

Article

A Lightweight Mining-Area Remote Sensing Scene Classification Framework via Knowledge Distillation and Channel-Aware Non-Local Attention

Wenxi He ¹, Zi Li ², Liangjun Wang ²  and Weitao Chen ^{2,*} ¹ Wuhan Geological Survey Center, Wuhan 430205, China; hewenxi@mail.cgs.gov.cn² School of Computer Science, China University of Geosciences, Wuhan 430074, China; zili.cug@gmail.com (Z.L.); ljwang8@cug.edu.cn (L.W.)

* Correspondence: wtchen@cug.edu.cn

Abstract

Mining-area remote sensing scene classification plays an important role in mineral resource monitoring and ecological environment assessment. However, high-accuracy deep learning models usually exhibit complex architectures, large parameter sizes, and high computational costs, which limit their deployment on resource-constrained edge devices. To address this issue, this paper proposes a knowledge distillation and channel-aware non-local attention network (KDANet) for lightweight mining-area remote sensing scene classification. Specifically, a pretrained multi-layer Transformer-based high-performance network (HTransNet) is adopted as the teacher model, while ShuffleNetV2 is used as the lightweight student model. The proposed channel-aware non-local attention module (CANLAM) adaptively enhances high-level features during the distillation process, enabling the student model to better capture critical channel-wise and spatial contextual information. Furthermore, a multi-perspective distillation strategy is constructed by jointly optimizing contrastive distillation loss, relational distillation loss, and feature distillation loss, guiding the student model to approximate the teacher model from both feature representation and output distribution perspectives. Experimental results on the previously constructed CUG_MA mining scene dataset demonstrate that, with a reduction of approximately 52% in parameters and 62% in computational cost, KDANet achieves an overall accuracy (OA) of 69.48%. The lightweight student model maintains competitive classification performance while significantly improving computational efficiency and outperforming several mainstream distillation approaches.

Keywords: mining area; remote sensing scene classification; knowledge distillation; lightweight network; channel-aware attention



Academic Editor: Antonio Miguel Martínez-Graña

Received: 5 June 2026

Revised: 9 July 2026

Accepted: 13 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the continuous advancement of high-resolution remote sensing imaging technologies, mining-area remote sensing scene classification has become a fundamental technique for supporting mineral resource monitoring, illegal mining detection, environmental impact assessment, and ecological restoration tracking. By automatically identifying and classifying surface scenes in mining areas, it provides essential data support for mining management and environmental supervision and thus holds significant practical and engineering value.

However, compared with general remote sensing scenes, mining areas typically exhibit complex land-cover compositions, highly fragmented spatial structures, and significant interference from human activities. On the one hand, different categories often share similar visual and textural characteristics; on the other hand, the same category may present substantial intra-class variations under different mining stages or regional conditions. This characteristic of high inter-class similarity and large intra-class variation makes it difficult for conventional feature-based methods to learn stable and discriminative feature representations, posing greater challenges to model generalization and robustness [1,2].

Under this context, deep learning methods have gradually become the mainstream solution for complex remote sensing scene classification tasks due to their powerful nonlinear feature modeling capability and end-to-end learning paradigm. Particularly in mining-area scenarios, deep models are able to capture subtle differences between land-cover types through hierarchical feature extraction and multi-scale information fusion, thereby alleviating the challenges caused by high inter-class similarity and intra-class variability and achieving more robust classification performance under complex backgrounds. Building upon this, early studies mainly relied on convolutional neural networks (CNNs) to construct hierarchical feature representations. By leveraging multi-scale convolutional structures, CNN-based methods effectively extract local spatial textures and semantic patterns, leading to significant performance improvements in various remote sensing scene classification tasks [3–5]. Subsequently, attention mechanisms were introduced into convolutional networks to enhance the modeling of salient regions and important feature channels, further improving discriminative performance [6]. In recent years, models such as Vision Transformer (ViT) [7] and Swin Transformer [8] have demonstrated superior performance in high-resolution remote sensing scene classification tasks by exploiting the advantages of self-attention in global context modeling [9]. Meanwhile, hybrid architectures that integrate CNNs and Transformers have also been widely adopted to balance local detail representation and global semantic understanding [10–12].

Despite the remarkable improvements in classification accuracy achieved by these deep learning models, they are typically associated with large parameter sizes and high computational complexity, leading to substantial inference costs. In practical mining scenarios such as field inspection, UAV-based monitoring, and real-time surveillance [13], computational resources and response latency are often strictly constrained, making it difficult to directly deploy high-complexity models. Therefore, how to reduce model complexity while maintaining classification accuracy, and thus achieve efficient lightweight modeling, has become a critical issue in mining-area remote sensing scene classification.

As an example, the HTransNet [14] model proposed in our previous work adopts a multi-layer Transformer architecture with cross-attention fusion mechanisms and achieves high classification accuracy on the CUG_MA mining scene dataset. However, due to the extensive use of self-attention layers and complex feature fusion modules, the model involves a substantial number of parameters and high computational complexity. Consequently, its inference process demands considerable computational and memory resources, making real-time deployment on unmanned aerial inspection platforms or mobile devices challenging. This highlights a persistent trade-off between high classification accuracy and computational efficiency in practical mining applications, where a more balanced solution is urgently required.

To reduce model complexity while maintaining performance, various lightweight strategies have been explored, including lightweight network design [15], model pruning [16], and knowledge distillation (KD) [17]. Among these approaches, KD constructs a teacher–student framework in which a compact student model learns the knowledge representations and output distributions of a high-performance teacher model. Without

substantially modifying the network architecture, KD enables effective knowledge transfer and has become an important direction in model compression research. Beyond the classical response-based distillation paradigm [17], feature distillation methods such as FitNet [18] and relational distillation approaches including RKD [19] and CRD [20] have been proposed to enhance the transfer of intermediate representations and sample structural information.

In remote sensing image classification tasks, distillation techniques have also been extensively investigated. Existing studies mainly focus on transferring predictive knowledge from high-performance teacher models to lightweight student networks, so as to improve classification accuracy under limited computational budgets [21,22]. Some recent works further attempt to enhance the distillation process by introducing class-aware constraints or adaptive optimization mechanisms, thereby improving the representational and discriminative capabilities of student models in complex remote sensing scenes [23,24]. Overall, current methods mainly emphasize output-level supervision, class-relation modeling, or holistic feature transfer, while relatively limited attention has been paid to the selective modeling of discriminative channels and spatial regions in complex scene classification tasks.

However, in mining-area scene classification characterized by complex surface structures and high inter-class similarity, existing distillation methods still face several limitations. First, most approaches primarily emphasize output-level alignment while insufficiently exploiting intermediate structural and semantic representations, limiting effective knowledge transfer. Second, selective transfer mechanisms for discriminative channels and spatial regions are often lacking, and indiscriminate feature alignment may introduce redundancy. Third, inter-sample relational modeling remains underexplored, which constrains the student model's ability to capture complex semantic relationships and generalize effectively.

To address these issues, this paper proposes a knowledge distillation and channel-aware non-local attention network (KDANet) for lightweight mining-area scene classification. The proposed framework integrates teacher–student knowledge distillation with channel-aware non-local attention to enhance discriminative feature transfer and improve the representation capability of the lightweight student model. The main contributions of this work are summarized as follows:

1. A teacher–student distillation framework, namely KDANet, is constructed using HTransNet as the teacher model and ShuffleNetV2 as the student model, enabling lightweight modeling for mining-area scene classification.
2. A channel-aware non-local attention module (CANLAM) is designed as an attention-enhanced feature distillation module. By jointly emphasizing discriminative channel responses and modeling long-range contextual dependencies, CANLAM facilitates more effective transfer of high-level semantic knowledge from the teacher model to the lightweight student model.
3. A multi-perspective distillation strategy is introduced by jointly incorporating contrastive distillation loss, intra-class/inter-class relational constraints, and feature distillation loss, thereby improving the representational capacity of the student model from both feature representation and output distribution perspectives.
4. Experimental results on the previously constructed CUG_MA dataset demonstrate that the proposed KDANet significantly reduces computational complexity while achieving classification performance comparable to that of the teacher model.

2. Study Area and Dataset

This study employs the large-scale mining-area remote sensing scene classification dataset CUG_MA, which was previously constructed by our research team and is used as the benchmark dataset in this work (see [14] for details and data access). The dataset covers

three representative mining provinces in China, namely Hubei, Jiangxi, and Heilongjiang, and includes nine typical mining scene categories, i.e., mining area, coal yard, hillock, mining architecture, gangue dump, tailing pond, refuse dump, mineral processing area, and transfer site. Representative samples of these categories are illustrated in Figure 1. It can be observed that certain categories exhibit similar spatial structures and textural patterns, while notable intra-class variations exist under different regional conditions, which increases the difficulty of the classification task.

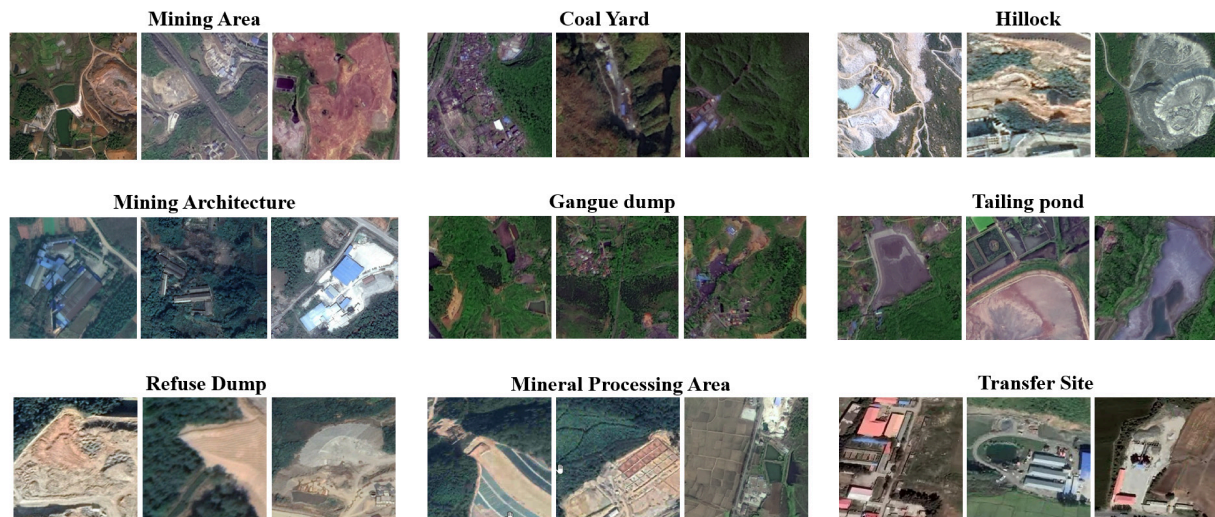


Figure 1. Representative samples of the nine mining scene categories in the CUG_MA dataset.

The dataset is constructed from multi-source high-resolution remote sensing imagery, including data acquired by WorldView-2 (WV-2), WorldView-3 (WV-3), GeoEye-1 (GE-1), SPOT-6, and SPOT-7 satellites. All images are composed of RGB bands, with spatial resolutions ranging from approximately 0.31 m to 2 m, and were primarily acquired in 2017. These sensors may have differences in spatial resolution, radiometric response, and spectral characteristics, which can introduce sensor-related variability into model training and evaluation. In this study, all images are converted into RGB image patches with a unified input format, and standard image normalization is applied during model training to reduce the influence of differences in pixel value ranges. However, no sensor-specific radiometric harmonization is performed. Therefore, the multi-sensor nature of the dataset increases the complexity of the classification task and provides a practical setting for evaluating the robustness of the proposed method under heterogeneous remote sensing imagery. The original images were annotated with the assistance of expert knowledge, followed by quality inspection and preprocessing. All samples were then cropped into patches of size 256×256 pixels, resulting in a total of 2649 images. The number of samples per category ranges from 71 to 463, providing relatively balanced coverage of diverse mining scene types under complex environments.

To ensure objective and fair evaluation, the dataset is split into training, validation, and test sets at a ratio of 6:2:2 using stratified sampling, maintaining consistent class distributions across subsets. All methods in this study are trained and evaluated on the same data split to guarantee fair comparison and reproducibility. It should be noted that this split follows a stratified sample-level evaluation protocol rather than a strict site-wise or region-wise split. Therefore, the reported results mainly reflect the model performance under the current benchmark setting of the CUG_MA dataset. To reduce potential comparison bias, all baseline and distillation methods are evaluated under the identical data partition and optimization settings. We acknowledge that a site-wise or

region-wise split would provide a stricter assessment of cross-region generalization, and this issue is further discussed in the Limitations Section. The dataset is publicly available at: <https://doi.org/10.5281/zenodo.15172547> (accessed on 12 July 2026).

3. Methodology

3.1. Overall Framework

The proposed KDANet is illustrated in Figure 2. The framework consists of three main components: a teacher model, a student model, and a CANLAM. It follows a teacher–student paradigm, where the lightweight student model is guided by joint distillation constraints at the feature, output, and relational levels during training. In this way, the student model learns discriminative knowledge from the teacher model, achieving competitive classification performance while significantly reducing model complexity.

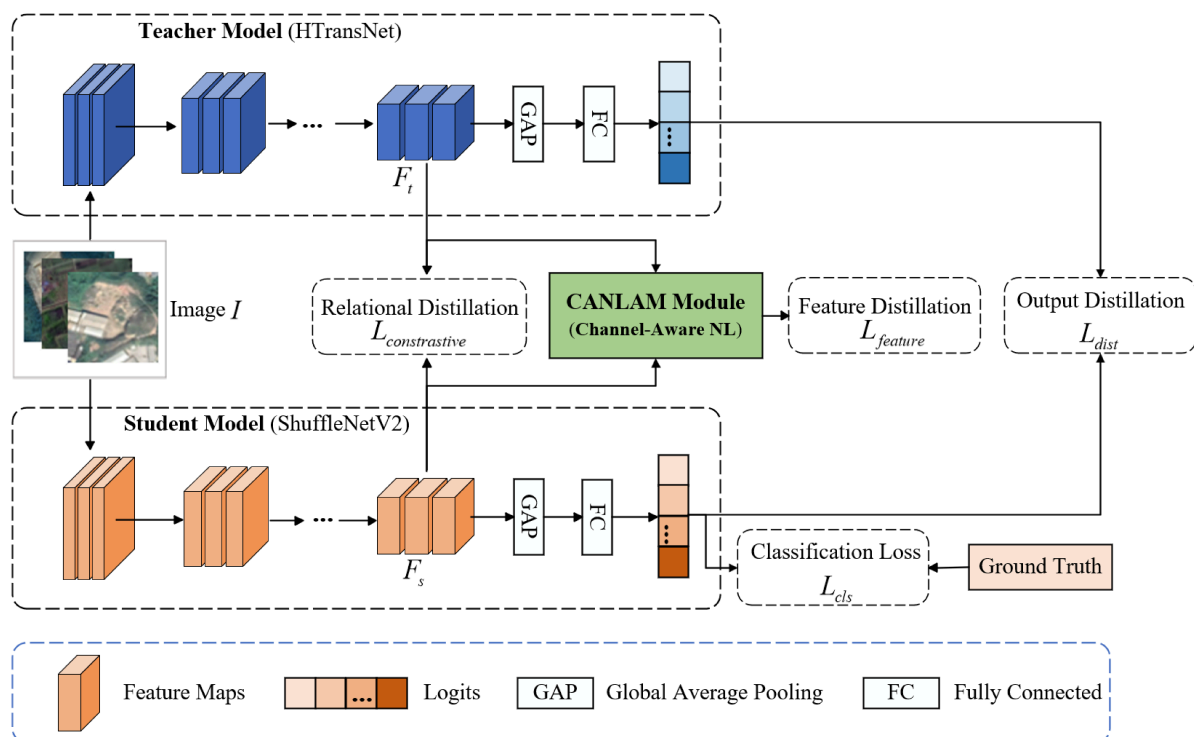


Figure 2. Overall architecture of the proposed KDANet. The input image is processed by both the teacher and student networks. The extracted high-level features are enhanced by CANLAM for feature distillation, while the classification logits are further used for output-level and relational distillation, together with classification loss.

3.1.1. Teacher Model

The teacher network adopts the HTransNet model proposed in our previous work [14]. Built upon a multi-layer Transformer architecture with cross-attention fusion mechanisms, HTransNet can effectively capture multi-scale global contextual information from mining scenes and produce highly discriminative high-level feature representations as well as classification outputs. HTransNet is selected as the teacher model because its hierarchical Transformer structure and attention-based feature modeling provide strong semantic representation capability for complex mining-area scenes. This enables it to serve as a reliable knowledge provider for transferring both feature-level representations and output-level distributions to the lightweight student model. During the distillation process, the parameters of the teacher model remain frozen, and the teacher acts solely as a knowledge provider to transfer feature representations and output distributions to the student model.

3.1.2. Student Model

The student network employs the lightweight architecture ShuffleNetV2 [15]. By adopting efficient design strategies such as channel shuffling, grouped convolution, and depthwise convolution, ShuffleNetV2 significantly reduces parameter size and computational cost while maintaining satisfactory representational capability. As the student model, it performs the final classification task and learns to approximate the high-level feature representations and output distributions of the teacher model through multiple distillation losses during training. ShuffleNetV2 is selected as the student model because it provides a good balance between computational efficiency and representation capability, making it suitable for resource-constrained scenarios such as UAV platforms and edge devices. Its clear complexity gap from the teacher model also helps evaluate whether the proposed distillation strategy can effectively transfer knowledge from a stronger teacher to a compact student. Although other lightweight architectures, such as MobileNetV3, EfficientNet-Lite, MobileViT, EfficientFormer, and EdgeNeXt, are also promising, this study adopts ShuffleNetV2 as a representative lightweight baseline, and the extension to other backbones will be considered in future work.

3.1.3. Distillation Mechanism

Given an input image I , the teacher and student networks extract high-level features denoted as F_t and F_s , respectively. These features are further passed through global average pooling and a fully connected layer to generate the corresponding classification logits. During training, the student model is supervised not only by the ground-truth labels through cross-entropy loss but also by the teacher model from three complementary perspectives: (1) feature-level distillation, which aligns the enhanced high-level feature representations to facilitate semantic knowledge transfer; (2) output-level distillation, which aligns the logits distributions of the teacher and student models to improve prediction consistency; (3) relational distillation, which models structural relationships among samples in the high-level feature space to enhance discriminative representation capability. To further improve feature distillation effectiveness, the proposed CANLAM is introduced to adaptively enhance the high-level features of both the teacher and student networks during feature-level distillation. This module strengthens the modeling of critical channel-wise responses and global contextual dependencies. The enhanced features are then used to compute the feature distillation loss within a unified feature space.

3.2. CANLAM

To enhance the selective representation capability of high-level semantic information during the feature distillation process, the CANLAM is proposed. During the distillation stage, the high-level features extracted from the teacher model and the student model are respectively fed into this module. Specifically, channel attention is first applied to adaptively reweight important feature channels, emphasizing discriminative semantic responses. Subsequently, a non-local attention mechanism is introduced to model long-range spatial dependencies across the feature map, thereby enhancing the effective representation of critical semantic information and global contextual cues.

It is worth emphasizing that CANLAM is not used as a standalone attention block for directly improving classification accuracy. Instead, it is introduced as an attention-guided feature refinement module within the teacher–student distillation framework. Existing channel attention or non-local attention mechanisms usually enhance the feature representation of a single network. In contrast, CANLAM is applied to the high-level features of both the teacher and student models, so that feature distillation is performed in an adaptively enhanced and semantically aligned feature space. In this way, the module serves

as a bridge for more effective knowledge transfer by jointly strengthening discriminative channel responses and global contextual dependencies during distillation.

Compared with existing attention mechanisms, CANLAM differs mainly in its functional role and optimization objective. SENet mainly performs channel-wise recalibration within a single network, while CBAM sequentially combines channel and spatial attention to enhance single-network feature representation. Non-Local Networks focus on capturing long-range dependencies within an individual feature map. In contrast, CANLAM integrates channel-aware feature selection and non-local contextual modeling into the teacher–student feature distillation process, enabling knowledge transfer in an attention-enhanced and semantically aligned feature space.

After the feature enhancement process, the feature distillation loss is computed in the unified enhanced feature space, which encourages the student model to better approximate the discriminative high-level representations learned by the teacher model. The proposed CANLAM consists of two sequential substructures, namely channel attention and non-local attention, and its overall architecture is illustrated in Figure 3.

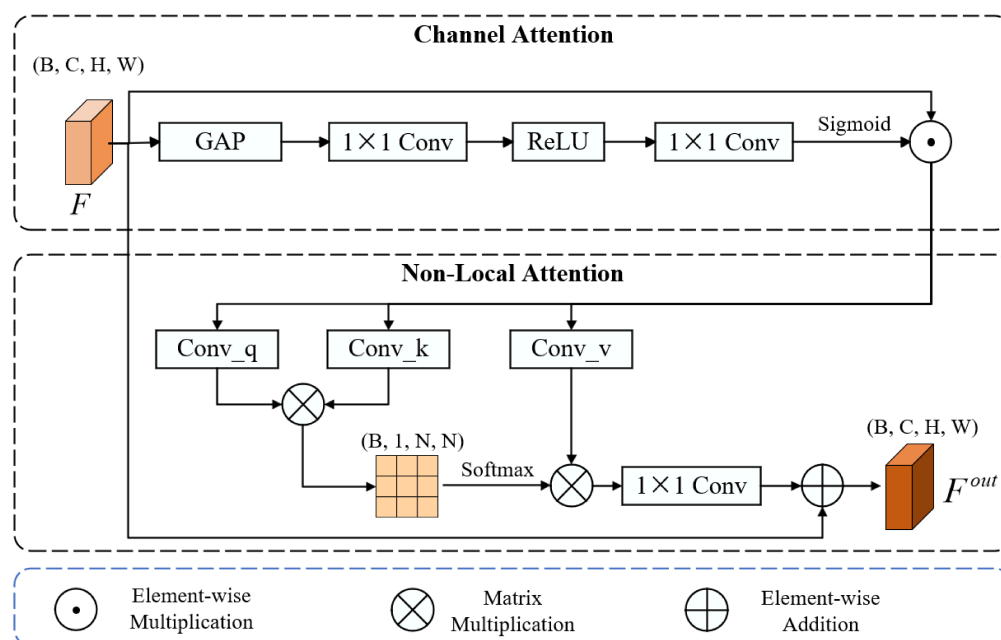


Figure 3. Architecture of the proposed CANLAM. The module sequentially performs channel attention and non-local attention to enhance critical semantic channels and model global contextual dependencies, producing refined features for feature-level distillation.

3.2.1. Channel Attention

In mining-area remote sensing scenes, different channels typically correspond to different types of semantic responses, such as texture, color, or edge-related features, and their contributions to the final classification results may vary. The channel attention mechanism learns the importance distribution across the channel dimension, enabling the adaptive enhancement of informative feature channels while suppressing redundant or noisy responses, thereby improving the discriminative capability of feature representations.

Given an input feature map $F \in \mathbb{R}^{B \times C \times H \times W}$, global average pooling is first applied along the spatial dimensions to obtain channel-wise statistical descriptors:

$$F_{avg} = \text{AvgPool}(F) \in \mathbb{R}^{B \times C \times 1 \times 1} \quad (1)$$

Subsequently, two 1×1 convolution layers are employed to construct a bottleneck structure, where the channel dimension is first reduced to C/r (with $r = 16$ in this work),

followed by a ReLU activation and then restored to the original channel dimension. A Sigmoid function is finally used to generate normalized channel attention weights:

$$Attn_{channel} = \sigma(Con v(ReLU(Con v(F_{avg})))) \quad (2)$$

where σ denotes the Sigmoid activation function. The channel attention weights are then applied to the original feature map through channel-wise element-wise multiplication to obtain the enhanced feature representation:

$$X = F \odot Attn_{channel} \quad (3)$$

Here, $\odot$ denotes element-wise multiplication. The resulting enhanced feature X serves as the input to the subsequent non-local attention module.

3.2.2. Non-Local Attention

After channel refinement, a non-local attention mechanism is introduced to model long-range dependencies among spatial positions, thereby enhancing global contextual representation. By computing pairwise similarity between any two positions in the feature map, this mechanism enables global information aggregation and propagation, which is particularly suitable for mining scenes characterized by spatially scattered objects and significant scale variations.

Given the refined feature map X , three independent 1×1 convolution layers are applied to generate the Query, Key, and Value representations:

$$\theta = Con v_q(X), \phi = Con v_k(X), g = Con v_v(X) \quad (4)$$

The spatial dimensions are then reshaped into $N = H \times W$, resulting in feature representations of size $\mathbb{R}^{B \times C \times N}$. The similarity matrix between Query and Key is computed, followed by Softmax normalization to obtain the attention weights:

$$attention_{map} = Softmax(\theta^T \phi) \quad (5)$$

Using the obtained attention weights, the value features are aggregated to produce a global context-enhanced representation:

$$weighted_g = attention_{map} \times g \quad (6)$$

Finally, the aggregated features are reshaped back to the original spatial dimensions and fused through a 1×1 convolution. A residual connection is introduced to preserve the original information:

$$F^{out} = Con v(weighted_g) + X \quad (7)$$

where ' $\times$ ' denotes matrix multiplication. Through this process, the module effectively enhances feature consistency and global contextual representation in the spatial domain.

3.3. Loss Function Design

To enable the student model to effectively inherit discriminative knowledge from the teacher model, KDANet optimizes a composite objective consisting of contrastive distillation loss, relational distillation loss, feature distillation loss, and classification loss. These losses jointly constrain the student model from the perspectives of feature embedding

distribution, inter-sample relational structure, attention-enhanced feature representation, and ground-truth category supervision. The overall training objective is formulated as:

$$L = \alpha L_{\text{contrastive}} + \beta L_{\text{dist}} + \gamma L_{\text{feature}} + \delta L_{\text{cls}} \quad (8)$$

where $L_{\text{contrastive}}$, L_{dist} , L_{feature} , and L_{cls} denote the contrastive distillation loss, relational distillation loss, feature distillation loss, and classification loss, respectively. The coefficients α , β , γ , and δ are used to balance the contributions of these four loss terms. The detailed mathematical formulations of each loss component are provided in the following subsections.

3.3.1. Contrastive Distillation Loss

The contrastive distillation loss $L_{\text{contrastive}}$ is used to align the feature embeddings of the student and teacher models in a contrastive manner. For a mini-batch containing N samples, let h_i^s and h_i^t denote the normalized feature embeddings of the i -th sample extracted from the student and teacher models, respectively. The teacher embedding of the same sample is treated as the positive sample, while teacher embeddings of other samples in the mini-batch are treated as negative samples. The contrastive distillation loss is formulated as:

$$L_{\text{contrastive}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(h_i^s, h_i^t) / \tau)}{\sum_{j=1}^N \exp(\text{sim}(h_i^s, h_j^t) / \tau)} \quad (9)$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function, and τ is the temperature parameter. By maximizing the similarity between the student and teacher embeddings of the same sample while suppressing the similarity between mismatched samples, this loss encourages the student model to approximate the discriminative embedding distribution of the teacher model.

3.3.2. Relational Distillation Loss

The relational distillation loss L_{dist} is designed to transfer the structural relationships among samples from the teacher model to the student model. Let p_i^s and p_i^t denote the class probability distributions predicted by the student and teacher models for the i -th sample, respectively. The pairwise relational similarity between two samples is computed using cosine similarity:

$$R_{ij}^s = \text{sim}(p_i^s, p_j^s), R_{ij}^t = \text{sim}(p_i^t, p_j^t) \quad (10)$$

where R_{ij}^s and R_{ij}^t represent the relational similarity between the i -th and j -th samples in the student and teacher prediction spaces, respectively. The relational distillation loss is then defined as:

$$L_{\text{dist}} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (R_{ij}^s - R_{ij}^t)^2 \quad (11)$$

By minimizing the discrepancy between the relational similarity matrices of the teacher and student models, this loss encourages the student model to preserve both intra-class and inter-class structural relationships learned by the teacher model.

3.3.3. Feature Distillation Loss

The feature distillation loss L_{feature} measures the discrepancy between the high-level feature representations of the teacher and student models after attention-guided refinement by CANLAM. Let F_i^t and F_i^s denote the high-level feature maps of the teacher and student

models for the i -th sample, respectively. After being processed by CANLAM, the enhanced feature representations are denoted as $\tilde{F}_i^t$ and $\tilde{F}_i^s$. The feature distillation loss is defined as:

$$L_{feature} = \frac{1}{N} \sum_{i=1}^N \|\tilde{F}_i^s - \tilde{F}_i^t\|_2^2 \quad (12)$$

where $\|\cdot\|_2^2$ denotes the squared Euclidean distance. This loss constrains the student model to approximate the attention-enhanced high-level semantic representations of the teacher model, thereby improving semantic consistency during feature-level knowledge transfer.

3.3.4. Classification Loss

The classification loss L_{cls} adopts the standard cross-entropy loss to supervise the student model using the ground-truth labels. Let $y_{i,c}$ denote the one-hot ground-truth label of the i -th sample for class c and $p_{i,c}^s$ denote the predicted probability of the student model for class c . The classification loss is formulated as:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log p_{i,c}^s \quad (13)$$

where C denotes the number of scene categories. This loss ensures that the student model maintains the basic discriminative capability for mining-area scene classification.

4. Experimental Results

4.1. Experimental Settings

4.1.1. Implementation Details

All experiments are conducted under a Linux environment. The hardware platform is equipped with an Intel Xeon E5-2620 v4 CPU with dual processors, an NVIDIA GeForce RTX 2080 Ti GPU with 11 GB memory, and 128 GB RAM. The software environment includes Ubuntu 18.04, Python 3.7, PyTorch 1.8, CUDA 10.2, and cuDNN 7.4.1.

4.1.2. Teacher and Student Configurations

The teacher model is the pre-trained multi-layer Transformer network HTransNet, which achieves an Overall Accuracy (OA) of 70.71% on the CUG_MA test set. The teacher model contains approximately 32.5M parameters and 1.73 GFLOPs. The student model adopts ShuffleNetV2, with approximately 0.78M parameters and 0.659 GFLOPs. Compared with the teacher model, the student model reduces the number of parameters by approximately 97.6% and the FLOPs by approximately 61.9%. During training, HTransNet provides high-level feature and output-level guidance for the student model, and CANLAM is used for attention-guided feature distillation between the teacher and student representations. During inference, the teacher model, CANLAM, and distillation losses are removed, and only the trained ShuffleNetV2 student model is retained for classification. Therefore, the deployed KDANet has the same Params and FLOPs as the student model. For fair comparison, all distillation methods are implemented using the same teacher–student architecture and are trained and evaluated under identical experimental settings.

4.1.3. Loss Weight Settings

The overall loss function is defined in Section 3.3. To balance the contributions of the distillation losses and the classification loss, the weighting coefficients are set as $\alpha = 0.10$, $\beta = 0.10$, $\gamma = 0.10$, and $\delta = 1.0$. The classification loss serves as the primary optimization objective, while the contrastive, relational, and feature distillation losses act as auxiliary constraints for knowledge transfer. Specifically, δ is kept as 1.0 to maintain the dominant role of supervised classification, whereas α , β , and γ are assigned relatively smaller values to

avoid overwhelming the classification objective. To further justify this setting, a sensitivity analysis of different loss weight combinations is conducted in Section 4.4.3.

4.1.4. Training Strategy and Hyperparameters

The model is trained for 200 epochs with a batch size of 16. The initial learning rate is set to 0.001, and the weight decay coefficient is 0.0001. Stochastic Gradient Descent (SGD) is used as the optimizer with a momentum of 0.9. A learning rate decay schedule is applied to gradually reduce the learning rate during training to ensure stable convergence. To improve generalization, data augmentation techniques including random horizontal flipping and random cropping are adopted. All experiments are repeated three times with different random seeds, and the average results are reported as the final performance. This repeated-run setting is used to reduce the influence of training randomness and evaluate model stability under the current benchmark split. It should be noted that it does not represent evaluation across independent data partitions. Therefore, the limitation of sample-level data partitioning and the need for more rigorous site-wise or region-wise evaluation are further discussed in Section 5.2.

4.2. Evaluation Metrics

To evaluate classification performance, OA, average accuracy (AA), and the Kappa coefficient are adopted. All metrics are computed on the test set.

Overall Accuracy is defined as

$$OA = \frac{\sum_{i=1}^C n_{ii}}{N} \quad (14)$$

where n_{ii} denotes the number of correctly classified samples in class i , C represents the total number of classes, and N is the total number of samples.

Average Accuracy is defined as

$$AA = \frac{1}{C} \sum_{i=1}^C \frac{n_{ii}}{\sum_{j=1}^C n_{ij}} \quad (15)$$

where n_{ij} denotes the number of samples belonging to class i and predicted as class j . This metric computes the mean of per-class accuracies.

The Kappa coefficient measures the agreement between predictions and ground-truth labels, defined as

$$Kappa = \frac{P_o - P_e}{1 - P_e} \quad (16)$$

where P_o denotes the observed agreement and P_e represents the expected agreement by chance.

To further provide category-level evaluation, precision, recall, and F1-score are also adopted in this study. For the i -th class, these metrics are defined as follows:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (17)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (18)$$

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (19)$$

where TP_i , FP_i , and FN_i denote the number of true positives, false positives, and false negatives for the i -th class, respectively. Precision reflects the reliability of samples predicted

as a given category, recall reflects the recognition ability for samples belonging to that category, and F1-score provides a balanced measure of precision and recall. These metrics are used to analyze the class-specific performance of KDANet in Section 4.3.

In addition, computational efficiency is assessed by reporting the number of floating-point operations (FLOPs) and the number of parameters (Params) to compare model complexity across methods.

4.3. Quantitative Results

To systematically evaluate the effectiveness and performance of the proposed distillation framework, ten representative methods from three mainstream categories of knowledge distillation are selected for comparison. The first category includes response-based distillation methods: KD [17], DKD [25], WKD-L [26], and LS-KD [27]. The second category consists of feature-based distillation methods: FitNet [18], AT [28], NST [29], and MGD [30]. The third category includes relational distillation methods: RKD [19] and PKT [31]. For fair comparison, all methods are implemented using the same teacher and student architectures and are trained and evaluated using identical data splits and optimization settings. The quantitative results are reported in Table 1. To reduce the influence of training randomness, all OA, AA, and Kappa values are averaged over three repeated runs with different random seeds.

Table 1. Quantitative comparison results on the CUG_MA dataset.

Models	OA (%)	AA (%)	Kappa (%)
Teacher Model [14]	70.71	73.34	66.36
Student Model [15]	56.32	56.54	50.81
KD [17]	58.75	55.74	52.33
DKD [25]	64.64	65.45	59.26
WKD-L [26]	63.69	62.24	58.18
LS-KD [27]	65.02	66.60	59.65
FitNet [18]	56.46	59.50	49.79
AT [28]	65.78	64.78	60.46
NST [29]	58.56	56.26	52.11
MGD [30]	57.32	56.33	50.17
RKD [19]	63.69	59.54	58.19
PKT [31]	67.43	66.73	62.56
(ours) KDANet	69.48	71.17	64.13

Table 1 summarizes the OA, AA, and Kappa coefficient of different distillation methods on the test set. The teacher model achieves 70.71% OA, 73.34% AA, and 66.36% Kappa. In contrast, the vanilla student model obtains 56.32%, 56.54%, and 50.81%, respectively, indicating that the lightweight network alone exhibits limited representational capability under complex mining scene conditions. KDANet achieves 69.48% OA, 71.17% AA, and 64.13% Kappa, achieving the best performance among all compared distillation approaches. Compared with the strongest competing method, PKT, KDANet improves the OA from 67.43% to 69.48%, corresponding to a gain of 2.05 percentage points. Although this improvement is moderate, KDANet still obtains the highest OA, AA, and Kappa under the same experimental setting. Compared with the teacher model, the OA decreases by only 1.23 percentage points, while the student model maintains substantially lower computational complexity, with 6.59×10^8 FLOPs and 0.78M parameters, corresponding to approximately 38% and 2.4% of the teacher model, respectively. This demonstrates that the proposed framework effectively preserves performance while significantly reducing model complexity.

Among response-based methods, KD provides limited improvement over the student baseline (58.75% OA), indicating that aligning output distributions alone is insufficient for complex scene understanding. More advanced variants such as DKD and LS-KD improve performance to 64.64% and 65.02%, respectively, yet remain inferior to KDANet.

For feature-based distillation, AT achieves the strongest performance (65.78% OA), suggesting that attention-guided feature alignment contributes positively to knowledge transfer. Other methods such as FitNet, NST, and MGD yield relatively modest gains.

In relational distillation, PKT attains 67.43% OA, highlighting the importance of structural information modeling, while RKD achieves 63.69%.

Overall, the proposed KDANet achieves the best performance among the three categories of distillation strategies under the same experimental setting. This improvement can be attributed to the integration of CANLAM, which strengthens critical channel and global contextual information, together with the joint optimization of contrastive, relational, and feature-level distillation losses, enabling more comprehensive knowledge transfer from the teacher model. Nevertheless, we avoid claiming statistical significance based solely on the current three repeated runs, and more extensive repeated evaluations and statistical tests will be considered in future work.

To further analyze the category-level classification behavior of KDANet, class-specific precision, recall, and F1-score are reported in Table 2. As shown in Table 2, KDANet achieves relatively high performance for categories with distinctive visual and spatial patterns. For example, Tailing Pond obtains the highest F1-score of 83.06%, followed by Mining Area with an F1-score of 81.94% and Hillock with an F1-score of 78.34%. These results indicate that KDANet can effectively recognize mining scene categories with relatively clear structural or textural characteristics.

Table 2. Class-specific performance of KDANet on the CUG_MA dataset.

Class	Precision (%)	Recall (%)	F1-Score (%)
Mining Area	82.84	81.06	81.94
Coal Yard	67.37	64.71	66.01
Hillock	79.17	77.52	78.34
Mining Architecture	68.43	68.26	68.34
Gangue Dump	69.38	67.43	68.39
Tailing Pond	83.52	82.61	83.06
Refuse Dump	66.79	65.32	66.05
Mineral Processing Area	70.50	68.73	69.60
Transfer Site	66.89	64.89	65.87

In contrast, categories such as Coal Yard, Refuse Dump, and Transfer Site obtain relatively lower F1-scores of 66.01%, 66.05%, and 65.87%, respectively. This may be attributed to the high visual similarity among material storage areas, waste accumulation regions, and transportation-related sites in mining scenes. In addition, variations in imaging conditions, spatial layouts, and object scales may further increase the intra-class variability of these categories. These category-level results provide a more detailed understanding of the strengths and limitations of KDANet, showing that CANLAM and the proposed distillation strategy improve the overall discriminative capability of the lightweight student model, while visually similar mining-related categories remain challenging.

4.4. Ablation Study

4.4.1. Ablation Study of Distillation Losses

To investigate the contribution of each distillation loss component, ablation experiments are conducted under identical teacher–student architectures and training settings.

The feature-level distillation loss $L_{feature}$, relational distillation loss L_{dist} , and contrastive distillation loss $L_{contrastive}$ are removed individually to evaluate their impact on overall performance. The Overall Accuracy (OA) under different loss combinations is reported in Table 3.

Table 3. Ablation study of distillation loss components on the CUG_MA dataset.

$L_{feature}$	L_{dist}	$L_{contrastive}$	OA (%)
✓	✓	✓	69.48
✓	✓	✗	68.29
✓	✗	✗	60.11
✗	✓	✗	65.11
✓	✗	✓	61.6
✗	✓	✓	67.2

Note: ✓ indicates that the corresponding loss component is included, whereas ✗ indicates that it is excluded.

As shown in Table 3, the complete model incorporating all three distillation losses achieves the highest OA of 69.48%. Removing any individual component leads to performance degradation, confirming the effectiveness of each constraint.

Among the three components, the relational distillation loss L_{dist} has the most significant impact. When L_{dist} is removed (while retaining $L_{feature}$ and $L_{contrastive}$), OA decreases to 61.60%, indicating that modeling intra-class and inter-class relationships plays a critical role in knowledge transfer.

Removing the feature-level distillation loss $L_{feature}$ results in an OA of 67.20%, suggesting that feature alignment in the attention-enhanced representation space contributes to improved semantic consistency. Since $L_{feature}$ operates on features processed by CANLAM, its removal also weakens the supervision imposed on the enhanced feature space.

In contrast, the independent impact of the contrastive loss $L_{contrastive}$ is relatively moderate, with OA decreasing to 68.29% when excluded. However, combined loss configurations demonstrate that contrastive regularization stabilizes feature distribution and enhances overall consistency.

Overall, the three losses exhibit complementary effects: L_{dist} primarily captures structural relationships, $L_{feature}$ refines semantic feature alignment, and $L_{contrastive}$ regularizes the embedding space. Their joint optimization enables stable performance improvement without compromising the lightweight design of the student model.

4.4.2. Ablation Study of CANLAM Components

To further evaluate the contribution of the proposed CANLAM, additional ablation experiments are conducted by removing the whole module or retaining only one of its two components. Specifically, four configurations are compared: without CANLAM, using only channel attention, using only non-local attention, and using the complete CANLAM. All experiments are conducted under the same teacher–student architecture, training strategy, and data split. For consistency with the ablation study of distillation losses, OA is used as the evaluation metric in this ablation analysis.

As shown in Table 4, removing CANLAM results in an OA of 64.52%, indicating that directly performing feature distillation without attention-guided feature refinement is less effective. When only channel attention is used, the OA increases to 67.71%, suggesting that channel-wise feature reweighting helps emphasize discriminative semantic responses. When only non-local attention is used, the OA reaches 66.35%, showing that long-range contextual dependency modeling also contributes to feature distillation. The complete CANLAM achieves the highest OA of 69.48%, outperforming the configuration without CANLAM by 4.96 percentage points. These results demonstrate that channel attention

and non-local attention provide complementary benefits and verify that the performance improvement of KDANet is not solely derived from the loss design but also from the proposed attention-guided feature distillation mechanism.

Table 4. Ablation study of CANLAM components on the CUG_MA dataset.

Configuration	Channel Attention	Non-Local Attention	OA (%)
w/o CANLAM	✗	✗	64.52
Only Channel Attention	✓	✗	67.71
Only Non-Local Attention	✗	✓	66.35
Full CANLAM	✓	✓	69.48

Note: ✓ indicates that the corresponding component is included, whereas ✗ indicates that it is excluded.

4.4.3. Sensitivity Analysis of Loss Weight Coefficients

To further justify the selection of the loss weighting coefficients, a sensitivity analysis is conducted by varying the weights of the three distillation losses while keeping the classification loss weight fixed. Since the classification loss provides the primary supervised learning objective, δ is fixed at 1.0 in all experiments, while α , β , and γ are adjusted to evaluate the influence of different distillation strengths. All experiments are conducted under the same teacher–student architecture, data split, and training strategy. The OA results under different weight combinations are reported in Table 5.

Table 5. Sensitivity analysis of loss weight coefficients on the CUG_MA dataset.

α	β	γ	δ	OA (%)
0.05	0.05	0.05	1.0	68.37
0.10	0.10	0.10	1.0	69.48
0.20	0.20	0.20	1.0	68.75
0.10	0.05	0.10	1.0	68.96
0.10	0.10	0.05	1.0	68.52

As shown in Table 5, different loss weight combinations lead to different classification performance, indicating that the balance among contrastive, relational, and feature-level distillation losses influences the optimization of KDANet. The setting $\alpha = 0.10$, $\beta = 0.10$, $\gamma = 0.10$, and $\delta = 1.0$ achieves the highest OA of 69.48% among the tested configurations. When the three distillation weights are reduced to 0.05, the OA decreases to 68.37%, suggesting that weak distillation constraints may provide insufficient guidance from the teacher model. When the weights are increased to 0.20, the OA decreases to 68.75%, indicating that excessive distillation constraints may interfere with the dominant supervised classification objective. In addition, reducing β or γ also leads to performance degradation, further demonstrating the necessity of balancing relational and feature-level distillation. Therefore, the adopted coefficient setting provides a suitable balance between classification supervision and knowledge distillation.

5. Discussion

5.1. Effectiveness and Mechanism Analysis

This study addresses the challenge of deploying high-accuracy models under resource-constrained conditions in mining area scene classification by integrating the CANLAM with a multi-perspective distillation framework. Experimental results demonstrate that the proposed method significantly reduces model complexity while maintaining performance close to that of the teacher model.

The ablation study indicates that the performance gain arises from the complementary interaction among feature-level, relational, and contrastive distillation, rather than from a single structural component. In complex mining scenes characterized by high inter-class visual similarity and subtle spatial differences, aligning output distributions alone is insufficient to transfer discriminative knowledge effectively. By incorporating channel-selective modeling and relational constraints, the student model achieves more stable semantic representations while preserving lightweight advantages. These findings suggest that multi-perspective distillation is well suited for complex remote sensing scenarios requiring efficient deployment.

From a practical application perspective, the OA of 69.48% should be interpreted in the context of fine-grained mining-area scene classification. The CUG_MA dataset contains nine mining-related categories with high inter-class similarity and notable intra-class variability, such as coal yard, gangue dump, refuse dump, mineral processing area, and transfer site. Under this challenging setting, KDANet maintains a relatively small performance gap compared with the teacher model while substantially reducing model complexity, indicating its potential value for rapid scene screening and auxiliary interpretation in large-scale mining-area monitoring. However, the current accuracy is not intended to fully replace expert interpretation in operational applications. Instead, KDANet is more suitable as an efficient preliminary classification tool, while visually similar or uncertain categories may still require expert verification or more fine-grained analysis.

5.2. Limitations and Future Work

Despite the promising results, several limitations remain. First, certain categories in the CUG_MA dataset contain relatively fewer samples, which may affect learning performance for underrepresented classes. Second, the effectiveness of knowledge distillation inherently depends on the quality of the teacher model. Third, the current framework is built upon single-modal optical imagery and does not exploit complementary information from DEM, SAR, or hyperspectral data. In addition, although standard image normalization is applied, sensor-related radiometric and spectral differences may still affect cross-sensor transferability, which will be further explored in future work. Fourth, the current evaluation is conducted on a stratified sample-level split of the CUG_MA dataset. Although this setting ensures fair comparison among different methods, it may not fully reflect cross-site or cross-region generalization under more challenging deployment scenarios. In future work, site-wise and region-wise evaluation protocols will be further considered to assess the transferability of the proposed framework across different mining regions and geographical conditions. Moreover, the current experiments are conducted only on the CUG_MA dataset, and the generalization capability of KDANet to other mining-area datasets or broader remote sensing scene classification domains still requires further validation. Finally, although Params and FLOPs are reported to evaluate model complexity, practical deployment efficiency may also be affected by hardware platforms, inference frameworks, memory access patterns, and batch settings. During inference, KDANet retains only the lightweight student model, while the teacher model and distillation losses are used only during training and introduce no additional inference burden. Future work will further evaluate inference time, FPS, and memory consumption on UAV platforms and edge devices to provide a more comprehensive deployment assessment.

Future work may explore category rebalancing strategies, multi-teacher or self-distillation mechanisms, multimodal distillation frameworks, validation on additional mining-area and public remote sensing scene classification datasets, more rigorous cross-region evaluation strategies, and feature visualization analyses to further enhance generalization capability, practical applicability, and model interpretability.

6. Conclusions

This study addresses the challenge that high-accuracy models for mining-area remote sensing scene classification often exhibit high computational complexity and are difficult to deploy in resource-constrained environments. To tackle this issue, a lightweight classification method, KDANet, is proposed by integrating knowledge distillation with CANLAM. By constructing a distillation framework with HTransNet as the teacher model and ShuffleNetV2 as the student model, effective knowledge transfer from a high-performance model to a lightweight network is achieved. At the methodological level, CANLAM enhances the representation of critical semantic information and global contextual dependencies in high-level features, thereby improving the efficiency of discriminative knowledge transfer during feature distillation. Meanwhile, a multi-perspective distillation strategy is adopted by jointly incorporating contrastive distillation loss, intra-class and inter-class relational loss, and feature distillation loss, enabling the student model to approximate the semantic representation capability of the teacher model from both feature representation and output distribution perspectives.

Experimental results on the previously constructed CUG_MA dataset demonstrate that, with reductions of approximately 52% in parameters and 62% in computational cost, KDANet achieves an OA of 69.48%. Compared with the teacher model, KDANet maintains only a small performance gap while outperforming several mainstream distillation approaches. Ablation experiments further verify the effectiveness and complementary benefits of CANLAM and the proposed multi-perspective distillation strategy.

Overall, KDANet provides an effective solution for lightweight remote sensing scene interpretation in complex mining environments and offers practical potential for deployment in resource-constrained scenarios. Future work will explore directions such as multi-teacher distillation, self-distillation mechanisms, and the integration of multimodal remote sensing data to further improve model generalization and practical applicability. The source code of this study is publicly available at <https://github.com/ljwang8/KDANet> (accessed on 12 July 2026).

Author Contributions: W.H., Z.L., L.W. and W.C. contributed to the study conception and design. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Project of Remote Sensing Monitoring of Natural Resources and Territorial Space in Central South China under Grant DD202607202807.

Data Availability Statement: The CUG_MA dataset used in this study is publicly available at <https://doi.org/10.5281/zenodo.15172547> (accessed on 12 July 2026). The source code is publicly available at <https://github.com/ljwang8/KDANet> (accessed on 12 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Chen, W.; Li, X.; He, H.; Wang, L. A review of fine-scale land use and land cover classification in open-pit mining areas by remote sensing techniques. *Remote Sens.* **2018**, *10*, 15. [CrossRef]
2. Maus, V.; Giljum, S.; Gutschlhofer, J.; da Silva, D.M.; Probst, M.; Gass, S.L.B.; Luckeneder, S.; Lieber, M.; McCallum, I. A global-scale data set of mining areas. *Sci. Data* **2020**, *7*, 289. [CrossRef] [PubMed]
3. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
4. Xia, G.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]

5. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
6. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
7. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
8. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision Transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
9. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.-S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [[CrossRef](#)]
10. Li, N.; Hao, S.; Zhao, K. A hybrid CNN–vision Transformer structure for remote sensing scene classification. *Remote Sens. Lett.* **2024**, *15*, 88–98. [[CrossRef](#)]
11. Ji, R.; Tan, K.; Wang, X.; Tang, S.; Sun, J.; Niu, C.; Pan, C. PatchOut: A novel patch-free approach based on a Transformer-CNN hybrid framework for fine-grained land-cover classification on large-scale airborne hyperspectral images. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *138*, 104457. [[CrossRef](#)]
12. Guo, X.; Feng, Q.; Guo, F. CMTNet: A hybrid CNN-transformer network for UAV-based hyperspectral crop classification in precision agriculture. *Sci. Rep.* **2025**, *15*, 12383. [[CrossRef](#)] [[PubMed](#)]
13. Du, W.; Sheng, Q.; Fu, X.; Chen, J.; Kang, J.; Pang, X.; Wan, D.; Yuan, W. Application of unmanned aerial vehicle remote sensing on dangerous rock mass identification and deformation analysis: Case study of a high-steep slope in an open pit mine. *J. Earth Sci.* **2025**, *36*, 750–763. [[CrossRef](#)]
14. Chen, W.; Li, Z.; Wang, H.; He, W.; Chen, Z.; Li, J. HTransNet: A hierarchical Transformer network with dual attention for large-scale mining scene classification using multi-category high-resolution remote sensing dataset. *J. Earth Sci.* **2025**. [[CrossRef](#)]
15. Ma, N.; Zhang, X.; Zheng, H.; Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 116–131.
16. Han, S.; Pool, J.; Tran, J.; Dally, W. Learning both weights and connections for efficient neural network. In Proceedings of the 29th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015.
17. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
18. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
19. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3967–3976.
20. Tian, Y.; Krishnan, D.; Isola, P. Contrastive representation distillation. *arXiv* **2019**, arXiv:1910.10699.
21. Song, H.; Wei, C.; Yong, Z. Efficient knowledge distillation for remote sensing image classification: A CNN-based approach. *Int. J. Web Inf. Syst.* **2024**, *20*, 129–158. [[CrossRef](#)]
22. Xu, K.; Deng, P.; Huang, H. Vision Transformer: An excellent teacher for guiding small networks in remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618715. [[CrossRef](#)]
23. Li, C.; Teng, X.; Ding, Y.; Lan, L. Instance-level scaling and dynamic margin-alignment knowledge distillation for remote sensing image scene classification. *Remote Sens.* **2024**, *16*, 3853. [[CrossRef](#)]
24. Wu, B.; Hao, S.; Wang, W. Class-aware self-distillation for remote sensing image scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 2173–2188. [[CrossRef](#)]
25. Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; Liang, J. Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 11953–11962.
26. Lv, J.; Yang, H.; Li, P. Wasserstein distance rivals Kullback-Leibler divergence for knowledge distillation. In Proceedings of the 38th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; pp. 65445–65475.
27. Sun, S.; Ren, W.; Li, J.; Wang, R.; Cao, X. Logit standardization in knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 15731–15740.
28. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
29. Huang, Z.; Wang, N. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv* **2017**, arXiv:1707.01219.

30. Yang, Z.; Li, Z.; Shao, M.; Shi, D.; Yuan, Z.; Yuan, C. Masked generative distillation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 53–69.
31. Passalis, N.; Tzelepi, M.; Tefas, A. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2030–2039. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.