remote sensing

MDPI

Article

Ensemble Network-Based Distillation for Hyperspectral Image Classification in the Presence of Label Noise

Youqiang Zhang [1,2,*], Ruihui Ding [3], Hao Shi [4], Jiaxi Liu [1], Qiqiong Yu [4], Guo Cao [4] and Xuesong Li [2,5]

[1] School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[2] Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tiangong University, Tianjin 300387, China
[3] Portland Institute, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[4] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[5] School of Computer Science and Technology, Tiangong University, Tianjin 300387, China
* Correspondence: zhangyq@njupt.edu.cn

Abstract: Deep learning has made remarkable strides in hyperspectral image (HSI) classification, significantly improving classification performance. However, the challenge of obtaining accurately labeled training samples persists, primarily due to the subjectivity of human annotators and their limited domain knowledge. This often results in erroneous labels, commonly referred to as label noise. Such noisy labels can critically impair the performance of deep learning models, making it essential to address this issue. While previous studies focused on label noise filtering and label correction, these approaches often require estimating noise rates and may inadvertently propagate noisy labels to clean labels, especially in scenarios with high noise levels. In this study, we introduce an ensemble network-based distillation (END) method specifically designed to address the challenges posed by label noise in HSI classification. The core idea is to leverage multiple base neural networks to generate an estimated label distribution from the training data. This estimated distribution is then used alongside the ground-truth labels to train the target network effectively. Moreover, we propose a parameter-adaptive loss function that balances the impact of both the estimated and ground-truth label distributions during the training process. Our approach not only simplifies architectural requirements but also integrates seamlessly into existing deep learning frameworks. Comparative experiments on four hyperspectral datasets demonstrate the effectiveness of our method, highlighting its competitive performance in the presence of label noise.

Keywords: ensemble network; distillation; hyperspectral image; classification

Citation: Zhang, Y.; Ding, R.; Shi, H.; Liu, J.; Yu, Q.; Cao, G.; Li, X. Ensemble Network-Based Distillation for Hyperspectral Image Classification in the Presence of Label Noise. *Remote Sens.* **2024**, *16*, 4247. https://doi.org/10.3390/rs16224247

Academic Editor: Salah Bourennane

Received: 14 September 2024
Revised: 6 November 2024
Accepted: 12 November 2024
Published: 14 November 2024

1. Introduction

Hyperspectral imaging provides rich spectral information, enabling fine-grained classification of ground objects with greater accuracy and detail. This capability is of immense importance across diverse domains including agricultural monitoring [1], forest inventory control [2], and urban planning and management [3]. For instance, Yuan et al. [4] used GaoFen-5 hyperspectral satellite imagery to assess and analyze the classification of complex urban functional zones located in the heart of Wuhan, China. Similarly, Rajamani et al. [5] utilized a convolutional neural network to analyze building footprints and detect roads. Fine-grained classification is essential for the effective use of hyperspectral remote sensing. As computer technology has advanced, methods for processing hyperspectral data have steadily improved.

In recent years, machine learning methods, particularly deep learning, have emerged as promising approaches for HSI classification tasks [6]. As a supervised method, deep learning heavily depends on accurate labels to train reliable models. Accurate labels are crucial because incorrect or noisy labels can cause the model to learn incorrect patterns,

leading to poor performance on unseen data. Labeling complex data like HSIs is often time-consuming and challenging, typically requiring human annotators to manually assign labels to each data point. This process can be subjective and prone to errors. Despite best efforts, label noise is unavoidable due to the limitations of expert knowledge and the inherent subjectivity of human annotation [7].

In the early stages, researchers proposed several traditional methods to address label noise in HSI classification [8–11]. More recently, significant efforts have focused on tackling label noise using deep learning methodologies. For example, Ghafari et al. [12] investigated the robustness of convolutional neural networks by evaluating the performance of different loss functions, including cross-entropy, pseudo-Huber, and correntropy, on noisy hyperspectral data. Xu et al. [13] developed a novel dual-channel residual network to mitigate the effects of noisy labels. Roy et al. [14] integrated spectral and spatial domain convolutional kernels within a heterogeneous kernel convolution framework designed specifically for HSI classification with label noise. Wang et al. [15] investigated attention mechanisms and introduced an end-to-end attentive-adaptive network architecture, combined with a noise-resistant loss function to enhance overall efficiency. Zhang et al. [16] proposed a triple contrastive learning framework that explored cluster-, instance-, and structure-level representations of HSI data. To address the challenges posed by label noise, Ma et al. [17] designed a noise-tolerant learning algorithm within a spatial pooling transformer network. Xu et al. [18] introduced a superpixel-guided sample selection network for HSI classification with noisy labels, aiming to prevent the spread of label noise and correct noisy labels using a superpixel technique. Wei et al. [19] developed a unified deep learning network that effectively leverages both labeled and unlabeled data, addressing the twin challenges of limited samples and label noise. Similarly, Wang et al. [20] introduced a dual-level deep spatial manifold representation network, specifically designed for HSI classification scenarios where training samples are scarce or corrupted by noise.

Deep learning-based methods for handling label noise in HSI classification offer promising solutions. These approaches generally avoid filtering out samples with label noise, instead directly utilizing noisy labeled data to train models. However, previous methods often rely on designing complex network architectures or employing multiple techniques to effectively manage label noise. Moreover, some approaches may unintentionally exacerbate the spread of label noise to correctly labeled samples, particularly when using hard label corrections. This issue becomes even more pronounced in high-noise scenarios.

To address the aforementioned problem, we propose an innovative approach called the ensemble network-based distillation (END) method, specifically designed for HSI classification. The main idea behind END is to leverage an ensemble network to estimate the label distribution of training data affected by label noise. This estimated label distribution is then combined with the ground-truth label distribution to train a student network. Unlike previous methods that rely on corrected samples with hard labels for training, our END method generates soft labels by distilling label distribution information from the ensemble network. During the training student network's training phase, we introduce a robust loss function that intelligently incorporates both the soft labels from the ensemble teacher and the original ground-truth labels. To effectively balance the influence of these two label sources, we include an adaptive parameter within the loss function.

The main contributions of our research are outlined as follows:

(1) We introduce an ensemble-based strategy that revolutionizes the way hyperspectral data's label distribution is evaluated, offering a more comprehensive and accurate perspective;
(2) The knowledge distillation technique is utilized to train the classification network, considering both the estimated label distribution and the ground-truth label distribution;
(3) A robust loss function with an adaptive parameter is designed for the classification network, avoiding the need to estimate the noise rate;
(4) Extensive experiments on real hyperspectral datasets demonstrate that our END method achieves competitive results.

The rest of this paper is organized as follows: Section 2 reviews related work. The proposed END method is introduced in Section 3. Section 4 reports the experimental results and analysis. Section 5 provides a discussion, and finally, Section 6 presents the conclusions of this work.

2. Related Work

2.1. Traditional Methods for HSI Classification

Over the past few decades, researchers have developed a variety of traditional methods for HSI classification. These methods primarily focus on extracting and analyzing spectral and spatial features and can be broadly categorized into three main groups:

(1) Spectral feature analysis methods: These methods focus on the spectral information of each pixel by analyzing reflectance or absorption characteristics across different bands for data classification. Common techniques include spectral representation and band selection [21,22], minimum distance classifiers [23], maximum likelihood classifiers [24], discriminant analysis [25], random forests [26,27], and support vector machines [27,28]. These approaches utilize statistical and machine learning models to leverage the spectral features of HSI data for identification and classification;

(2) Spatial feature analysis methods: These methods leverage spatial information to improve classification accuracy. Typical approaches include texture analysis-based classification techniques [29], Markov random fields [30], and morphological filtering techniques [31]. These approaches utilize spatial structural features of the image, such as edges, shapes, and textures, to assist in classification and enhance the spatial consistency and accuracy of the results;

(3) Spectral–spatial joint analysis methods: To further enhance classification performance, some methods integrate both spectral and spatial features for joint analysis. An illustrative example is object-based image analysis (OBIA) [32], which segments the image into distinct objects and utilizes both their spectral and spatial attributes for classification, providing a comprehensive and nuanced approach to image interpretation. Additionally, sparse representation and multi-scale segmentation methods [33,34] are commonly employed in spectral–spatial HSI classification;

Traditional HSI classification methods have strengths in handling high-dimensional data and capturing subtle differences, but they also face challenges, such as high computational complexity. With the advancement of deep learning technologies, there is a growing focus on applying deep learning methods to HSI classification to address the limitations of traditional approaches.

2.2. Deep Learning-Based Methods for HSI Classification

In recent years, the rapid advancement of deep learning technology has led to its increased application in HSI classification. Leveraging powerful feature extraction and learning capabilities, deep learning methods have significantly improved HSI classification performance. These deep learning-based methods for HSI classification can be broadly categorized as follows:

(1) Convolutional neural networks (CNNs): Early 2D-CNN methods [35–37] primarily used two-dimensional convolution operations to extract either spectral or spatial features. While effective, these methods often addressed only one dimension at a time, potentially missing the high-dimensional characteristics of hyperspectral data. To overcome this limitation, 3D-CNN methods [38,39] employ three-dimensional convolution operations, allowing for the simultaneous processing of both spectral and spatial features. This approach significantly enhances classification accuracy by leveraging the full potential of hyperspectral data;

(2) Recurrent neural networks (RNNs) and long short-term memory networks (LSTMs): RNNs are well-suited for processing sequential data and capturing long-range dependencies in the spectral dimension. For instance, Mou et al. [40] were among the first to apply RNNs to HSI classification, effectively modeling spectral sequence information.

LSTMs, a specialized type of RNN, handle long-range dependencies more effectively and mitigate issues like gradient vanishing. As a result, they are widely used in HSI classification to extract spectral dependencies [41,42];

(3) Deep belief networks (DBNs) and sparse auto-encoders (SAEs): DBNs exploit layer-wise unsupervised training methods to carefully extract complex features from HSIs. For example, Zhong et al. [43] introduced a novel diversified DBN framework specifically designed for HSI classification, demonstrating the versatility and efficacy of this approach. Auto-encoders perform feature extraction and data reconstruction through encoding and decoding processes. SAEs enhance these capabilities by improving feature compression and dimensionality reduction [44];

(4) Generative adversarial networks (GANs): GANs utilize adversarial training between a generator and a discriminator to create realistic HSI data for data augmentation and sample expansion. They hold significant potential for addressing challenges related to HSI data labeling [45,46];

(5) Advanced models: Attention mechanisms, Transformers, and Mamba models have been widely applied in HSI classification. For instance, Sun et al. [47] introduced a spectral–spatial attention network that enhances performance by integrating an attention module to extract key features from hyperspectral images. Liu et al. [48] analyzed the properties of HSI and developed a scaled dot-product central attention mechanism for spectral–spatial feature extraction, leading to the creation of a central attention network. Scheibenreif et al. [49] proposed a spatial-spectral factorization for Transformers, which reduces computational load while improving performance on hyperspectral data through self-supervised learning. More recently, Mamba models [50] have been investigated in HSI classification due to their strong long-distance modeling capabilities and linear computational complexity. Xu et al. [51] combined orientational learning with CNN to develop an orientational clustering method for open-set HSI classification;

(6) Hybrid models: Hybrid models integrate deep learning techniques with other methodologies, such as active learning, semi-supervised learning, and transfer learning. For example, Di et al. [52] incorporated active learning into a Siamese network to reduce labeling costs. Wu et al. [53] merged semi-supervised learning with deep learning to leverage both limited labeled data and abundant unlabeled data for training effective deep neural networks. Zhong et al. [54] applied deep transfer learning for cross-scene HSI classification.

In summary, deep learning technology offers powerful tools and methods for HSI classification. Its continued development and optimization are anticipated to drive further advancements and expand applications in this field.

2.3. Label Noise Learning in HSI Classification

Label noise refers to errors or inconsistencies in the labeling of training samples, which can significantly degrade the performance of classification models. This issue has been widely investigated in the fields of machine learning and computer vision. Recent surveys on label noise learning can be found in [55–57]. Several studies related to our research have emerged in recent years. For example, Li et al. [58] introduced a unified distillation framework that leverages auxiliary information, such as a small clean dataset and label relationships within a knowledge graph, to mitigate the risks associated with learning from noisy labels. Lukov et al. [59] proposed soft label smoothing (SLS) to address noisy labels by adjusting high-confidence class probabilities and assigning lower probabilities to low-confidence classes. Algan and Ulusoy [60] developed MetaLabelNet (MLN), a label-noise-robust algorithm that trains soft labels using meta-objectives, optimizing gradients to minimize meta-data loss through a single-layer perceptron. Similarly, Wu et al. [61] proposed a meta-learning model that automatically estimates soft labels using meta-gradient descent, adapting label corrections iteratively based on a small amount of noise-free meta-data.

In HSI, label noise can arise from several sources [7]. Firstly, human annotators may introduce labeling errors due to limited experience. Secondly, confusion between similar spectral signatures of different classes can result in incorrect labeling. Finally, variability in sensor performance and environmental conditions during data acquisition can also contribute to label noise. To address the issue of label noise in HSI classification, researchers have employed various strategies, which can be broadly categorized into two main approaches.

(1) Label noise detection and cleaning methods: These methods focus on identifying and correcting noisy labels before training. Common techniques include using cross-validation to detect outliers and applying majority voting among multiple classifiers. For example, Tu et al. [62–64] developed a series of outlier detection methods specifically for label noise. Kang et al. [65] introduced an innovative method based on constrained energy minimization to identify and correct mislabeled training samples, thereby improving data quality. Leng et al. [66] proposed a spectral–spatial sparse graph-based adaptive label propagation technique that facilitates the recognition and iterative correction of "polluted" samples, refining the dataset for better classification results. Bahraini et al. [67] suggested a modified mean-shift method to detect and remove mislabeled samples from the training set;

(2) Label noise robust classification models: These methods aim to enhance classification performance by developing models that are resilient to label noise. Key approaches include designing noise-tolerant loss functions, applying regularization techniques, and developing novel learning architectures. For example, Kang et al. [68] explored deep metric learning and introduced a robust normalized softmax loss function specifically for remote sensing images. Damodaran et al. [69] developed a loss function based on optimal transport theory to improve deep learning under label noise. Zhang et al. [70] implemented a co-learning strategy using a dual network architecture to address the challenges of HSI classification with noisy labels. Liao et al. [71] proposed a meta-learning framework that employs joint positive and negative learning to adaptively reweight samples, enhancing classification robustness. Fang et al. [72] introduced a deep reinforcement learning method to address label noise in HSI classification.

In summary, label noise robust classification models provide several advantages over label noise detection and cleaning methods. Firstly, they utilize the entire training data without discarding noisy samples. Secondly, robust classification models, especially those incorporating deep learning techniques, can achieve superior performance by extracting more complex features for improved modeling.

2.4. Knowledge Distillation in HSI Classification

Knowledge distillation is a model compression technique designed to transfer knowledge from a large, complex model (often referred to as the teacher) to a smaller, more efficient model (the student). This process involves two key stages. First, the teacher model is trained on labeled data, generating soft labels—probability distributions over classes—that capture class similarities. These soft labels provide richer information than hard labels alone. In the second stage, the student model learns by mimicking the teacher's soft labels in conjunction with the original hard labels, enabling it to understand both the ground truth and the nuanced relationships learned by the teacher. This dual-learning approach enhances the student model's performance and generalization ability.

In recent years, researchers have investigated various knowledge distillation methods for HSI classification. Notable contributions include the following: Due to the limited availability of labeled HSI samples, deep learning methods have remained underutilized. To address this issue, Yue et al. [73] proposed a self-supervised learning method with adaptive distillation that leverages abundant unlabeled samples for training. Zhao et al. [74] introduced a lifelong learning strategy to develop universal HSI classification models, continuously updating model weights through spectral–spatial feature distillation. To tackle classification problems with a limited number of samples while maximizing the use

of unlabeled data, Chi et al. [75] proposed a novel self-supervised learning method that incorporates knowledge distillation for HSI classification. Additionally, Feng et al. [76] developed a method to enhance cross-domain learning for HSI classification by treating meta-knowledge extraction and source domain debiasing as a synergistic process through decoupled knowledge distillation.

The aforementioned knowledge distillation methods for HSI classification primarily focus on limited labeled samples, domain adaptation, and universal models. Our method first employs knowledge distillation to address the issue of label noise in HSI classification. While similar to these methods in utilizing knowledge distillation to generate soft labels, our approach differs by integrating knowledge distillation within an ensemble learning framework. In this framework, the resampling strategy and out-of-bag (OOB) error estimation are used to produce more robust soft labels. Our goal is to enhance the model's robustness under label noise conditions, ensuring that the student model can still effectively distinguish categories in its final output.

3. Proposed Method

In this paper, we propose an ensemble network-based distillation (END) method for HSI classification under label noise conditions. The overall framework of the END method is depicted in Figure 1. The approach begins by applying ensemble learning to construct multiple base networks, each trained on distinct subsets of the data. These base networks are then used to estimate the labels of the training samples while accounting for label noise. By aggregating the predictions from these base networks, we generate an estimated distribution (ED) that captures the collective consensus.

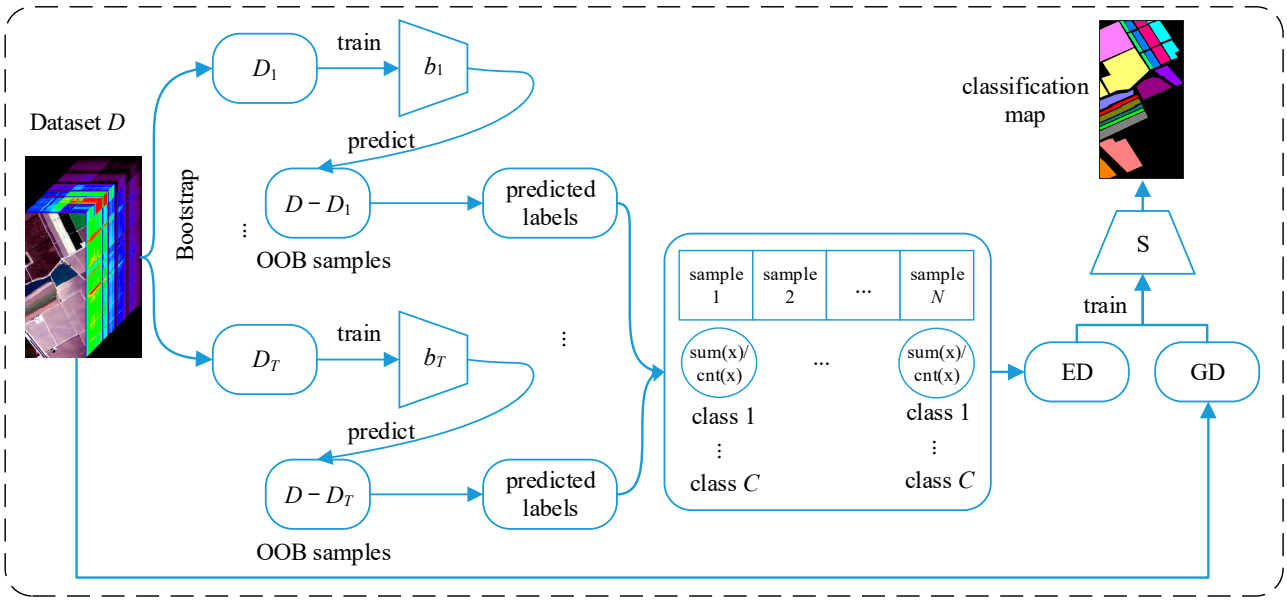


Figure 1. The framework of the proposed END method. First, *T*-based neural networks are trained on resampling datasets. Next, the estimated label distribution of the training data is computed by predicting out-of-bag (OOB) samples. Finally, the estimated distribution (ED) is combined with the ground-truth distribution (GD) to train a student network (S).

This ED is subsequently combined with the ground-truth distribution (GD) to train a student network. The student network learns from both the ED and GD, harnessing the benefits of ensemble learning while mitigating the effects of label noise. To balance the influence of the ED and GD during training student networks, we introduce a robust loss function with an adaptive parameter. This parameter dynamically adjusts the weighting between the ED and GD, enabling the student network to learn accurate representations despite the presence of noisy labels.

By integrating ensemble learning with knowledge distillation, the proposed END method not only strengthens the model's robustness against label noise but also enhances its overall classification accuracy, delivering more reliable and precise results even in challenging scenarios. Additionally, the adaptive nature of the loss function ensures the method's effectiveness across different noise levels and data distributions, making it a versatile and powerful tool for HSI classification tasks.

To ensure that the effectiveness of our method in handling label noise is not merely a consequence of using advanced base network models, we initially implemented our approach with a 2D-CNN as the base network. In subsequent experiments, we will also validate our method using a 3D network, specifically the spectral–spatial residual network (SSRN) [38].

In the subsequent subsections, we delve into a detailed exploration of the ensemble network-based distillation process and the robust loss function, providing a comprehensive understanding of their complexities and contributions to the proposed framework.

3.1. Ensemble Network-Based Distillation

Ensemble learning has demonstrated strong performance in HSI classification [27,77], with research indicating its robustness in the presence of label noise, making it particularly effective for handling noisy data [78,79]. To estimate the label distribution of the training data, we construct an ensemble network by integrating multiple deep networks, thereby leveraging their diverse strengths to improve classification accuracy and stability.

To facilitate this, we employ the bootstrap sampling method, which is well-suited for ensemble learning. Bootstrap sampling generates multiple subsets of the training data by randomly sampling with replacement. This approach fosters the creation of diverse training sets and naturally produces a set of out-of-bag (OOB) samples—data points excluded from their respective bootstrap samples. These OOB samples provide additional valuable information, as the OOB error, derived from predictions made on these samples, closely approximates the model's generalization error [80,81]. This makes OOB samples particularly useful for evaluating the ensemble network's performance and reliability.

In our method, bootstrap sampling is used to generate multiple training subsets, each assigned to a different base network within the ensemble. Once trained, the predictions on the OOB samples are used to estimate the label distribution of the training data. This OOB-based estimation offers a more reliable and accurate assessment of the label distribution, which is critical for effectively managing noisy labels in subsequent stages of our method. By combining these predictions, we construct an estimated label distribution that serves as a robust foundation for refining the classification model through the ensemble network-based distillation approach.

The following part of this subsection details the steps involved in obtaining the estimated distribution based on the ensemble network.

(1) Given the hyperspectral dataset $D = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N}$, where x_i is a sample described by spectral bands, and the label corresponding to x_i is $y_i \in \{1, \cdots, C\}$;

(2) Set 0 matrices $ED_{N\times C}$, $P_{N\times C}$, and $Q_{N\times C}$, where ED, P, and Q represents the estimated distribution, prediction matrix, and sample count matrix, respectively;

(3) Obtain T subsets $D_1, D_2, \cdots, D_T$ of dataset D by employing a bootstrap sampling strategy;

(4) Train T base neural networks $b_1, b_2, \cdots b_T$ on subsets $D_1, D_2, \cdots, D_T$;

(5) For each set $D - D_t$, every OOB sample x_j in $D - D_t$ is classified by its corresponding neural network b_i. Update p and Q through the following equations:

$$P(\mathrm{x}_j) = P(\mathrm{x}_j) + b(\mathrm{x}_j), \tag{1}$$

$$Q(\mathrm{x}_j) = Q(\mathrm{x}_j) + 1. \tag{2}$$

(6) Compute estimated distribution through the following formula:

$$ED = P./Q, \tag{3}$$

where the symbol "./" denotes element-wise division operation.

Rather than categorizing samples as simply correct or incorrect and discarding those deemed incorrect, we adopt a more nuanced strategy. Using the estimated distribution (ED), we assign probability scores to all possible classes for each sample, offering a deeper insight into the ensemble's decision-making process. By labeling each sample with a probability distribution across all classes, rather than using a binary correct/incorrect label, we capture the uncertainty and variability inherent in the ensemble's predictions. This approach ensures that no potentially valuable data are discarded, allowing the model to learn from the full spectrum of available information and enhancing its overall robustness during training.

3.2. Noise Robust Loss Function with an Adaptive Parameter

After deriving the ED distilled from the ensemble network, we combine it with the GD to train the student network. To effectively guide this training, we require a loss function that accommodates both estimated and ground-truth labels. While cross-entropy (CE) loss is widely recognized for its efficacy in model convergence, previous studies [82,83] have shown that it lacks robustness in the presence of label noise. In contrast, reverse cross-entropy (RCE) has been proven to be more resistant to noisy labels [83]. This suggests that combining CE and RCE could result in a noise-robust loss function.

Building on these findings, we design a hybrid loss function that merges CE and RCE to improve noise robustness. To ensure a balanced influence between the soft labels (derived from the ED) and the ground-truth labels, we incorporate an adaptive parameter that dynamically adjusts this balance throughout the training process. Furthermore, we apply a normalized version of CE (NCE) loss to the estimated label distribution, as normalization has been shown to further enhance robustness against label noise.

The following content will provide a detailed procedure for the design and implementation of this loss function.

For a sample x_i, its predicted label distribution from a classifier is denoted as $p(c|\mathrm{x}_i)$, and the ground-truth label distribution over observed labels is denoted as $q(c|\mathrm{x}_i)$. The CE loss for sample x_i is as follows:

$$l_{CE}(\mathrm{x}_i) = -\sum_{c=1}^{C} q(c|\mathrm{x}_i)\log p(c|\mathrm{x}_i). \tag{4}$$

The NCE is written as follows.

$$l_{NCE}(\mathrm{x}_i) = \frac{-\sum_{c=1}^{C} q(c|\mathrm{x}_i)\log p(c|\mathrm{x}_i)}{-\sum_{j=1}^{C}\sum_{c=1}^{C} q(y=j|\mathrm{x}_i)\log p(c|\mathrm{x}_i)}. \tag{5}$$

RCE is the reverse version of CE, the RCE loss is as follows:

$$l_{RCE}\ (\mathrm{x}_i) = -\sum_{c=1}^{C} p(c|\mathrm{x}_i)\log q(c|\mathrm{x}_i). \tag{6}$$

Thus, we define the sample-wise mixed loss for the student network (S) as follows:

$$\begin{array}{c} END_{MIX}(\mathrm{x}_i) = \frac{1}{2} l_{NCE}[S(\mathrm{x}_i), (1-\lambda)ED(\mathrm{x}_i) + \lambda y_i] \\ + \frac{1}{2} l_{RCE}[S(\mathrm{x}_i), (1-\lambda)ED(\mathrm{x}_i) + \lambda y_i] \end{array} \tag{7}$$

In Equation (7), $S(\mathrm{x}_i)$ represents the prediction made by the student network for the sample x_i. The parameter λ is defined as follows:

$$\lambda = \frac{\sum_{i=1}^{N} 1(y_i == arg\max_{c} ED(c|\mathrm{x}_i))}{N}, \tag{8}$$

where N is the total number of samples, y_i is the ground-truth label, $arg\max_c ED(c|\mathrm{x}_i)$ is the class label with the highest probability in the estimated distribution for the sample x_i, and the indicator function $1(\cdot)$ is a binary operator that evaluates a given condition. It returns a value of 1 if the condition is true and 0 if the condition is false. Essentially, λ represents the proportion of samples for which the predicted label from the ensemble network matches the ground-truth label.

The parameter λ plays a crucial role in determining how the student network is trained in relation to the ensemble model's confidence in a given sample. Specifically, when the ensemble model exhibits low confidence, indicated by a lower value of λ, the student network is guided predominantly by the ground-truth labels during training. This reduces the reliance on the estimated distribution, ensuring that the model aligns more closely with the true class labels. On the other hand, when the ensemble model shows high confidence in a sample, as indicated by a higher value of λ, the estimated distribution takes precedence in the training process, even if it contradicts the ground-truth label. This strategy allows the student model to focus more on the information provided by the ensemble model's probability estimates, which is believed to reflect a more reliable and confident prediction. In this way, the system dynamically adjusts its reliance on either the ground-truth labels or the estimated distribution, depending on the confidence level of the ensemble model.

The above analysis provides key insights into the END method, and Algorithm 1 presents the pseudocodes of the END method.

Algorithm 1 Ensemble Network-Based Distillation with Robust Loss Function

Input: Ensemble size : T; Training dataset $D = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N}$.
Initialization: $ED_{N\times C} = 0$, $P_{N\times C} = 0$, $Q_{N\times C} = 0$.
$GD = groundtruth(D)$. //obtain the ground-truth distribution
For t = 1 to T **do**
  $D_t = Resampling(D)$. //bootstrap resampling
  $b_t = Train(D_t)$. //train base network
  For $\mathrm{x}_j \in D - D_t$ **do**
    $P\left(\mathrm{x}_j\right) = P\left(\mathrm{x}_j\right) + b\left(\mathrm{x}_j\right)$. // the cumulative prediction results for x_j
    $Q\left(\mathrm{x}_j\right) = Q\left(\mathrm{x}_j\right) + 1$. //frequency count for x_j
  End
End
$ED = P./Q$. //compute the estimated distribution
$S = Train(D, GD + ED, END_{MIX})$. //train the student network S through (7)
Output: The student network S.

4. Experimental Results and Analysis

4.1. Hyperspectral Datasets and Experimental Settings

We evaluated the proposed END method against several state-of-the-art approaches using the Salinas Valley, Houston, and Pavia University datasets to demonstrate its effectiveness. A detailed description of these datasets is provided below.

(1) Salinas Valley (SV) dataset: Collected in 1998 over Salinas Valley, CA, USA, this dataset consists of 512 × 217 pixels and contains 224 spectral bands, covering a wavelength range from 400 to 2500 nm. After removing 20 water absorption bands, 204 spectral bands remain. With a spatial resolution of 3.7 m, the dataset covers 16 land cover classes;

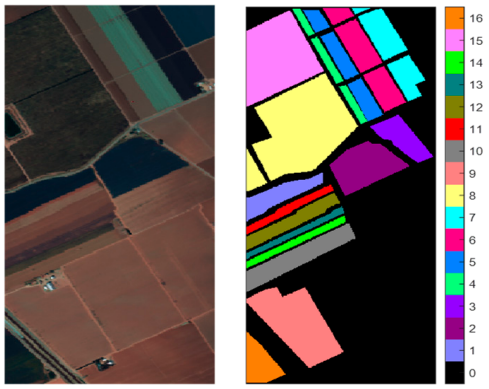
(2) Houston (HOU) dataset: Acquired by the ITRES CASI-1500 sensor, this dataset was part of the 2013 IEEE GRSS Data Fusion Competition. It consists of 349 × 1905 pixels and includes 144 spectral bands, spanning wavelengths from 364 to 1046 nm. The HOU dataset has a spatial resolution of 2.5 m and covers 15 land cover classes, providing a comprehensive representation of urban and suburban landscapes;

(3) Pavia University (PU) dataset: Acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over the University of Pavia in northern Italy, this dataset contains an image matrix of 610 pixels in width and 340 pixels in height. It includes 103 spectral bands, covering a wavelength range from 430 to 860 nm.

Table 1 provides a comprehensive overview of the class names along with the corresponding number of labeled samples for each of the three datasets. Figures 2–4 illustrate the false-color images of these HSI datasets along with their respective reference maps.

Table 1. Class names along with the number of labeled samples for three hyperspectral datasets.

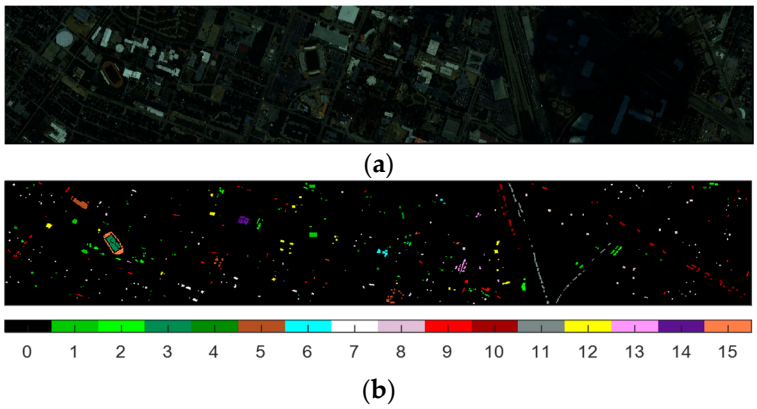
No	SV		HOU		PU	
	Class Name	**Samples**	**Class Name**	**Samples**	**Class Name**	**Samples**
1	Brocoli_green_weeds_1	2009	Healthy grass	1251	Asphalt	6631
2	Brocoli_green_weeds_2	3726	Stressed grass	1254	Meadows	18,649
3	Fallow	1976	Synthetic grass	697	Gravel	2099
4	Fallow_rough_plow	1394	Trees	1244	Trees	3064
5	Fallow_smooth	2678	Soil	1242	Painted metal sheets	1345
6	Stubble	3959	Water	325	Bare Soil	5029
7	Celery	3579	Residential	1268	Bitumen	1330
8	Grapes_untrained	11,271	Commercial	1244	Self-Blocking Bricks	3682
9	Soil_vinyard_develop	6203	Road	1252	Shadows	947
10	Corn_senesced_green_weeds	3278	Highway	1227		
11	Lettuce_romaine_4wk	1068	Railway	1235		
12	Lettuce_romaine_5wk	1927	Parking Lot 1	1233		
13	Lettuce_romaine_6wk	916	Parking Lot 2	469		
14	Lettuce_romaine_7wk	1070	Tennis Court	428		
15	Vinyard_untrained	7268	Running Track	660		
16	Vinyard_vertical_trellis	1807				



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

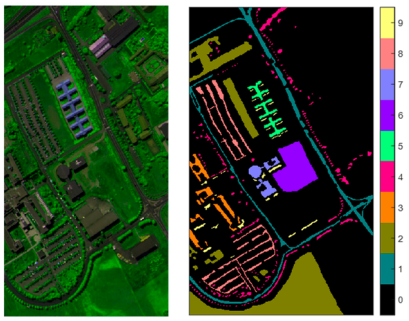
Figure 2. False-color image and reference map of the SV. (**a**) False-color image created from a combination of three spectral bands. (**b**) Reference map detailing the ground truth classification.

To ensure a rigorous quantitative evaluation of our experimental results, we employed three key metrics: overall accuracy (OA), average accuracy (AA), and the Kappa coefficient (κ). Each metric serves as a crucial indicator to assess the performance and accuracy of our classification outcomes. Label noise was simulated using a uniform distribution, where randomly selected samples were reassigned to different class labels with a random probability. To assess the effect of varying noise levels on classification performance, we conducted experiments using training data with multiple levels of label noise.



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

Figure 3. False-color image and reference map of the HOU. (**a**) False-color image created from a combination of three spectral bands. (**b**) Reference map detailing the ground truth classification.



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

Figure 4. False-color image and reference map of the PU. (**a**) False-color image created from a combination of three spectral bands. (**b**) Reference map detailing the ground truth classification.

Our proposed END method was built on a 2D CNN as the base network. The training process was optimized using the Adam optimizer with a learning rate of 0.001. The base networks in the ensemble were trained for 30 epochs, while the student network was trained for 100 epochs. The implementation was performed using PyTorch, leveraging an NVIDIA RTX 3070 GPU equipped with CUDA 11, which significantly accelerated the training process. Following preliminary experiments, the ensemble size was set to 30 for optimal performance.

4.2. Comparison with the State-of-the-Art Methods

We conducted a comparative evaluation of our proposed END method against several state-of-the-art approaches, including MSSG [10], DCRN [13], AAN [15], TCRL [16], SLS [59], and MLN [60]. Additionally, we compared the proposed method with the base network 2D-CNN and a traditional ensemble method, random forests (RF) [80]. For consistency, we adhered to the parameter settings specified in their respective publications. In the case of MSSG, the segmentation scale parameter was optimized to 0, replicating the best-performing configuration from the original experiments. To ensure standardization of the training data, we randomly sampled 50 instances from each class and ran each method ten times to obtain average performance metrics. With a fixed noise rate (r) of 0.3, each algorithm was executed 10 times, and the average results were used for comparison. Tables 2–4 provide a detailed analysis of the classification performance on the SV, HOU, and PU datasets. Figures 5–7 display the classification maps for the three HSI datasets after a single run of each method.

Table 2. Classification accuracy (%) achieved by comparison methods on the SV dataset at a noise level of 0.3.

Class	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
1	97.86	98.59	99.48	99.67	99.45	99.48	99.25	96.87	99.75
2	99.73	99.59	99.81	99.69	99.77	99.73	99.80	99.68	99.80
3	98.29	92.35	99.47	98.93	96.28	98.32	99.42	97.31	97.52
4	99.63	96.43	96.90	99.19	98.90	99.62	99.52	93.60	98.92
5	97.77	98.67	94.92	99.17	99.07	97.39	97.94	95.54	99.14
6	97.08	99.77	99.00	99.65	99.59	99.66	99.78	97.33	99.72
7	99.09	99.21	99.46	99.58	99.58	98.72	99.21	98.84	99.66
8	61.02	89.16	89.06	84.86	86.87	77.29	80.72	65.52	93.27
9	98.34	99.72	99.61	99.46	99.39	98.52	98.87	97.18	99.72
10	87.87	90.15	96.01	93.25	94.10	89.06	90.52	85.15	94.87
11	93.81	93.73	95.97	98.10	95.81	94.43	97.19	88.81	98.13
12	98.98	99.58	99.89	99.94	99.94	97.31	99.46	98.05	99.94
13	99.54	97.95	98.76	98.09	98.57	98.11	98.80	97.83	99.30
14	93.50	97.25	92.41	93.65	93.91	94.40	97.53	91.54	97.28
15	80.16	69.39	72.34	76.87	77.05	70.58	81.75	75.69	95.56
16	95.56	95.11	97.00	97.25	98.30	96.83	98.07	91.33	98.62
OA	87.29	92.54	94.33	95.14	95.77	89.71	92.98	86.66	97.28
AA	93.64	94.79	95.63	96.10	96.04	94.34	96.11	91.89	98.20
$\kappa \times 100$	85.92	91.73	93.77	94.66	95.36	88.55	92.22	85.19	96.97

Table 3. Classification accuracy (%) achieved by comparison methods on the HOU dataset at a noise level of 0.3.

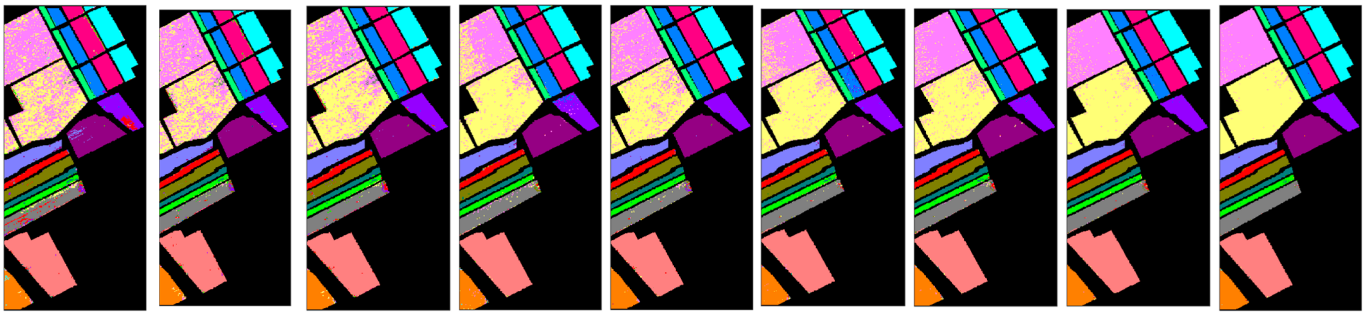
Class	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
1	82.79	94.08	94.48	85.93	92.33	90.30	96.88	84.37	95.36
2	95.44	97.85	98.09	96.57	98.56	97.70	95.08	98.06	98.33
3	95.13	99.99	100	100	100	95.35	96.08	97.78	100
4	94.34	97.43	98.07	99.20	96.30	93.64	93.78	88.24	97.75
5	97.37	99.44	99.19	99.28	97.58	96.03	95.54	95.17	98.07
6	84.11	98.15	97.23	95.38	94.77	93.84	95.74	95.08	99.69
7	79.29	89.91	88.80	88.01	78.94	79.12	84.30	75.72	81.70
8	57.79	62.06	68.65	74.20	76.21	73.95	75.52	54.82	78.38
9	65.36	71.73	77.96	81.07	83.71	79.44	68.05	61.93	87.14
10	74.54	60.88	65.12	82.07	82.31	72.80	72.49	73.27	87.45
11	72.44	67.45	71.34	72.63	82.75	74.14	73.07	65.98	84.37
12	59.42	59.69	66.10	78.51	74.86	63.14	73.48	61.76	82.40
13	48.38	59.28	62.05	91.47	97.23	50.49	65.79	42.82	95.10
14	97.53	97.43	99.30	99.30	99.77	95.44	98.48	95.59	99.77
15	97.16	98.94	98.94	98.94	99.24	95.51	93.11	90.63	99.55
OA	79.36	82.11	84.48	87.78	88.47	83.02	84.24	77.61	90.78
AA	80.07	83.62	85.69	89.50	90.30	83.39	85.16	78.77	92.34
$\kappa \times 100$	77.69	80.65	83.22	86.80	87.54	81.64	82.96	75.80	90.04

Table 4. Classification accuracy (%) achieved by comparison methods on the PU dataset at a noise level of 0.3.

Class	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
1	80.17	90.91	92.20	92.84	93.17	86.83	87.10	80.56	93.45
2	82.78	87.55	98.17	98.76	99.00	88.13	90.94	75.52	99.23
3	83.54	90.71	95.00	95.05	96.14	89.92	89.48	80.12	97.05
4	93.28	97.68	96.08	95.92	96.15	96.54	97.76	95.03	97.39
5	99.59	99.48	99.41	99.55	99.70	98.08	98.66	99.01	99.48
6	83.12	89.24	82.00	82.60	87.79	86.52	87.10	79.09	92.07

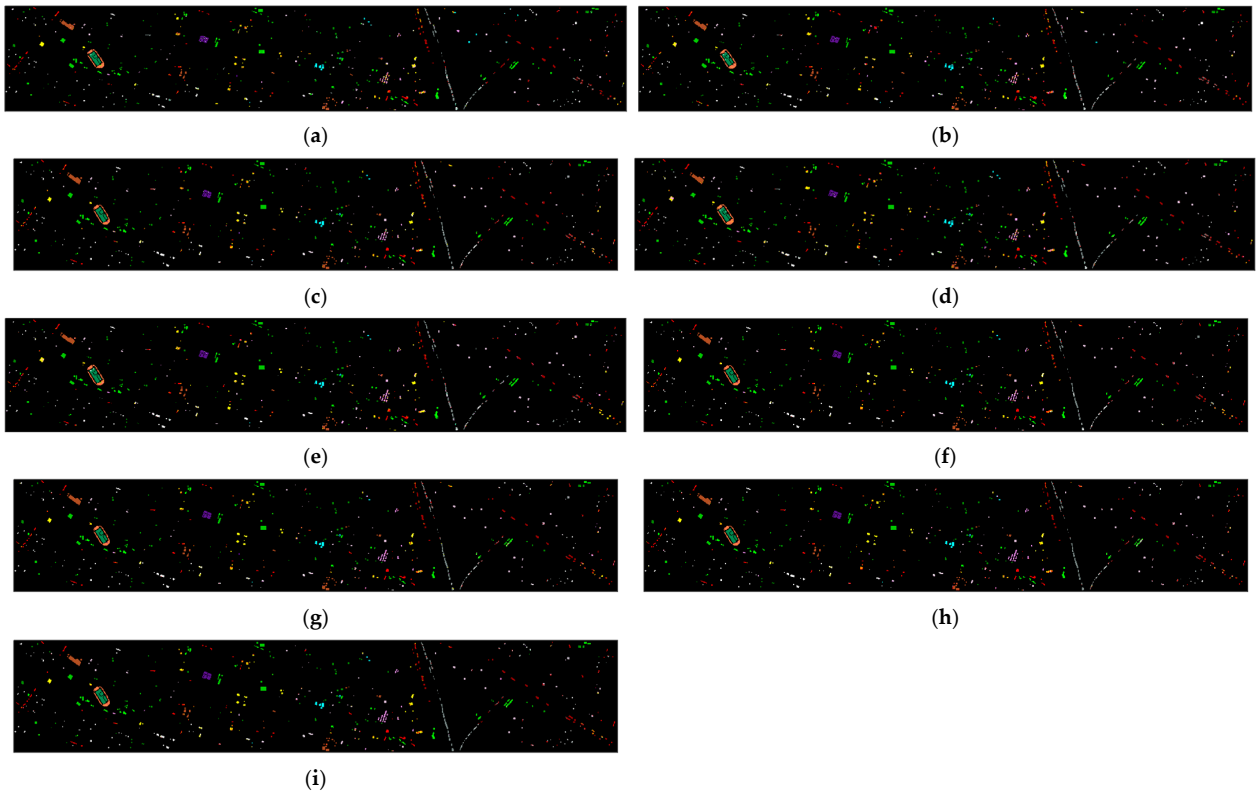
Table 4. *Cont.*

Class	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
7	91.24	95.71	94.36	95.71	96.24	91.23	96.15	83.59	97.89
8	84.40	80.61	89.79	93.64	93.89	78.20	86.21	81.00	95.65
9	99.95	100	100	100	99.79	99.77	99.19	99.78	100
OA	84.22	89.64	94.28	95.07	95.95	88.16	90.36	80.25	96.93
AA	88.68	92.43	94.11	94.90	95.76	90.57	92.51	85.97	96.91
$\kappa \times 100$	79.13	86.30	92.38	93.43	94.61	83.99	86.35	74.67	95.92



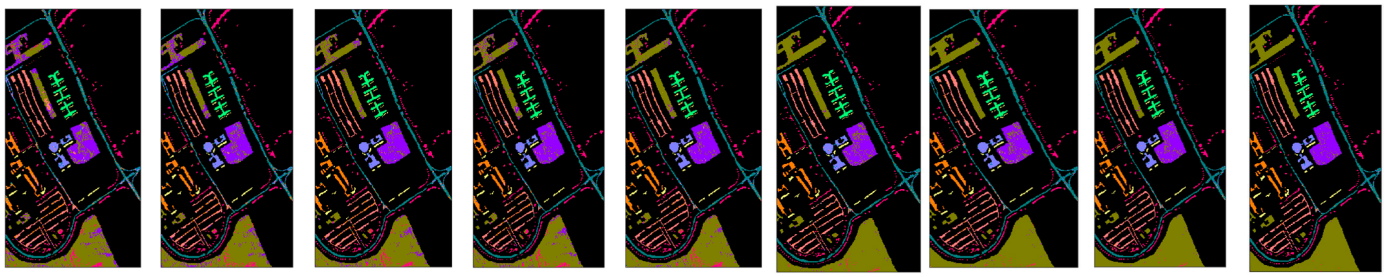
(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**)

Figure 5. Classification maps generated for the SV image utilizing various comparison methods. (**a**) RF: OA = 86.66. (**b**) 2D-CNN: OA = 87.29%. (**c**) SLS: OA = 89.71%. (**d**) MSSG: OA = 92.53%. (**e**) MLN: OA = 92.98%. (**f**) DCRN: OA = 94.36%. (**g**) AAN: OA = 95.14%. (**h**) TCRL: OA = 95.78%. (**i**) END: OA = 97.30%.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**)

Figure 6. Classification maps generated for the HOU image utilizing various comparison methods. (**a**) RF: OA = 77.61%. (**b**) 2D-CNN: OA = 79.36%. (**c**) MSSG: OA = 82.12%. (**d**) SLS: OA = 83.02%. (**e**) MLN: OA = 84.24%. (**f**) DCRN: OA = 84.47%. (**g**) AAN: OA = 87.79%. (**h**) TCRL: OA = 88.47%. (**i**) END: OA = 90.77%.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**)

Figure 7. Classification maps generated for the PU image utilizing various comparison methods. (**a**) RF: OA = 80.26%. (**b**) 2D-CNN: OA = 84.21%. (**c**) SLS: OA = 88.16%. (**d**) MSSG: OA = 89.65%. (**e**) MLN: OA = 90.36%. (**f**) DCRN: OA = 94.26%. (**g**) AAN: OA = 95.07%. (**h**) TCRL: OA = 95.98%. (**i**) END: OA = 96.94%.

Several key insights can be drawn from Tables 2–4 and Figures 5–7. First, the OA and Kappa coefficients of these methods on the SV and PU datasets demonstrate a clear ranking: END, TCRL, AAN, DCRN, MLN, MSSG, SLS, 2D-CNN, and RF. In contrast, for the HOU dataset, the only difference is that the OA and Kappa coefficients of MSSG are lower than those of SLS. The AA ordering of DCRN, TCRL, and MLN is inconsistent across the three datasets in the comparison methods. However, the AA ordering for the other methods is consistent and follows this sequence: END, AAN, MSSG, SLS, 2D-CNN, and RF. The above ranking underscores the END method's superior ability to manage label noise effectively while achieving high classification accuracy. Second, the END method consistently delivers the highest class-specific accuracy across a broader range of classes compared to the other methods, further affirming its robustness. Notably, it performs exceptionally well in challenging scenarios. For example, on the SV dataset, the END method significantly improves accuracy for class 8 by approximately 4% and for class 15 by about 14% compared to the second-best performing method. Finally, the visual results presented in Figures 5–7 further validate the effectiveness of the END method. The classification maps generated by the END approach offer more precise and detailed representations of the underlying land cover classes compared to those produced by other methods, highlighting its superior performance.

4.3. Classification Performance Across Varying Noise Levels

In this section, we delve into experiments that varied the level of label noise, with the noise rate (r) systematically increasing from 0.1 to 0.5 in steps of 0.1, while maintaining consistency in all other parameters as established in the previous section. Tables 5–7 comprehensively present the impact of these varying noise levels on the performance of the comparison methods, reporting the OA, AA, and Kappa coefficients for the SV, HOU, and PU datasets, respectively.

Table 5. Classification accuracy (in %) obtained by comparison methods with different noise rates (r) when applied to the SV dataset.

r	Metric	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
0.1	OA	88.56	92.97	94.68	95.94	96.32	90.54	93.75	87.93	97.94
	AA	94.68	95.84	95.92	96.88	96.97	95.23	96.64	92.86	98.41
	$\kappa \times 100$	86.89	92.46	93.96	94.83	95.82	89.15	92.78	86.63	97.76
0.2	OA	87.82	92.83	94.55	95.71	96.11	90.12	93.23	87.12	97.66
	AA	94.11	95.55	95.77	96.73	96.88	94.84	96.35	92.35	98.32
	$\kappa \times 100$	86.33	92.21	93.88	94.75	95.47	88.89	92.45	85.75	97.39
0.3	OA	87.29	92.54	94.33	95.14	95.77	89.71	92.98	86.66	97.28
	AA	93.64	94.79	95.63	96.10	96.04	94.34	96.11	91.89	98.20
	$\kappa \times 100$	85.92	91.73	93.77	94.66	95.36	88.55	92.22	85.19	96.97

Table 5. *Cont.*

r	Metric	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
0.4	OA	86.54	91.83	93.84	94.58	94.91	89.05	92.22	85.87	96.83
	AA	93.15	93.95	95.06	95.75	95.82	93.85	95.77	91.24	97.25
	$\kappa \times 100$	85.06	90.96	92.56	92.51	93.64	88.06	91.55	84.66	96.07
0.5	OA	85.62	91.63	93.12	93.64	93.85	88.12	91.18	85.08	96.38
	AA	92.86	93.76	94.22	94.66	94.87	93.20	95.13	90.46	96.88
	$\kappa \times 100$	83.82	90.84	91.89	91.84	92.97	87.54	91.04	83.33	95.72

Table 6. Classification accuracy (in %) obtained by comparison methods with different noise rates (*r*) when applied to the HOU dataset.

r	Metric	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
0.1	OA	81.03	83.67	86.14	88.46	89.57	84.12	85.26	78.93	91.38
	AA	80.96	84.78	87.04	90.31	90.94	84.26	86.15	79.67	93.45
	$\kappa \times 100$	78.65	84.15	85.62	87.96	88.91	83.64	84.46	76.86	91.02
0.2	OA	80.15	82.98	85.36	88.13	89.06	83.68	84.87	78.34	91.06
	AA	80.45	84.11	86.29	90.08	90.76	83.87	85.64	79.25	93.04
	$\kappa \times 100$	78.14	83.97	84.67	87.84	88.46	82.78	83.76	76.42	90.85
0.3	OA	79.36	82.11	84.48	87.78	88.47	83.02	84.24	77.61	90.78
	AA	80.07	83.62	85.69	89.50	90.30	83.39	85.16	78.77	92.34
	$\kappa \times 100$	77.69	80.65	83.22	86.80	87.54	81.64	82.96	75.80	90.04
0.4	OA	78.12	81.26	83.68	87.12	87.75	81.96	83.42	76.45	90.23
	AA	79.33	82.62	84.57	88.93	89.34	82.45	84.46	78.02	91.35
	$\kappa \times 100$	76.87	79.95	82.76	86.18	86.89	80.35	82.03	74.96	89.58
0.5	OA	76.84	80.53	82.47	86.28	86.82	80.84	82.26	74.82	89.48
	AA	78.12	81.74	83.38	87.84	87.41	81.65	83.10	76.98	90.81
	$\kappa \times 100$	75.72	79.14	81.97	85.49	85.67	79.26	80.87	73.71	88.64

Table 7. Classification accuracy (in %) obtained by comparison methods with different noise rates (*r*) when applied to the PU dataset.

r	Metric	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
0.1	OA	85.26	90.82	94.93	95.58	96.47	90.04	91.45	81.15	97.26
	AA	89.21	93.66	94.71	95.29	96.13	91.58	93.70	86.62	97.18
	$\kappa \times 100$	80.04	88.06	92.89	93.97	95.28	85.11	87.72	75.28	96.21
0.2	OA	84.81	90.35	94.74	95.42	96.32	88.65	90.97	80.76	97.11
	AA	88.95	93.12	94.58	95.18	95.98	91.15	93.24	86.33	97.06
	$\kappa \times 100$	79.65	87.22	92.76	93.83	94.97	84.66	86.94	75.05	96.08
0.3	OA	84.22	89.64	94.28	95.07	95.95	88.16	90.36	80.25	96.93
	AA	88.68	92.43	94.11	94.90	95.76	90.57	92.51	85.97	96.91
	$\kappa \times 100$	79.13	86.30	92.38	93.43	94.61	83.99	86.05	74.67	95.92
0.4	OA	83.41	88.67	93.41	94.36	95.14	87.25	89.52	79.65	96.52
	AA	88.12	91.53	93.34	94.21	94.82	89.62	91.62	85.42	96.57
	$\kappa \times 100$	78.61	85.54	91.66	92.74	94.02	83.14	85.23	73.95	95.57
0.5	OA	82.25	87.26	92.33	93.17	94.06	85.93	88.25	78.76	96.02
	AA	87.26	87.55	92.44	93.28	93.95	88.45	90.48	84.69	95.97
	$\kappa \times 100$	77.57	84.48	90.63	91.25	93.07	81.79	83.96	73.14	95.03

Several key observations can be drawn from Tables 5–7. First, across all three HSI datasets, there is a noticeable decline in OA, AA, and Kappa coefficients as the noise rate increases, underscoring the significant challenge that label noise presents in HSI classification. Second, the proposed END method consistently outperforms other approaches

in terms of classification accuracy, regardless of the noise level. This demonstrates the robustness of the END method in effectively managing label noise compared to alternative techniques. Lastly, the END method exhibits low sensitivity to increasing noise levels, with only minimal reductions in performance observed. This adaptability highlights its effectiveness in handling various noise conditions, making it particularly well-suited for real-world HSI classification tasks, where label noise is often unavoidable. Overall, the robustness and flexibility of the END method position it as a highly reliable solution for classification tasks under noisy conditions.

4.4. Evaluation of the Loss Function

To assess the effectiveness of the proposed loss function, we conducted comparisons with CE loss, NCE loss, and RCE loss. The loss functions for the proposed method can be written as follows:

$$END_{CE}(\mathrm{x}_i) = l_{CE}[S(\mathrm{x}_i), (1-\lambda)ED(\mathrm{x}_i) + \lambda y_i], \tag{9}$$

$$END_{NCE}(\mathrm{x}_i) = l_{NCE}[S(\mathrm{x}_i), (1-\lambda)ED(\mathrm{x}_i) + \lambda y_i], \tag{10}$$

$$END_{RCE}(\mathrm{x}_i) = l_{RCE}[S(\mathrm{x}_i), (1-\lambda)ED(\mathrm{x}_i) + \lambda y_i], \tag{11}$$

where the parameter λ is set the same as in (8), acting as an adaptive parameter, and $S(\mathrm{x}_i)$ represents the prediction of the student network.

In this experiment, we formed training sets of various sizes by randomly selecting 10 to 80 samples for each class, while maintaining a fixed noise rate r = 0.3. Figure 8 illustrates the OA curves of the END method with different versions of loss functions plotted against the number of training samples.

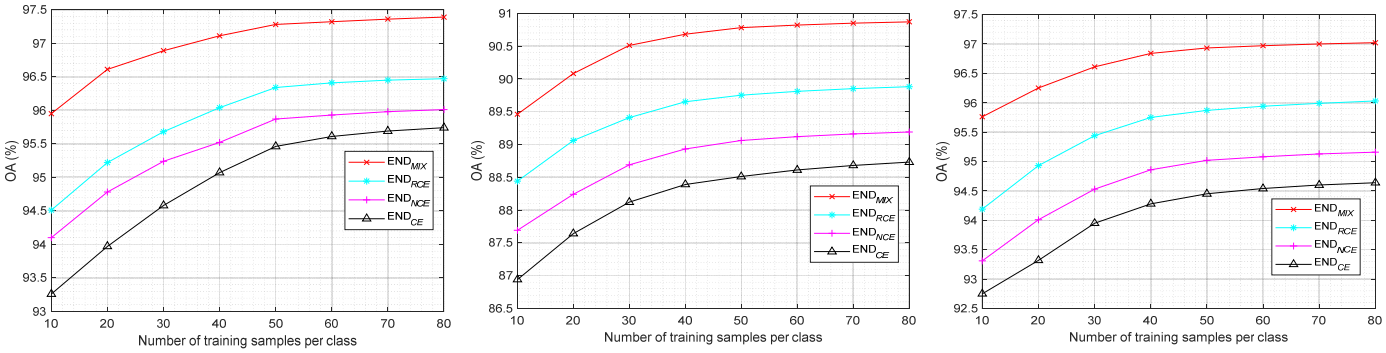


Figure 8. OA curves of the END method with different versions of loss functions plotted against the number of training samples. (**a**) SV; (**b**) HOU; (**c**) HOU.

The results in Figure 8 indicate, first, that the number of training samples has a notable influence on classification accuracy regardless of the chosen loss function. Specifically, under a fixed noise rate (r = 0.3), the OA tends to increase as the number of training samples grows. Second, the END_{MIX} loss significantly improves the model's classification accuracy in the presence of label noise. This effectiveness can be attributed to two factors. On the one hand, the designed loss function simultaneously leverages the complementary advantages of cross-entropy and reverse cross-entropy, allowing it to fit both noisy and clean data more effectively. On the other hand, the adaptive parameter λ effectively balances the roles of ED and GD in the target network during training. This adaptability addresses varying levels of noise, enabling the model to achieve strong classification performance across different noise scenarios.

4.5. Ablation Study

To fully investigate the significance of each component in the proposed method, we removed various elements of the END method for our experiments. Detailed information is provided in Table 8. In this section, we conducted experiments using a 2D-CNN as the base network as well as a 3D network, specifically the spectral–spatial residual network (SSRN) [38]. The parameter settings for the SSRN adhere to those outlined in the original paper, while the parameters for the other methods remain consistent with our previous experiments. Table 9 presents the OA of three methods applied to the SV, HOU, and PU datasets.

Table 8. Description of the END method and its streamlined versions.

Method	Ensemble	Distillation	Description
Single	✗	✗	Train a single network model and use it for classification
Bagging	✓	✗	Train an ensemble network model and use it for classification
END	✓	✓	The trained ensemble network is used to guide the training of the target network

Table 9. Classification accuracy (OA, %) of the three methods on the SV, HOU, and PU datasets.

Method	Base Network	SV	HOU	PU
Single	2D-CNN	87.29	79.36	84.22
	SSRN	89.75	84.47	87.61
Bagging	2D-CNN	90.35	85.74	89.74
	SSRN	92.42	88.59	92.58
END	2D-CNN	97.28	90.78	96.93
	SSRN	98.16	92.33	97.85

Table 9 presents several important findings. First, SSRN consistently outperforms 2D-CNN in classification accuracy across all three datasets, due to the enhanced feature extraction capabilities of the 3D network. Second, while Bagging does improve classification accuracy in the presence of label noise, the degree of improvement remains relatively modest. Finally, the END method demonstrates the most substantial enhancement in accuracy, primarily because it utilizes knowledge distillation from the ensemble network. In other words, END employs ensemble learning not directly for classification, but to estimate a label distribution, which is then combined with the ground-truth distribution to effectively train the classification network.

4.6. Analysis of the Ensemble Size

In the previous experiment, we used an ensemble size of 30. To explore the influence of ensemble size on classification performance and determine the optimal size, we conducted experiments with ensemble sizes ranging from 5 to 50 in increments of 5. Figure 9 illustrates the relationship between classification accuracy and ensemble size.

The results in Figure 9 indicate that classification accuracy increases rapidly with the rise in ensemble size before stabilizing. As shown in Figure 9, setting the ensemble size to around 20 or more yields satisfactory classification results. This suggests that the proposed method does not require constructing a large number of base networks, which effectively reduces computational costs.

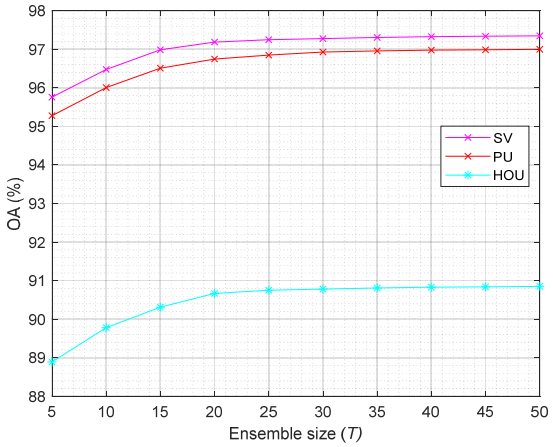


Figure 9. OA curves of SV, HOU, and PU datasets under different ensemble sizes.

4.7. Analysis of the Training Epoch for Base Network

To obtain a relatively accurate estimated distribution distilled from the ensemble, it is crucial to determine an appropriate epoch for terminating the training of the base networks early, in order to prevent overfitting to noisy samples. In our experiment, we set the maximum epoch to 100. Starting from the 10th epoch, at the end of each subsequent epoch, we use the resulting ED and ground-truth distribution to train the student network. The classification accuracy is shown in Figure 10.

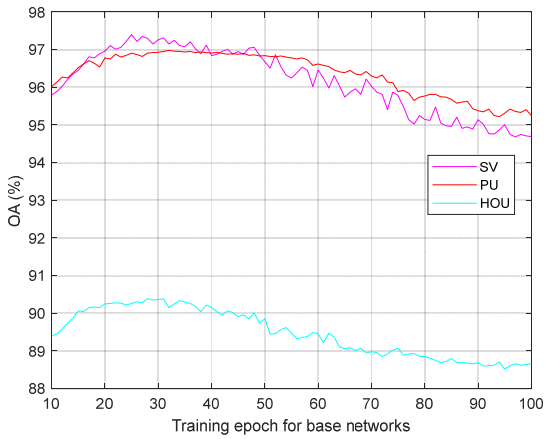


Figure 10. Classification accuracy curves of SV, HOU, and PU datasets with the increase in training epochs.

As shown in Figure 10, the OA exhibits oscillations over time. Generally, the OA initially increases, stabilizes, and then decreases as the number of epochs progresses. This trend can be attributed to the model's ability to learn more effectively from "clean" samples in the early stages, leading to more accurate estimated distributions. However, as the model begins to incorporate noisy data, convergence becomes more difficult, resulting in less accurate estimated distributions. Therefore, it is recommended to terminate the training early, ideally between 20 and 50 epochs, to prevent overfitting to noisy samples.

4.8. Experimental Results on Toulouse Dataset

To evaluate the robustness to label noise on a large dataset, we conducted experiments using the Toulouse HSI dataset [84]. This dataset was captured by an AisaFENIX 1K camera over the city of Toulouse, France. It contains approximately 380,000 labeled pixels and includes 310 spectral bands, spanning wavelengths from 0.4 to 2.5 µm. The Toulouse dataset has a spatial resolution of 1 m and covers 32 land cover classes. The ground truth of the Toulouse data consists of eight spatially disjoint splits.

We adopted the original data division in [84] for our experiments, and the experimental settings were consistent with previous studies. Table 10 reports the classification accuracy of the comparison methods on the Toulouse dataset. The results presented in Table 10 demonstrate that our proposed method achieves the highest classification accuracy among all comparison methods. Furthermore, it outperforms the second-place method by approximately 2 percent, indicating that our approach is effective for disjoint HSI classification with noisy labels.

Table 10. Classification accuracy (%) achieved by comparison methods on the Toulouse dataset at a noise level of 0.3.

Method	2D-CNN	MSSG [10]	DCRN [13]	AAN [15]	TCRL [16]	SLS [59]	MLN [60]	RF [80]	END
OA	73.48	76.68	79.25	81.45	82.95	77.85	78.64	71.35	85.64
AA	74.11	77.32	80.06	82.63	83.47	77.96	79.22	71.98	86.23
$\kappa \times 100$	71.32	75.43	78.24	80.55	81.68	75.98	77.13	70.06	83.82

5. Discussion

Some important findings from previous experimental results warrant further discussion.

The experimental results in Section 4.2 indicate that the END method outperforms other approaches, particularly when classifying challenging classes. When the labels of these challenging classes are contaminated by noise, it becomes increasingly difficult to accurately classify such samples. The superior performance of the END method in handling noisy labels suggests that its ensemble-based approach effectively mitigates the impact of label noise. In future research, it would be valuable to focus on improving classification accuracy specifically for these challenging classes. Designing methods that better address noisy labels in difficult classes could significantly enhance overall performance in HSI classification tasks.

The experimental results in Section 4.3 demonstrate that other approaches, compared to the END method, are more sensitive to high levels of label noise. The proposed END method addresses this issue by using ensemble learning and knowledge distillation to estimate the label distribution and an adaptive loss function to balance the estimated distribution with the ground-truth distribution. This allows the END method to handle high levels of noise more robustly. Future research should focus on further exploring techniques and methods to enhance performance in high-noise environments.

As shown in Section 4.4, classification accuracy drops significantly across all loss functions when fewer than 40 labeled samples per class are available. This indicates that, under conditions of limited labeled samples, label noise has a more pronounced negative impact on HSI classification. The primary challenge lies in the difficulty of extracting the true label distribution when there are already too few labeled samples. Future research should address both the scarcity of labeled samples and the presence of label noise. For instance, exploring unsupervised methods to infer the underlying label distribution could prove beneficial. This approach could leverage unlabeled data to enhance the model's robustness in noisy environments, improving overall classification performance.

The ablation study in Section 4.5 demonstrates that while the direct performance improvement from ensemble learning alone is limited, the knowledge distilled from the ensemble model can provide valuable guidance for training the target network. This knowledge distillation approach enhances the model's robustness, especially in noisy label scenarios, by effectively transferring ensemble-derived insights. Future research can further explore techniques to maximize the effectiveness of ensemble learning in distillation, aiming to distill richer and more targeted knowledge that can boost classification accuracy and generalization in various contexts.

6. Conclusions

In this paper, we introduce an ensemble network-based distillation method for HSI classification. The proposed END method leverages ensemble learning to effectively estimate the distribution of training data. Additionally, a novel loss function, more robust to label noise, is designed for training the target network. Experimental results demonstrate that the proposed method can effectively alleviate the influence of label noise, even at high noise rates. Several key conclusions can be drawn from the experimental results and analysis:

- While ensemble learning offers limited performance gains for HSI classification with noisy labels, the knowledge distilled from the ensemble model can serve as valuable guidance for training the target network;
- The predictions for OOB samples generated through resampling effectively estimate the label distribution of the training data. By building multiple base networks and aggregating their outputs for OOB samples, this method improves noise robustness and provides a refined label distribution for subsequent learning;
- The tailored loss function enhances the model's resilience against label noise, further bolstering its robustness. By dynamically balancing the importance of the estimated label distribution and the ground-truth label distribution, the loss function enables the student network to learn accurate representations.

In future research, several avenues can be explored to further refine the END method. First, incorporating resampling strategies, such as boosting, could help mitigate the impact of noisy labels by focusing on difficult-to-classify samples. Moreover, future studies could explore applying advanced deep learning architectures, such as transformers or graph neural networks, in conjunction with the proposed method. This would open new possibilities for improving the model's adaptability and scalability in various HSI classification tasks.

Author Contributions: Conceptualization, R.D.; methodology, Y.Z. and H.S.; software, J.L. and Q.Y.; validation, R.D. and J.L.; formal analysis, Q.Y. and X.L.; writing—original draft preparation, Y.Z.; writing—review and editing, R.D. and X.L.; visualization, H.S.; supervision, Y.Z. and X.L.; funding acquisition, Y.Z. and G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the National Natural Science Foundation of China under grants 62201282 and 62203231, in part by the Open Project of Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems under grant TJKL-AITS-20241003, in part by the Natural Science Foundation of Jiangsu Province under grant BK20231456, and in part by the Basic Research Project of the Natural Science Foundation of the Jiangsu Higher Education Institutions under grants 22KJB510037 and 22KJB520006.

Data Availability Statement: The Salinas Valley and Pavia University datasets used in this study are accessible at [https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes] (accessed on 10 June 2023). The Houston dataset is available at [https://hyperspectral.ee.uh.edu/?page_id=459] (accessed on 8 May 2024). The Toulouse dataset is available at [https://www.toulouse-hyperspectral-data-set.com/] (accessed on 29 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lu, B.; Dao, P.D.; Liu, J.; He, Y.; Shang, J. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sens.* **2020**, *12*, 2659. [CrossRef]
2. Anderson, J.E.; Plourde, L.C.; Martin, M.E.; Braswell, B.H.; Smith, M.L.; Dubayah, R.O.; Hofton, M.A.; Blair, J.B. Integrating waveform lidar with hyperspectral imagery for inventory of a northern temperate forest. *Remote Sens. Environ.* **2008**, *112*, 1856–1870. [CrossRef]
3. Heiden, U.; Heldens, W.; Roessner, S.; Segl, K.; Esch, T.; Mueller, A. Urban structure type characterization using hyperspectral remote sensing and height information. *Landsc. Urban Plann.* **2012**, *10*, 361–375. [CrossRef]

4. Yuan, J.; Wang, S.; Wu, C.; Xu, Y. Fine-grained classification of urban functional zones and landscape pattern analysis using hyperspectral satellite imagery: A case study of Wuhan. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 3972–3991. [CrossRef]
5. Rajamani, T.; Sevugan, P.; Ragupathi, S. Automatic building footprint extraction and road detection from hyperspectral imagery. *J. Electron. Imaging.* **2023**, *32*, 011005. [CrossRef]
6. Ahmad, M.; Shabbir, S.; Roy, S.K.; Hong, D.; Wu, X.; Yao, J.; Khan, A.M.; Mazzara, M.; Distefano, S.; Chanussot, J. Hyperspectral image classification—Traditional to deep models: A survey for future prospects. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 968–999. [CrossRef]
7. Jiang, J.; Ma, J.; Wang, Z.; Chen, C.; Liu, X. Hyperspectral image classification in the presence of noisy labels. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 851–865. [CrossRef]
8. Tu, B.; Zhou, C.; Liao, X.; Xu, Z.; Peng, Y.; Ou, X. Hierarchical structure-based noisy labels detection for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2183–2199. [CrossRef]
9. Li, Z.; Yang, X.; Meng, D.; Cao, X. An adaptive noisy label-correction method based on selective loss for hyperspectral image-classification problem. *Remote Sens.* **2024**, *16*, 2499. [CrossRef]
10. Jiang, J.; Ma, J.; Liu, X. Multilayer spectral–spatial graphs for label noisy robust hyperspectral image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 839–852. [CrossRef]
11. Yang, S.; Jia, Y.; Ding, Y.; Wu, X.; Hong, D. Unlabeled data guided partial label learning for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5503405. [CrossRef]
12. Ghafari, S.; Tarnik, M.G.; Yazdi, H.S. Robustness of convolutional neural network models in hyperspectral noisy datasets with loss functions. *Comput. Electr. Eng.* **2021**, *90*, 107009. [CrossRef]
13. Xu, Y.; Li, Z.; Li, W.; Du, Q.; Liu, C.; Fang, Z.; Zhai, L. Dual-channel residual network for hyperspectral image classification with noisy labels. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5502511. [CrossRef]
14. Roy, S.K.; Hong, D.; Kar, P.; Wu, X.; Liu, X.; Zhao, D. Lightweight heterogeneous kernel convolution for hyperspectral image classification with noisy labels. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5509705. [CrossRef]
15. Wang, L.; Zhu, T.; Kumar, N.; Li, Z.; Wu, C.; Zhang, P. Attentive-adaptive network for hyperspectral images classification with noisy labels. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5505514. [CrossRef]
16. Zhang, X.; Yang, S.; Feng, Z.; Song, L.; Wei, Y.; Jiao, L. Triple contrastive representation learning for hyperspectral image classification with noisy labels. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 500116. [CrossRef]
17. Ma, J.; Zou, Y.; Tang, X.; Zhang, X.; Liu, F.; Jiao, L. Spatial pooling transformer network and noise-tolerant learning for noisy hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5509719. [CrossRef]
18. Xu, H.; Zhang, H.; Zhang, L. A superpixel guided sample selection neural network for handling noisy labels in hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9486–9503. [CrossRef]
19. Wei, W.; Xu, S.; Zhang, L.; Zhang, J.; Zhang, Y. Boosting hyperspectral image classification with unsupervised feature learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5502315. [CrossRef]
20. Wang, C.; Zhang, L.; Wei, W.; Zhang, Y. Toward effective hyperspectral image classification using dual-level deep spatial manifold representation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5505614. [CrossRef]
21. Kang, X.; Zhu, Y.; Duan, P.; Li, S. Two dimensional spectral representation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5502809. [CrossRef]
22. Sun, W.; Du, Q. Hyperspectral band selection: A review. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 118–139. [CrossRef]
23. Zhao, Q.; Jia, S.; Li, Y. Hyperspectral remote sensing image classification based on tighter random projection with minimal intra-class variance algorithm. *Pattern Recognit.* **2021**, *111*, 107635. [CrossRef]
24. Peng, J.; Li, L.; Tang, Y.Y. Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 1790–1802. [CrossRef]
25. Li, X.; Zhang, L.; You, J. Locally weighted discriminant analysis for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 109. [CrossRef]
26. Zhang, Y.; Cao, G.; Li, X.; Wang, B.; Fu, P. Active semi-supervised random forest for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 2974. [CrossRef]
27. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [CrossRef]
28. Chen, Y.N.; Thaipisutikul, T.; Han, C.C.; Liu, T.J.; Fan, K.C. Feature line embedding based on support vector machine for hyperspectral image classification. *Remote Sens.* **2021**, *13*, 130. [CrossRef]
29. Mirzapour, F.; Ghassemian, H. Improving hyperspectral image classification by combining spectral, texture, and shape features. *Int. J. Remote Sens.* **2015**, *36*, 1070–1096. [CrossRef]
30. Li, W.; Prasad, S.; Fowler, J.E. Hyperspectral image classification using Gaussian mixture models and Markov random fields. *IEEE Geosci. Remote Sens. Lett.* **2013**, *11*, 153–157. [CrossRef]
31. Samat, A.; Li, E.; Wang, W.; Liu, S.; Lin, C.; Abuduwaili, J. Meta-XGBoost for hyperspectral image classification using extended MSER-guided morphological profiles. *Remote Sens.* **2020**, *12*, 1973. [CrossRef]

32. Hossain, M.D.; Chen, D. Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 115–134. [CrossRef]
33. Fang, L.; Li, S.; Kang, X.; Benediktsson, J.A. Spectral–spatial hyperspectral image classification via multiscale adaptive sparse representation. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7738–7749. [CrossRef]
34. Li, S.; Ni, L.; Jia, X.; Gao, L.; Zhang, B.; Peng, M. Multi-scale superpixel spectral–spatial classification of hyperspectral images. *Int. J. Remote Sens.* **2016**, *37*, 4905–4922. [CrossRef]
35. Yu, S.; Jia, S.; Xu, C. Convolutional neural networks for hyperspectral image classification. *Neurocomputing* **2017**, *219*, 88–98. [CrossRef]
36. Gao, Q.; Lim, S.; Jia, X. Hyperspectral image classification using convolutional neural networks and multiple feature learning. *Remote Sens.* **2018**, *10*, 299. [CrossRef]
37. Zhang, M.; Li, W.; Du, Q. Diverse region-based CNN for hyperspectral image classification. *IEEE Trans. Image Process.* **2018**, *27*, 2623–2634. [CrossRef]
38. Zhong, Z.; Li, J.; Luo, Z.; Chapman, M. Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 847–858. [CrossRef]
39. Zhou, J.; Zeng, S.; Xiao, Z.; Zhou, J.; Li, H.; Kang, Z. An enhanced spectral fusion 3D CNN model for hyperspectral image classification. *Remote Sens.* **2022**, *14*, 5334. [CrossRef]
40. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
41. Liu, Q.; Zhou, F.; Hang, R.; Yuan, X. Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification. *Remote Sens.* **2017**, *9*, 1330. [CrossRef]
42. Mei, S.; Li, X.; Liu, X.; Cai, H.; Du, Q. Hyperspectral image classification using attention-based bidirectional long short-term memory network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5509612. [CrossRef]
43. Zhong, P.; Gong, Z.; Li, S.; Schönlieb, C.B. Learning to diversify deep belief networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3516–3530. [CrossRef]
44. Tao, C.; Pan, H.; Li, Y.; Zou, Z. Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2438–2442.
45. Feng, J.; Feng, X.; Chen, J.; Cao, X.; Zhang, X.; Jiao, L.; Yu, T. Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification. *Remote Sens.* **2020**, *12*, 1149. [CrossRef]
46. Zhu, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Generative adversarial networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5046–5063. [CrossRef]
47. Sun, H.; Zheng, X.; Lu, X.; Wu, S. Spectral–spatial attention network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 323–3245. [CrossRef]
48. Liu, H.; Li, W.; Xia, X.; Zhang, M.; Gao, C.; Tao, R. Central attention network for hyperspectral imagery classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 8989–9003. [CrossRef]
49. Scheibenreif, L.; Mommert, M.; Borth, D. Masked vision transformers for hyperspectral image classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 2166–2176.
50. Li, Y.; Luo, Y.; Zhang, L.; Wang, Z.; Du, B. MambaHSI: Spatial-spectral mamba for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5524216. [CrossRef]
51. Xu, H.; Chen, W.; Tan, C.; Ning, H.; Sun, H.; Xie, W. Orientational clustering learning for open-set hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5508605. [CrossRef]
52. Di, X.; Xue, Z.; Zhang, M. Active learning-driven siamese network for hyperspectral image classification. *Remote Sens.* **2023**, *15*, 752. [CrossRef]
53. Wu, H.; Prasad, S. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Trans. Image Process.* **2017**, *27*, 1259–1270. [CrossRef] [PubMed]
54. Zhong, C.; Zhang, J.; Wu, S.; Zhang, Y. Cross-scene deep transfer learning with spectral feature adaptation for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2861–2873. [CrossRef]
55. Song, H.; Kim, M.; Park, D.; Shin, Y.; Lee, J.G. Learning from noisy labels with deep neural networks: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 8135–8153. [CrossRef] [PubMed]
56. Shi, J.; Zhang, K.; Guo, C.; Yang, Y.; Xu, Y.; Wu, J. A survey of label-noise deep learning for medical image analysis. *Med. Image Anal.* **2024**, *95*, 103166. [CrossRef]
57. Shin, J.; Won, J.; Lee, H.S.; Lee, J.W. A review on label cleaning techniques for learning with noisy labels. *ICT Express* **2024**, in press. [CrossRef]
58. Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; Li, L.J. Learning from noisy labels with distillation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1910–1918.
59. Lukov, T.; Zhao, N.; Lee, G.H.; Lim, S.N. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 648–665.
60. Algan, G.; Ulusoy, I. MetaLabelNet: Learning to generate soft-labels from noisy-labels. *IEEE Trans. Image Process.* **2022**, *31*, 4352–4362. [CrossRef]

61. Wu, Y.; Shu, J.; Xie, Q.; Zhao, Q.; Meng, D. Learning to purify noisy labels via meta soft label corrector. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 10388–10396.
62. Tu, B.; Zhou, C.; Kuang, W.; Guo, L.; Ou, X. Hyperspectral imagery noisy label detection by spectral angle local outlier factor. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1417–1421. [CrossRef]
63. Tu, B.; Zhang, X.; Kang, X.; Zhang, G.; Li, S. Density peak-based noisy label detection for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1573–1584. [CrossRef]
64. Tu, B.; Zhou, C.; He, D.; Huang, S.; Plaza, A. Hyperspectral classification with noisy label detection via superpixel-to-pixel weighting distance. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4116–4131. [CrossRef]
65. Kang, X.; Duan, P.; Xiang, X.; Li, S.; Benediktsson, J.A. Detection and correction of mislabeled training samples for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5673–5686. [CrossRef]
66. Leng, Q.; Yang, H.; Jiang, J. Label noise cleansing with sparse graph for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 1116. [CrossRef]
67. Bahraini, T.; Azimpour, P.; Yazdi, H.S. Modified-mean-shift-based noisy label detection for hyperspectral image classification. *Comput. Geosci.* **2021**, *155*, 104843. [CrossRef]
68. Kang, J.; Fernandez-Beltran, R.; Duan, P.; Kang, X.; Plaza, A.J. Robust normalized softmax loss for deep metric learning-based characterization of remote sensing images with label noise. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 8798–8811. [CrossRef]
69. Damodaran, B.B.; Flamary, R.; Seguy, V.; Courty, N. An entropic optimal transport loss for learning deep neural networks under label noise in remote sensing images. *Comput. Vis. Image Underst.* **2020**, *191*, 102863. [CrossRef]
70. Zhang, Y.; Sun, J.; Shi, H.; Ge, Z.; Yu, Q.; Cao, G.; Li, X. Agreement and disagreement-based co-learning with dual network for hyperspectral image classification with noisy labels. *Remote Sens.* **2023**, *15*, 2543. [CrossRef]
71. Liao, Q.; Zhao, L.; Luo, W.; Li, X.; Zhang, G. Joint negative–positive-learning based sample reweighting for hyperspectral image classification with label noise. *Pattern Recognit. Lett.* **2024**, *183*, 98–103. [CrossRef]
72. Fang, C.; Zhang, G.; Li, J.; Li, X.; Chen, T.; Zhao, L. Intelligent agent for hyperspectral image classification with noisy labels: A deep reinforcement learning framework. *Int. J. Remote Sens.* **2024**, *45*, 2939–2964. [CrossRef]
73. Yue, J.; Fang, L.; Rahmani, H.; Ghamisi, P. Self-supervised learning with adaptive distillation for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5501813. [CrossRef]
74. Zhao, W.; Peng, R.; Wang, Q.; Cheng, C.; Emery, W.J.; Zhang, L. Life-long learning with continual spectral-spatial feature distillation for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5544214. [CrossRef]
75. Chi, Q.; Lv, G.; Zhao, G.; Dong, X. A novel knowledge distillation method for self-supervised hyperspectral image classification. *Remote Sens.* **2022**, *14*, 4523. [CrossRef]
76. Feng, S.; Zhang, H.; Xi, B.; Zhao, C.; Li, Y.; Chanussot, J. Cross-domain few-shot learning based on decoupled knowledge distillation for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5534414. [CrossRef]
77. Ullah, F.; Ullah, I.; Khan, R.U.; Khan, S.; Khan, K.; Pau, G. Conventional to deep ensemble methods for hyperspectral image classification: A comprehensive survey. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 3878–3916. [CrossRef]
78. Lu, Y.; Bo, Y.; He, W. An ensemble model for combating label noise. In Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM), Tempe, AZ, USA, 21–25 February 2022; pp. 608–617.
79. Shao, H.C.; Wang, H.C.; Su, W.T.; Lin, C.W. Ensemble learning with manifold-based data splitting for noisy label correction. *IEEE Trans. Multimed.* **2022**, *24*, 1127–1140. [CrossRef]
80. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
81. Bylander, T. Estimating generalization error on two-class datasets using out-of-bag estimates. *Mach. Learn.* **2002**, *48*, 287–297. [CrossRef]
82. Zhang, Z.; Sabuncu, M.R. Generalized cross entropy loss for training deep neural networks with noisy labels. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018; pp. 8792–8802.
83. Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 322–330.
84. Thoreau, R.; Risser, L.; Achard, V.; Berthelot, B.; Briottet, X. Toulouse Hyperspectral data set: A benchmark data set to assess semi-supervised spectral representation learning and pixel-wise classification techniques. *ISPRS J. Photogramm. Remote Sens.* **2024**, *212*, 323–337. [CrossRef]