ISSN 2959-6157

Knowledge Distillation for Urban Tree Image Classification Based on the Diffusion Model

Ruijing Shi

Internet of Things Engineering, Nankai University, Tianjin, China
2113322@mail.nankai.edu.cn

Abstract:

As urban areas continue to emerge as significant contributors to global carbon emissions, the classification of urban street trees has gained increasing importance in carbon sequestration research. By precisely identifying and categorizing different tree species, the carbon absorption capabilities of urban vegetation can be evaluated more accurately. This understanding is essential for addressing the growing environmental challenges posed by carbon emissions, as urban trees play a crucial role in absorbing carbon dioxide and mitigating the effects of climate change. In this study, a Feature Distillation Based on Diffusion Model (FDBD) framework was proposed, utilizing state-of-the-art image classification technology to enhance the accuracy of urban tree species identification. The framework utilized knowledge distillation, a process where a smaller, more efficient "student" model is trained to mimic the performance of a larger, more complex "teacher" model, significantly reducing computational demands while maintaining high accuracy. The model's effectiveness has been validated through combination experiments with a selected backbone model, achieving promising results. This approach not only enhances the understanding of urban trees' carbon sequestration potential but also provides crucial insights for policymakers. By facilitating precise tree classification, it empowers urban planners to implement more informed and targeted strategies for carbon reduction, ultimately promoting more sustainable urban environments.

Keywords: Knowledge distillation; diffusion model; urban tree classification; carbon sequestration.

1. Introduction

As of 2024, addressing carbon emissions has become a critical priority for governments and organizations worldwide. Urban areas, which are significant sources of carbon emissions, are increasingly focusing on sustainable practices to mitigate their environmental impact. Among these, urban greening stands out as an effective strategy due to its ability to act as a carbon sink. Accurately estimating the carbon sequestration potential of urban vegetation is essential for informed city planning and climate action.

Classification of city street tree holds significant importance in carbon emissions research. By accurately identifying and classifying different tree species in urban areas, researchers can gain a deeper understanding of these species' abilities to absorb and sequester carbon. Given the substantial differences in carbon absorption efficiency among various tree species, utilizing this dataset allows for a more precise assessment of the role urban greenery plays in carbon offsetting. This in-depth analysis not only reveals the impact of urban vegetation on carbon emissions but also provides scientific support for developing more effective carbon reduction strategies.

Traditionally, environmental monitoring relied on manual data collection and basic image processing techniques to identify and classify environmental features e.g. vegetation, water bodies, and pollution levels. While these methods provided foundational insights, they were often limited by their inability to process large-scale data and adapt to varying environmental conditions. Artificial intelligence technology can assist in dealing with these problems. The potential of artificial intelligence technology in promoting environmental sustainability has been fully demonstrated in recent years [1], showing remarkable performance across various fields [2]. In the application of image classification models, for instance, Zhang et al. used Convolutional Neural Networks (CNNs) to automatically classify and monitor coral reef ecosystems through underwater imagery highlights the potential of this model in expanding environmental monitoring efforts [3]. Additionally, Tabak applied CNNs to automatically detect wildlife in camera trap images [4], has significantly contributed to wildlife monitoring and improved species conservation efficiency. Furthermore, Zhang utilized deep learning technology to classify land cover types for monitoring desertification processes, is of critical importance for the ecological protection of arid regions [5].

The primary purpose of the urban street tree classification dataset is plant identification, but it holds significant potential in carbon emissions research and related applications. For example, researchers can estimate the carbon sequestration potential of different tree species through accurate classification and identification. This is crucial because different species vary in their ability to absorb and store carbon dioxide, and precise identification helps assess the impact of urban greenery on offsetting carbon emissions. Additionally, long-term monitoring of urban tree species and changes in their health can provide valuable insights into the impact of urbanization on carbon emissions and aid in predicting future emission trends under climate change. Policymakers can also use this data to support carbon reduction policies, such as protecting tree species with high carbon sequestration capacity and encouraging their planting through urban planning [6].

As the demand for higher accuracy in task processing increases, the depth and complexity of neural networks also rises, leading to greater challenges in computation and deployment. Knowledge distillation is an effective technique that addresses this issue by transferring knowledge from a complex, large-scale network (the teacher) to a simpler, more efficient network (the student). This process enables the student network to retain the high performance of the teacher network while maintaining a lower computational complexity. This study aims to employ a state-of-the-art knowledge distillation-based network, which has demonstrated superior performance in image classification tasks, to train a model capable of classifying images in an urban tree dataset. By leveraging the strengths of the teacher network, the resulting student model is expected to achieve high accuracy in tree classification while remaining computationally efficient.

2. Method

2.1 Preliminaries

2.1.1 Knowledge Distillation (KD)

Knowledge distillation [7] is a model compression technique in machine learning where a smaller, more efficient model (the student) is trained to replicate the behavior of a larger, more complex model (the teacher). The idea is that a teacher model trained on a large dataset can capture complex patterns and generalizations. The student model, while simpler, can learn from the teacher's output, including its predictions and probability distributions across categories, not just hard labels. If defining the student's output as $F^{(s)}$, the teacher's output as $F^{(t)}$, and a distance function as d, then the loss function for knowledge distillation can be obtained as follows:

$$\mathcal{L}_{kd} = d\left(F^{(s)}, F^{(t)}\right) \tag{1}$$

2.1.2 Diffusion Model

Diffusion models are a class of generative models used in machine learning and artificial intelligence. They work by gradually transforming a simple distribution (such as Gaussian noise) into a more complex distribution that represents the target data (such as images) through a series of iterative steps. During training, the model is taught to invert the noise added to the data in small increments, which allows it to generate new samples starting from random noise and applying the learned reverse steps. Just like in the classic stable model [8], the loss formula for the diffusion model is as follows:

$$\mathcal{L}_{\text{diff}} = \| ?_t - \Phi_\theta\left(z_t, t\right) \|_2^2 \tag{2}$$

Where $?_t \in \mathcal{N}(0, I)$ represents the actual noise added to the image, and $\Phi_\theta\left(z_t, t\right)$ is the noise distribution predicted by the model during training, so as to reduce their L2 loss during training.

2.2 Feature Distillation Based on Diffusion Model (FDBD)

The main significance of knowledge distillation technology is that it can significantly reduce model parameters while retaining output accuracy. The key to its technology is to make the content learned by students as close as possible to the content of the teacher. Specifically, Student Features (SF) usually contain more noise compared to Teacher Features (TF), which makes the overall results worse. Therefore, removing noise from SF to reduce the difference between SF and TF can effectively improve the distillation efficiency. At the same time, the impact of noise on KD has attracted sufficient attention from scholars. Previous studies [9] directly remove the background information to remove the influence of noise in the classification process to purify the information learned by students. However, background information also contains useful visual cues that have a strong relationship to the object [10]. Directly deleting background information will seriously limit student's feature representation capabilities. Considering the above two points, this study used the diffusion model to KD's model, which is able to remove irrelevant noise from SF while retaining discriminative information as much as possible. Using distilled denoised SF to enable students to learn more discriminative knowledge. To achieve this goal, this study used softmax to normalize different channels of the feature map and used KL divergence to evaluate the differences between teachers and students.

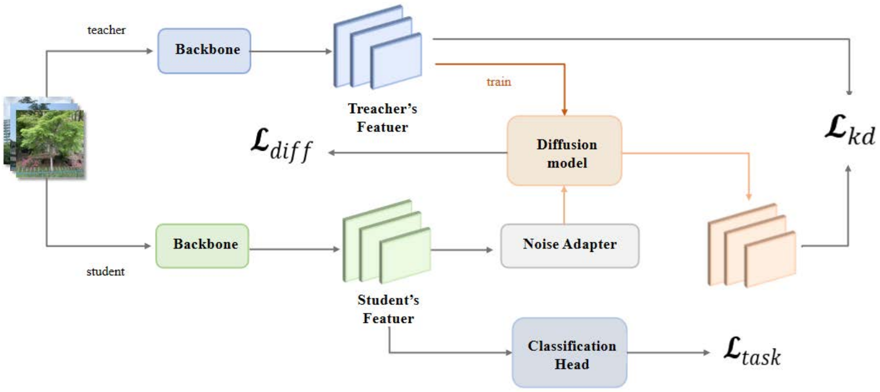

Fig. 1 The architecture of the proposed model (Photo/Picture credit: Original).

As shown in Fig. 1, this work refers to the image classification knowledge distillation framework based on the state-of-the-art (SOTA) DiffKD model [11]. Based on this framework, this work designs a training architecture that contains three key loss functions to drive the training process: diffusion loss, knowledge distillation loss, and task-specific classification loss.

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{ae}} + \lambda_3 \mathcal{L}_{\text{diffkd}} \tag{3}$$

Phase 1: Training the Diffusion Model. The first stage of the process focuses on training the diffusion model. The diffusion model is trained using the feature representations extracted by the teacher model. The teacher's feature, which is cleaner and less noisy due to the teacher model's larger capacity and higher capability, serves as the input to the diffusion model. This process is driven by the $\mathcal{L}_{\text{diff}}$, The calculation of this loss is described in Eq. (1), which aims to minimize the discrepancy between the noisy and clean features. In this stage, the diffusion model learns to denoise the teacher's features. Once this model is trained, it is frozen in the subsequent stages to ensure that its parameters remain unchanged.

Phase 2: Knowledge Distillation and Student Model Training. In the second stage, the student model's features (which are noisier due to the student's smaller capacity) are passed through the frozen diffusion model. The goal here is to denoise the student's features so that they more closely resemble the teacher's features. The Noise Adapter module is a crucial component that bridges the gap between the noisy student feature and the diffusion model, ensuring that the input noise level matches what the diffusion model expects.

Once the student's features are denoised, they are used for knowledge distillation. The denoised student features are compared with the teacher's features, and the knowledge distillation loss ($\mathcal{L}_{diffKD}$) is calculated. Based on the general distillation loss shown in Eq. (2), the distance calculation formula used in it here is KL, and $\overset{?}{F}^{(stu)}$ represents the denoised version of the student model features after the diffusion model obtained in the first stage, and $F^{(tea)}$ represents the features of the same image processed by the teacher model. This loss measures the discrepancy between the two sets of features and encourages the student to learn from the teacher by minimizing this gap.

$$\mathcal{L}_{\text{diffkd}} = d_{KL}\left(\overset{?}{F}^{(stu)}, F^{(tea)}\right) \tag{4}$$

In parallel, the student model also performs the main task of image classification and calculates the task-specific loss ($\mathcal{L}_{task}$) based on the prediction accuracy of the student model. This loss ensures that the student not only learns the teacher's knowledge, but also improves its performance on the target task.

3. Experiments

3.1 Dataset Preparation

In this work, the Urban Street Tree Classification dataset provided by Yang et al. from Zhejiang Agriculture and Forestry University in 2022 was utilized [12]. This dataset shown in Fig. 2 comprises 4,804 high-quality RGB images spanning 23 categories, specifically intended for the classification of urban street trees.

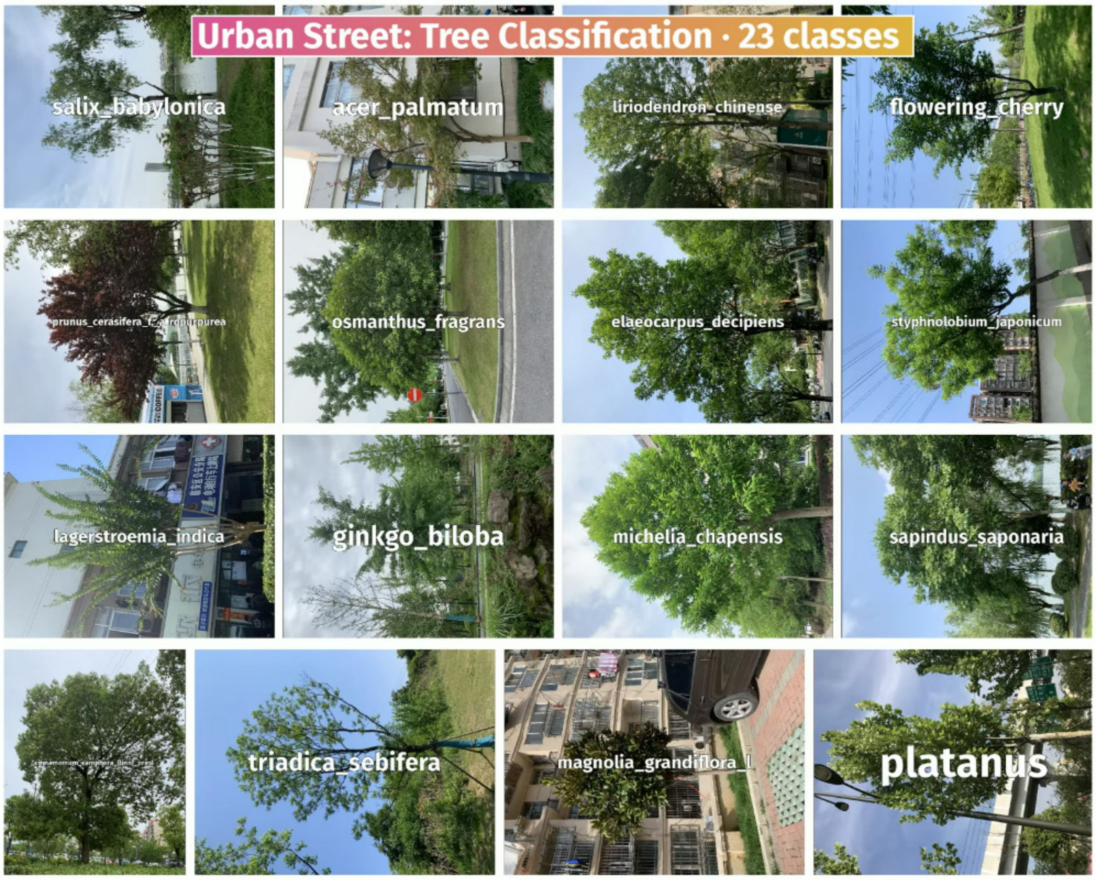

Fig. 2 The sample images of the collected dataset [12].

3.2 Implementation Details

In this experiment, ResNet18, ResNet34, and ResNet50 were primarily employed as backbones. ResNet18 and ResNet34 were used as student models, while ResNet34 and ResNet50 were paired as teacher models, forming two experimental combinations (Table 2). In addition to validating the effectiveness of the FDBD method applied in this work, this paper also evaluated the performance of the student models when trained independently without FDBD. Furthermore, this study tested another knowledge distillation model and compared its performance with the FDBD method used in this study.

The experiment is configured with a batch size of 32. Learning rate decay is applied every 30 epochs, with a decay rate of 0.1. The model is trained for 100 epochs in total shown in Table 1. The dropout rate is set to 0.0, and logging occurs every 50 iterations. The initial learning rate is set to 0.1, with a minimum learning rate of 1.0e-05. The optimizer used is stochastic gradient descent (SGD) with a momentum of 0.9. Weight decay is applied with a rate of 1.0e-4. The optimizer does not use beta values, and epsilon is set to 1.0e-08. The learning rate schedule follows a step-wise decay pattern, with no warm-up period and an initial warm-up learning rate of 0.2. The seed for random initialization is set to 42, and 8 workers are utilized for data loading.

Additionally, when using the knowledge distillation strategy, both the original loss and KD loss weights are set to 1.0. An autoencoder (AE) is employed with 1024 channels, and the temperature for the softmax function is set to 1.

Table 1. Main parameter configuration

Epochs	Total BS	Initial LR	Optimizer	WD	LR scheduler	Data augmentation
100	256	0.1	SGD	1×10^{-4}	×0.1 every 30 epochs	crop + flip

3.3 The Performance of Models

In the results section of this paper, the performance of two student-teacher model combinations is evaluated using several metrics, including Top-1 and Top-5 accuracy, DIST [13], and DiffKD. These metrics are used to measure the performance of the Feature Distillation Based on Diffusion (FDBD) framework for the urban street tree classification task.

Table 2. Evaluation results

Student (teacher)		Tea.	Stu.	DIST	DiffKD
R18(R34)	TOP1	85.983%	85.855%	84.232%	85.100%
	TOP5	96.795%	95.680%	96.718%	96.745%
R34(R50)	TOP1	88.498%	86.983%	85.349%	86.290%
	TOP5	97.851%	97.595%	97.176%	97.394%

ResNet18 (Student) with ResNet34 (Teacher) Top-1 Accuracy: The teacher model achieves 85.983%, while the student model achieves 85.855%. With the DIST and DiffKD methods applied, the accuracy slightly decreases to 84.232% and 85.100%, respectively. Top-5 Accuracy: The teacher model reaches 96.795%, and the student model performs closely at 95.680%. After applying DIST and DiffKD, the accuracy is slightly improved to 96.718% and 96.745%.

ResNet34 (Student) with ResNet50 (Teacher) Top-1 Accuracy: The teacher achieves 88.498%, while the student reaches 86.983%. With DIST and DiffKD, the accuracy is slightly lower, at 85.349% and 86.290%. Top-5 Accuracy: The teacher achieves 97.851%, while the student model scores 97.595%. After applying DIST and DiffKD, the performance remains strong at 97.176% and 97.394%.

These results demonstrated that although there is a minor drop in the student models' performance compared to their respective teachers, the performance remains competitive. The DiffKD method slightly outperforms the DIST method in most cases, demonstrating the effectiveness of the FDBD framework in reducing computational complexity while maintaining classification accuracy. The use of knowledge distillation effectively transfers the knowledge from a more complex teacher model to a simpler student model without a significant loss in performance

4. Conclusion

In this work, the application of knowledge distillation was explored to urban street tree classification, aiming to enhance the efficiency and accuracy of image classification models while maintaining lower computational complexity. By leveraging the strengths of diffusion models, this paper introduced a framework that distills knowledge from a larger, more complex teacher network to a smaller student network. The proposed FDBD successfully reduced noise and preserved key discriminative features in the student model. Numerous experiments demonstrated the effectiveness of this approach, with the student model achieving performance levels comparable to or surpassing the baseline, particularly in top-1 and top-5 accuracy metrics. The future work will focus on applying this framework to other environmental monitoring tasks, as well as further optimizing the model to enhance its generalization capabilities across different datasets and environmental conditions. This research contributes valuable insights into carbon sequestration research, offering a more computationally efficient method for urban tree classification.

References

[1] Nishant R, Kennedy M, Corbett J. Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. International Journal of Information Management, 2020, 53: 102104.Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[2] Collins A C. Harnessing Innovations in AI and Robotics for Environmental Conservation: A Comprehensive Overview. 2024.

[3] Gonzalez-Rivero M, Beijbom O, Rodriguez-Ramirez A, et al. Monitoring of coral reefs using artificial intelligence: A feasible and cost-effective approach. Remote Sensing, 2020, 12(3): 489.

[4] Villa A G, Salazar A, Vargas F. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. Ecological informatics, 2017, 41: 24-32.

[5] Feng K, Wang T, Liu S, et al. Monitoring desertification using machine-learning techniques with multiple indicators derived from MODIS images in Mu Us Sandy Land, China. Remote Sensing, 2022, 14(11): 2663.

[6] Wang X, Wang Y, Zhou C, et al. Urban forest monitoring based on multiple features at the single tree scale by UAV. Urban Forestry & Urban Greening, 2021, 58: 126958.

[7] Hinton G. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531, 2015.

[8] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.

[9] Bylander T. Learning linear threshold functions in the presence of classification noise. Proceedings of the seventh annual conference on Computational learning theory. 1994: 340-347.

[10] Guo J, Han K, Wang Y, et al. Distilling object detectors via decoupled features. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 2154-2164.

[11] Huang T, Zhang Y, Zheng M, et al. Knowledge diffusion for distillation. Advances in Neural Information Processing Systems, 2024, 36.

[12] Github, Urban Street Tree Classification, 2024, https://github.com/dataset-ninja/urban-street-tree-classification

[13] Huang T, You S, Wang F, et al. Knowledge distillation from a stronger teacher. Advances in Neural Information Processing Systems, 2022, 35: 33716-33727.