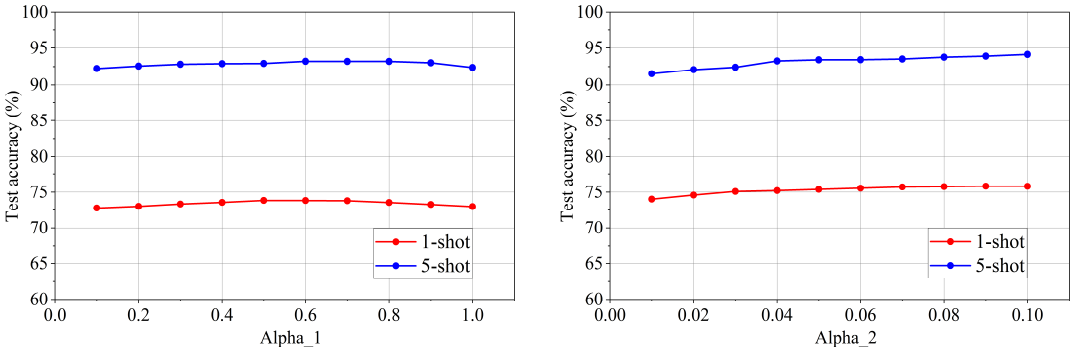
1. On the MiniImageNet dataset, our proposed method achieves the best classification performance. Compared with the best performing method in the meta-learning-based category, HGNN, our method outperforms it by 2.23% and 0.9% on 1-shot and 5-shot classification tasks, respectively. In the transfer learning-based category, compared with the best performing method, CGCS, our method outperforms it by 2.33% and 1.26% on 1-shot and 5-shot classification tasks, respectively.
2. On the CIFAR-FS dataset, our proposed method also achieves the top performance. Our method outperforms the best performing method, PSST, by 2.67% and 0.42% on 1-shot and 5-shot classification tasks, respectively.
3. On the CUB-200-2011 dataset, our proposed method achieves the highest classification performance. Our method outperforms the best performing method, HGNN, by 1.42% and 0.99% on 1-shot and 5-shot classification tasks, respectively.

3.3. Experiments on EEI-100 Dataset

Furthermore, we compare the performance of our proposed method with the SOTA methods on EEI-100 dataset. The experimental process employs the same parameter selection strategy as before.

3.3.1. Parametric Analysis Experiment

By following the approach outlined in Section 3.2.2, the values of parameters α_1 and α_2 are determined to optimize the performance of the model on EEI-100. Experimental results show that α_1 has the optimal value of 0.6 within the range of [0.1,1], as illustrated in Figure 4(a). Similarly, α_2 has the optimal value of 0.1 within the range of [0.01,0.1], as illustrated in Figure 4(b).



(**a**) (**b**)

Figure 4. Test experiments under different values of α_1 and α_2 on EEI-100 dataset. (**a**) 1-shot and 5-shot test accuracy under different values of α_1; (**b**) 1-shot and 5-shot test accuracy under different values of α_2.

3.3.2. Comparison Experiment with Existing Methods

To demonstrate the superiority of our proposed method in the classification of electrical equipment images, this section presents a comparative experiment with three existing methods, namely CGCS, Neg-Cosine, and HGNN, on the EEI-100 dataset. These three methods have recently achieved good performance on public datasets. The classification accuracy of the test set is presented in Table 3. Specifically, our method achieves the highest classification accuracy (up to 94.12%) compared with the other methods.

Table 3. Comparison results (%) of the experiment on EEI-100 dataset

Method	1-shot	5-shot
CGCS	72.85±0.68	89.68±0.27
Neg-Cosine	74.57±0.63	90.54±0.25
HGNN	75.61±0.62	93.54±0.24
Our method	75.80±0.67	94.12±0.20

4. Conclusion and Future Work

In conclusion, this paper presents a novel few-shot electrical image classification algorithm based on knowledge distillation. By leveraging the few-shot learning method and employing global and local knowledge distillation, our algorithm achieves high classification accuracy with only a limited number of image samples. The results obtained on the newly introduced EEI-100 dataset demonstrate that our method achieves a remarkable prediction accuracy of 94.12% using just 5-shot images.

The lightweight and high-performance nature of our model enables its practical application in the online inspection of electrical equipment in smart grids, effectively enhancing the efficiency of detection and maintenance in the power system. Furthermore, the training and deployment of our model do not impose significant hardware requirements, making it accessible to a wide range of researchers.

As future work, we plan to explore a pre-training method to separate the foreground and background, as different backgrounds may negatively affect distillation. Additionally, we plan to use a multi-stage fusion of global and local features during the distillation process. This can provide a better understanding of the underlying structure of the complex model and the relationship between different stages of the models.

Author Contributions: Conceptualization, B.Z. and J.G.; methodology, B.Z.; software, C.Y.; validation, B.Z. and J.Z.; formal analysis, B.Z.; investigation, X.Z.; resources, J.Z.; data curation, X.Z.; writing—original draft preparation, B.Z.; writing—review and editing, J.G.

Funding: This research was supported in part by the National Natural Science Foundation of China under Grant No. U2066203 and No. 61973178, and in part by the Key Research & Development Program of Jiangsu Province under Grant No. BE2021063.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset proposed in this paper can be obtained from the author with a reasonable request.

Conflicts of Interest: The authors declare no conflict of interest in preparing this article.

Appendix A

In this appendix, more images of the EEI-100 dataset proposed by us are presented. However, we regret to inform that due to the fact that some image data were collected in specific scenarios, the device information in the pictures cannot be disclosed. Therefore, we are unable to fully release the entire dataset here.

Abbreviations

SOTA	State-of-the-art
UAV	Unmanned aerial vehicle
CNN	Convolutional neural network
FSL	Few-shot learning
FSC	Few-shot classification

References

1. Peng, J.; Sun, L.; Wang, K.; Song, L. ED-YOLO power inspection UAV obstacle avoidance target detection algorithm based on model compression. *Chinese Journal of Scientific Instrument* **2021**, *10*, 161-170.
2. Geoffrey, H.; Oriol V.; Jeff, D. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531*, Mar 2015.
3. Adriana, R.; Nicolas, B.; Samira, E.K.; Antoine, C.; Carlo, G.; Yoshua, B. FitNets: Hints for Thin Deep Nets. *arXiv:1412.6550*, Dec 2014.
4. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep network. Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 7 Aug 2017.
5. Li, Z.; Zhou, F.; Chen, F.; Li, H. Meta-SGD: Learning to Learn Quickly for Few-Shot Learning. *arXiv:1707.09835*, Jul 2017.
6. Ravi, S.; Larochelle, H. Optimization as a model for few-shot learning. 5th International Conference on Learning Representations (ICLR), Toulon, France, 24 Apr 2017.
7. Wu, Z.; Li, Y.; Guo, L; Jia, K. PARN: Position-Aware Relation Networks for few-shot learning. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 Oct 2019.
8. Gidaris, S.; Bursuc, A.; Komodakis, N.; Perez, P.; Cord, M.; Ecole, L. Boosting few-shot visual learning with self-supervision. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 Oct 2019.
9. Zhang, H.; Zhang, J.; Koniusz, P. Few-shot learning via saliency-guided hallucination of samples. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), California, USA, 16 June 2019.
10. Hou, R.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Cross attention network for few-shot classification. Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, 8 Dec 2019.

11. Guo, Y.; Cheung, N. Attentive weights generation for few shot learning via information maximization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 13 June 2020.
12. Li, H.; Eigen, D.; Dodge, S.; Zeiler, M.; Wang, X. Finding task-relevant features for few-shot learning by category traversal. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), California, USA, 16 June 2019.
13. Nguyen, V.N.; Løkse, S.; Wickstrøm, K.; Kampffmeyer, M.; Roverso, D.; Jenssen, R. SEN: a novel feature normalization dissimilarity measure for prototypical few-Shot learning networks. Proceedings of the 16th European conference on computer vision (ECCV), Glasgow, Scotland, 23 Aug 2020.
14. Wertheime, D.; Tang, L.; Hariharan, B. Few-shot classification with feature map reconstruction networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), online, 19 June 2021.
15. Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; Luo, J. Revisiting local descriptor based image-to-class measure for few-shot learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), California, USA, 16 June 2019.
16. Zhang, C.; Cai, Y.; Lin, G.; Shen, C. DeepEMD: few-shot image classification with differentiable Earth Mover's distance and structured classifiers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 13 June 2020.
17. Chen, Y.; Liu, Y.; Kira, Zsolt.; Wang, Y.F.; Huang, J. A closer look at few-shot classification. Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, USA, 6 May 2019.
18. Liu, B.; Cao, Y.; Lin, Y.; Zhang, Z.; Long, M.; Hu, H. Negative margin matters: understanding margin in few-shot classification. Proceedings of the 16th European conference on computer vision (ECCV), Glasgow, Scotland, 23 Aug 2020.
19. Mangla, P.; Singh, M.; Sinha, A.; Kumari, N.; Balasubramanian, V.; Krishnamurthy, B. Charting the right manifold: Manifold mixup for few-shot learning. The IEEE Winter Conference on Applications of Computer Vision (WACV), Hawaii, USA, 7 Janu 2020.
20. Su, J.; Maji, S.; Hariharan, B. When does self-supervision improve few-shot learning. Proceedings of the 16th European conference on computer vision (ECCV), Glasgow, Scotland, 23 Aug 2020.
21. Shao, S.; Xing, L.; Wang, Y.; Xu, R.; Zhao, C.; Wang, Y.J.; Liu, B. MHFC: multi-head feature collaboration for few-shot learning. Proceedings of the 29th ACM International Conference on Multimedia (MM), Online, 20 Otc 2021.
22. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. 5th International Conference on Learning Representations (ICLR), Toulon, France, 24 Apr 2017.
23. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), California, USA, 16 June 2019.
24. Peng, B.; Jin, X.; Liu, J; Zhou, S.; Wu, Y.; Liu, Y.; Li, D.; Zhang, Z. Correlation congruence for knowledge distillation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 Oct 2019.
25. Bogdan, T.N.; Bruno, M.; Rafael, W.; Victor, B.G.; Vanderlei, Z.; Lourival, L. A Computer Vision System for Monitoring Disconnect Switches in Distribution Substations. *IEEE Transactions on Power Delivery* **2022**, *37*, 833-841.
26. Zhang, Z.D.; Zhang, B.; Lan, Z.C.; Lu, H.C.; Li, D.Y.; Pei, L.; Yu, W.X. FINet: An Insulator Dataset and Detection Benchmark Based on Synthetic Fog and Improved YOLOv5. *IEEE Transactions on Instrumentation and Measurement* **2022**, *71*, 1-8.
27. Xu, Y.; Li, Y.; Wang, Y.; Zhong, D.; Zhang, G. Improved few-shot learning method for transformer fault diagnosis based on approximation space and belief functions. *Expert Systems with Applications* **2021**, *167*, 114105.
28. Yi, Y.; Chen, Z.; Wang, L. Intelligent Aging Diagnosis of Conductor in Smart Grid Using Label-Distribution Deep Convolutional Neural Networks. *IEEE Transactions on Instrumentation and Measurement* **2022**, *71*, 1-8.
29. Zhou, B.; Zhang, X.; Zhao, J; Zhao, F.; Yan, C.; Xu, Y.; Gu, J. Few-shot electric equipment classification via mutual learning of transfer-learning model. IEEE 5th International Electrical and Energy Conference (CIEEC), Nanjing, China, 27 May 2022.

Appendix A Few images of EEI-100 dataset


1 2 3 4 5 6


7 8 9 10 11 12


13 14 15 16 17 18


19 20 21 22 23 24


25 26 27 28 29 30


31 32 33 34 35 36


37 38 39 40 41 42


43 44 45 46 47 48