remote sensing

MDPI

Article

Pair-Wise Similarity Knowledge Distillation for RSI Scene Classification

Haoran Zhao [1], Xin Sun [1,2,*], Feng Gao [1] and Junyu Dong [1]

[1] College of Information Science and Engineering, Haide College and Institute of Advanced Ocean Study, Ocean University of China, Qingdao 266100, China; zhaohaoran@stu.ouc.edu.cn (H.Z.); gaofeng@ouc.edu.cn (F.G.); dongjunyu@ouc.edu.cn (J.D.)
[2] The Department of Aerospace and Geodesy, Technical University of Munich, 80333 München, Germany
* Correspondence: sunxin1984@ieee.org

Abstract: Remote sensing image (RSI) scene classification aims to identify the semantic categories of remote sensing images based on their contents. Owing to the strong learning capability of deep convolutional neural networks (CNNs), RSI scene classification methods based on CNNs have drawn much attention and achieved remarkable performance. However, such outstanding deep neural networks are usually computationally expensive and time-consuming, making them impossible to apply on resource-constrained edge devices, such as the embedded systems used on drones. To tackle this problem, we introduce a novel pair-wise similarity knowledge distillation method, which could reduce the model complexity while maintaining satisfactory accuracy, to obtain a compact and efficient deep neural network for RSI scene classification. Different from the existing knowledge distillation methods, we design a novel distillation loss to transfer the valuable discriminative information, which could reduce the within-class variations and restrain the between-class similarity, from the cumbersome model to the compact model. This method could obtain the compact student model with higher performance compared with existing knowledge distillation methods in RSI scene classification. To be specific, we distill the probability outputs between sample pairs with the same label and match the probability outputs between the teacher and student models. Experiments on three public benchmark datasets for RSI scene classification, i.e., AID, UCMerced, and NWPU-RESISC datasets, verify that the proposed method could effectively distill the knowledge and result in a higher performance.

Keywords: knowledge distillation; imaging science; scene classification; geosciences; convolutional neural network

Citation: Zhao, H.; Sun, X; Gao, F; Dong, J. Pair-Wise Similarity Knowledge Distillation for RSI Scene Classification. *Remote Sens.* **2022**, *14*, 2483. https://doi.org/10.3390/rs14102483

Academic Editors: Zhou Zhang, Zhengxia Zou, Bin Pan and Xia Xu

Received: 4 April 2022
Accepted: 19 May 2022
Published: 22 May 2022

1. Introduction

The tremendous development of imaging science has significantly improved the resolution and quality of images. For example, images derived from remote sensing technology, which has seen huge improvements in recent years, are of high resolution and quality. This has resulted in the requirement for researchers to investigate meaningful aspects of remote sensing image processing. Among these aspects, remote sensing image scene classification is a fundamental and crucial task for remote sensing image analysis and applications such as geosciences, environment monitoring [1] and urban planning [2]. RSI scene classification proposes to correctly classify remote sensing samples with different semantic classes. In the remote sensing community, RSI scene classification has been studied by designing handcrafted features for the earlier years. However, these methods perform poorly for complex scenes or massive data. In recent years, CNNs have been extensively applied to extract more discriminative features of RSI and have achieved great success [3–5]. However, these high-performance CNNs are usually designed with large parameter sets and high complexity, which comes with a high computation cost and makes

it impossible to be applied on edge-computing devices, such as the embedded systems used on drones.

Some typical network compression techniques, such as network weight pruning [6], low-rank decomposition [7], and knowledge distillation [8], have been proposed to reduce the computational complexity and high storage requirement of CNNs. Different from other methods, knowledge distillation (KD) does not directly compress the parameters of the large network. It obtains a lightweight student network by transferring valuable information from the complicated teacher network. To be specific, KD [8–10] aims to obtain a lightweight student model with high accuracy. To this end, it transfers the dark knowledge from the complicated teacher model to the student model. Thus, it is a natural and effective method to obtain a lightweight deep neural network. It has been used to address some practical applications such as remote sensing scene classification [11,12], object detection from drone images [13], and compact cloud detection [14].

To obtain an outstanding and compact model for RSI scene classification, we aim to distill the valuable knowledge from the complicated teacher model to the compact student, as shown in Figure 1. However, RSI scene classification usually faces the challenge of big intra-class variations and high inter-class similarities. Although the complicated teacher model is powerful when applied to solve such challenges, the discriminative information contained in the teacher model will be attenuated and even lost when applying existing KD methods. Thus, existing KD methods can hardly obtain satisfactory results for RSI scene classification. In other words, these methods lose some valuable discriminative knowledge in the process of distillation. Although the complicated teacher model has strong capabilities to classify scene images, such a powerful ability fails to distill information to the compact student by existing KD methods. Thus, traditional KD loss cannot satisfactorily transfer valuable knowledge to the compact student model for solving the challenge of big intra-class variations and high inter-class similarities.

To efficiently train and obtain a compact student model with high performance, this paper will design a pair-wise similarity knowledge distillation approach for RSI scene classification. To obtain a compact model with high performance for image scene classification, we focus on transferring the valuable discriminative information from the cumbersome model to the compact model. To be specific, we first distill the probability outputs of the student model between sample pairs with the same label. We expect that similar samples with the same label will produce similar incorrect outputs. This way, the within-class variations could be reduced. Second, we create virtual samples by mixup technology, which represents the similarity correlation between samples with different labels. Then, we match the probability outputs between the teacher model and the student model to restrain the between-class similarity. Finally, we effectively distill the teacher model's powerful ability, which could reduce the within-class variations and restrain the between-class similarity, to the student model. We summarize the main contributions as follows:

- We introduce a pair-wise similarity knowledge distillation approach to obtain a lightweight model with high performance for RSI scene classification. It could effectively distill the teacher model's powerful ability to the student model, which could reduce the intra-class variations and restrain the inter-class similarities;
- We employ the sample pairs and virtual samples as the training data. We reduce the intra-class variations and restrain the inter-class similarities by forcing similar incorrect outputs and matching the outputs, causing the student model to well absorb the discriminative knowledge from the teacher model;
- We verify the proposed pair-wise similarity knowledge distillation framework on AID, UCMerced and NWPU-RESISC datasets. The experimental results show that the proposed approach can significantly improve the lightweight model's performance in terms of RSI scene classification.

The rest of this paper is organized as follows. In Section 2, traditional related methods are reviewed. In Section 3, we introduce the pair-wise similarity knowledge distillation

method. Furthermore, we present a detailed comparison in Section 4 and discuss our findings in Section 5. Finally, we conclude this paper in Section 6.

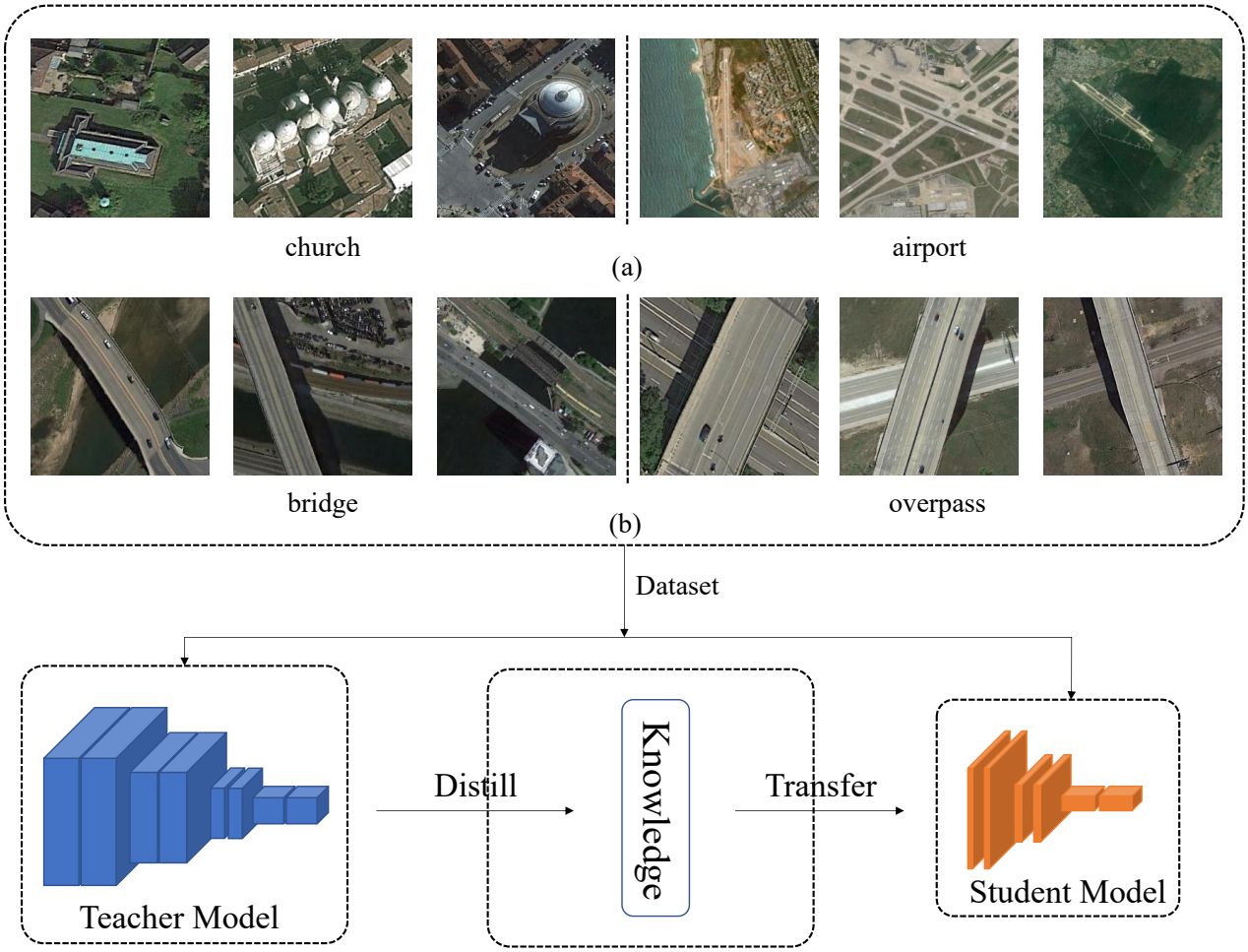

Figure 1. The motivation of the proposed method. (**a**) Various building styles of churches and different shapes of airports can be seen. (**b**) We observe that the images of the bridge seems similar to those of the overpass. Unlike the complicated deep models, light models are unable to deal with such challenges as (**a**) within-class diversity and (**b**) between-class similarity. We aim to obtain a lightweight network by distilling valuable knowledge from the complicated network.

2. Related Work

2.1. Remote Sensing Image Scene Classification

RSI scene classification is a fundamental and important task for intelligent geosciences observation. It can be widely used for some practical applications, such as image processing for geosciences and urban planning. However, RSI scene classification is a difficult issue due to large intra-class variations and high inter-class similarities in remote sensing datasets. To solve these problems, a number of approaches have been introduced in the past decades.

In the early years, most of the RSI scene classification approaches were based on hand-crafted features. Generally, these methods extract low-level hand-crafted features that represent the images by some classical feature descriptor, such as scale-invariant feature transformation (SIFT) [15] or the histogram of oriented gradients (HOG) [16]. However, these methods have local characteristics and cannot directly represent the entire image. Thus, some feature-encoding methods, including bag-of-visual-words (BoVW) [17], are used to encode the obtained features for representing an entire scene image. Then, the classifier [18,19] is trained using the encoded features for scene recognition. Afterwards,

due to the limited representation capability of hand-crafted features, some unsupervised learning methods [20–23] are used to extract features for scene classification.

Recently, CNNs [24,25] have achieved huge success in the computer vision community due to their powerful feature representation. Therefore, CNNs have been extended to RSI scene classification [3,5,26] and have achieved a dominant position in the field of RSI scene classification. For example, Cheng et al. [3] introduced a metric learning regularization item for optimizing the discriminative loss function. Hua et al. [27] introduced the relationship network to predict label existences using label relationships.

Although CNN-based methods have achieved remarkable success in RSI classification, such top-performing methods are computationally expensive and time-consuming. It is not possible to deploy such complicated models on edge devices, such as drones. Thus, some model compression methods are introduced to solve such problems. Chen et al. [28] introduced the standard knowledge distillation [8] into RSI scene classification. They boosted the compact model's accuracy by matching the high-temperature softmax outputs from the large model. However, they ignored and thus experienced a loss in the valuable discriminative information for restraining the intra-class variations and inter-class similarities in the process of distillation.

2.2. Knowledge Distillation

Although CNNs have top-performing results in various tasks [29], it is a challenging task to deploy such complicated models into edge-computing devices. To tackle these problems, CNN compression and acceleration [6,7] have been proposed for some real-time applications on resource-constrained devices, such as drones. There are three mainstream categories: network pruning [6], network quantization [30], and knowledge distillation [8].

Knowledge distillation is one of the most efficient and practical approaches among these methods. The idea of transferring knowledge from an ensemble of models to a compact model was proposed by Bucilua et al. [9]. Caruana et al. [31] further extended the method by matching the output logits between the teacher and student models. The process of knowledge distillation was first defined by Hinton et al. [8]. They introduced a temperature hyper-parameter to soften the logits before softmax. Thus, the teacher model could provide the softened outputs containing extra dark knowledge in the process of distillation. Finally, the compact student model was trained by imitating the distribution of the teacher model's softened outputs.

Some recent works [32–34] extend KD by distilling different kinds of dark knowledge. Unlike standard KD [8], they use the outputs from the intermediate layers, i.e., feature outputs instead of the probability outputs from the last layers, as dark knowledge. Moreover, the relationships between different feature maps and data samples are explored as dark knowledge [35,36].

For example, Park et al. [35] explored the relationships between image samples and introduced a novel relational knowledge distillation (RKD) approach. Tung et al. [36] fully utilized the similar activation of input pairs to achieve similarity-preserving (SP) knowledge distillation. Peng et al. [37] employed the kernel approach to obtain the high order correlation of images. Zhao et al. [38] proposed the gradual distillation of the student network by a curriculum-learning strategy.

To the best of our knowledge, there are few works that have successfully applied knowledge distillation for RSI scene classification. Due to the challenge of intra-class variations and between-class similarities, existing KD methods cannot be efficiently used for RSI scene classification. When distilling the information from the teacher model to the student model using the existing methods, the valuable discriminative information for within-class diversity and between-class similarity is easy to lose, which reduces the accuracy of the student network. Thus, we customize a pair-wise similarity knowledge distillation method for RSI scene classification.

3. Materials and Methods

In this section, we will present the proposed pair-wise similarity knowledge distillation for RSI scene classification. As shown in Figure 2, we distill the pair-wise knowledge to the student model, which could restore the discriminative information for within-class diversity and between-class similarity.

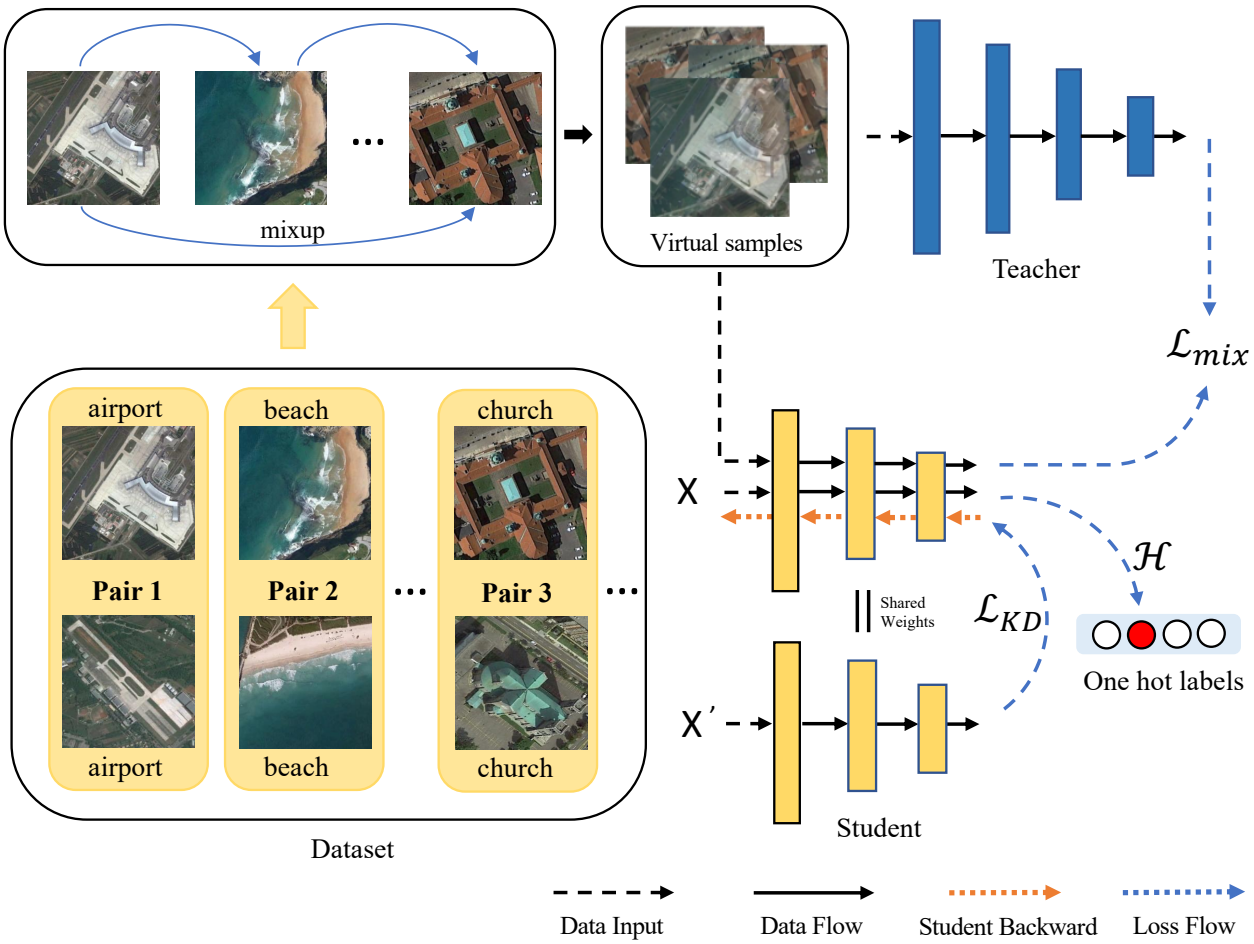

Figure 2. The overall framework for our pair-wise similarity knowledge distillation.

3.1. The Standard Teacher–Student Distillation

We first introduce the standard knowledge distillation paradigm before presenting our pair-wise similarity knowledge distillation for RSI scene classification. We define the complicated teacher and compact student models as T and S, respectively. Furthermore, we denote the initial training data consisting of tuples of input images and labels as $(x, y) \in D$. Given an input x_i, the output of the last fully connected layer is given as logits z_i. The final probability can be evaluated by a softmax function,

$$P_i = \frac{exp(z_i)}{\sum exp(z_i)} \tag{1}$$

Furthermore, Hinton et al. [8] introduced a temperature t to soften the probability and employ the soft targets as dark knowledge, which contains more valuable information than one-hot labels. The softened probability is as follows:

$$P_i = \frac{exp(z_i/t)}{\sum exp(z_i/t)} \tag{2}$$

where a higher temperature t could produce a softer probability distribution over classes. Then, they tried to match the soft targets between the teacher model and student model using the Kullback–Leibler divergence loss,

$$\mathcal{L}_{KD} = t^2 KL(P_{z_t}, P_{z_s}) \tag{3}$$

KL means the Kullback–Leibler divergence. P_{z_t} and P_{z_s} represent the soften probabilities of the teacher and student networks. Thus, the student model will be trained under the supervision of standard cross-entropy loss and knowledge distillation loss,

$$\mathcal{L}_{Student} = (1-\delta)\mathcal{L}_{CE} + \delta\mathcal{L}_{KD} \tag{4}$$

where δ is a hyper-parameter to balance the above two items.

3.2. Pair-Wise Similarity Knowledge Distillation

The discriminative information for the within-class diversity and between-class similarity is vital for RSI scene classification. Existing knowledge distillation methods ignore such valuable knowledge when transferring the dark knowledge from the teacher to the student. Thus, we design a pair-wise similarity knowledge distillation method for RSI scene classification. It could efficiently distill the teacher model's knowledge to the student mode, for solving the within-class diversity and between-class similarity.

First, we randomly sample different data points with the same label in pairs. Then, we feed the samples of data pairs to the student model, and expect the student networks to produce similar incorrect predictions. In this way, we could reduce the within-class variations of the student network. Formally, we define the input sample pair as (x, x'), which has the same label y. We define the first loss function as follows:

$$\mathcal{L}_{KD}(x, x', \theta) = T^2 KL(P(y|x'; \tilde{\theta}), P(y|x; \theta)) \tag{5}$$

where $\tilde{\theta}$ is a fixed copy of the parameter θ. In the training phase, we only update the parameter θ. KL means the Kullback–Leibler divergence. The standard KD method [8] also needs to match two probability outputs. The difference is that we create two probability outputs from the single student model by self-distillation.

Second, we create virtual samples by mixing two images with different labels. We expect to capture and transfer the similarity correlation between samples with different labels. To be specific, the virtual sample $(\hat{x}, \hat{y})$ is produced by mixing two random samples (x_i, y_i) and (x_j, y_j). The weighted linear interpolation is as follows:

$$\hat{x} = \lambda x_i + (1-\lambda)x_j \tag{6}$$

$$\hat{y} = \lambda y_i + (1-\lambda)y_j \tag{7}$$

where $\lambda \in [0, 1]$ is the merging coefficient. It is randomly produced from the $Beta(\alpha, \alpha)$ distribution. As the two samples (x_i, y_i) and (x_j, y_j) should be equivalent, the sampling probability needs to be between $[0, 1]$. The $Beta(\alpha, \alpha)$ distribution meets this requirement. Moreover, the $Beta(\alpha, \alpha)$ distribution is flexible in order to obtain the diversified $[0, 1]$ probability distribution by adjusting the parameter α. Thus, when we adjust and choose the value of α, we could randomly sample a probability value λ drawn from the $Beta(\alpha, \alpha)$

distribution. Furthermore, y is the label of x, which is a convexed combination of the labels from x_i and x_j. Therefore, the mix loss can be formulated as:

$$\mathcal{L}_{mix} = \lambda \mathcal{H}(y_{s_i}, y) + (1 - \lambda)\mathcal{H}(y_{s_j}, y) + \mathcal{L}_{KD}(S(x, P_{z_t}), T(x, P_{z_s})) \quad (8)$$

where y_{s_i} and y_{s_j} are the predictive distribution of two mixed images, respectively. y is the ground truth label. λ is a hyper-parameter controlling the mixed ratio of two images. $\mathcal{H}$ denotes the standard cross-entropy loss. In this way, we could increase the between-class variations by transferring the similarity information between classes.

In addition, we force the student model to learn the hard targets from the ground truth labels using the standard cross-entropy loss, as in the original KD method. Here, we introduce the total loss as follows:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{KD} + \beta \mathcal{L}_{mix} + \gamma \mathcal{H}(y_s, y) \quad (9)$$

Note that, unlike traditional KD methods, we employ the pair-wise regularization loss to restore discriminative information, such as intra-class variations and inter-class similarities, for RSI scene classification. To be specific, we force the student model to output a similar incorrect probability distribution when feeding different images in a pair with the same label. This is a benefit that allows the student model to solve within-class diversity. Then, we employ the mixup technique to mix samples with different labels. Hence, the student model tries to match the soft predictions from the teacher model containing the inter-class information. Afterwards, the student model is able to solve the inter-class similarity for remote sensing image scene classification.

3.3. Training Procedure

We first train the teacher model by the standard supervised learning strategy. Then, we distill the student network using the proposed method by Equation (9). To be specific, we first feed the sample pairs with the same label to the student model, which is trained through self-distillation by Equation (5). Note that the paired samples are fed to the student model, which has the parameters θ. The gradient is not propagated through $\tilde{\theta}$, which is a fixed copy of the student model. Different from the standard KD method, we feed the sample pair to the single student model and make two probability outputs. We force the student model to produce similar incorrect predictions. To be specific, P in Equation (5) means the the predictive distribution by a softmax classifier. x and x', are the samples with the same label y. Thus, x and x' produce two predictive distributions by P in Equation (5). The predictive distribution contains the ground-truth prediction and the incorrect predictions. We expect that the different samples (x and x') with the same class label (y) not only have the same ground-truth prediction but also have similar incorrect predictions. Thus, we employ the Kullback–Leibler divergence to measure and control the similarity of two probability distributions. When training converges, we obtain similar incorrect predictions for the sample pair x and x'.

Second, we use the paired images with different labels to create virtual images by the mixup technique. Thus, the information of different classes is mixed into the virtual images. We feed the virtual samples to the teacher and student models, respectively, and force the student model to mimic the predictive distributions from the teacher model by Equation (8). Furthermore, the student model also matches the ground-truth label by a standard cross-entropy loss. The whole training procedure with our total loss in Equation (9) is illustrated in Algorithm 1.

Algorithm 1 Pair-wise Similarity Knowledge Distillation

Input: Teacher network T, Student network S with the parameters θ, $\tilde{\theta}$ is a fixed copy of parameters θ, training samples $X = \{(x_i, y_i)_{i=1}^{N}\}$.
Output: parameters θ.
Initialize: T, S and the hyper-parameters.
Stage 1: Train the teacher model.

1: **Repeated:**
2: computing $\mathcal{H}(y_{true}, P_{z_t})$.
3: updating parameters of T by gradient back-propagation.
4: **Until:** $\mathcal{H}(y_{true}, P_{z_t})$ converged.

Stage 2: Distilling the student via pair-wise similarity KD.

1: Creating the Virtual Samples by Mixup.
2: **Repeated:**
3: feed data in pair to our framework.
4: computing $\mathcal{L}_{total}$ by Equation (9).
5: updating the parameters θ of student by gradient back-propagation.
6: **Until:** $\mathcal{L}_{total}$ converged.
7: **Return** the student's parameters θ.

4. Results

We evaluate the effectiveness of our approach by conducting extensive experiments on three popular RSI scene classification benchmark datasets: the NWPU-RESISC45 dataset [39], Arial Image dataset (AID) [40], and UC Merced Land-Use dataset [41]. We implement the proposed method with Pytorch on NVIDIA 2080Ti GPUs.

4.1. Datasets

NWPU-RESISC45. The NWPU-RESISC45 dataset is currently the largest RSI scene classification dataset. It has 31,500 samples chosen from more than 100 countries and regions. It contains 45 scene categories and each category has 700 samples with a size of 256 × 256 pixels in the RGB space. Each pixel has a spatial resolution ranging from 300 to 20 cm/pixel. It is a challenging dataset due to the large inter-class similarities.
AID. The aerial image dataset (AID) comprises 10,000 images with 30 categories of scene samples. Each category consists of approximately 200 to 400 images with a size of 600 × 600 pixels in RGB space. Each pixel has a spatial resolution ranging from 800 to 50 cm/pixel.
UC Merced Land-Use. The UC Merced Land-Use dataset has 2100 images and consists of 21 scene categories. Each category has 100 images with a size of 256 × 256 pixels. Each pixel has a spatial solution of 30 cm in the RGB color space.

We randomly sample some images from these three datasets, as shown in Figure 3. Qe observe that some categories from these RSI datasets have high between-class similarities. For example, the forest and sparse residential area (from the UC Merced Land-Use dataset in the first line), and the stadium and playground (from the AID dataset in the second line). This presents great challenges to the CNNs, especially the small and compact models with low capability.

The general information is shown in Table 1. We present the information of these remote sensing image scene classification datasets with an 80% training ratio. We randomly split the dataset into an 80% training set and 20% testing set.

Table 1. Different datasets for RSI scene classification with the 80% training ratio. We present the detailed information of each dataset.

Dataset	Resolution	Size	Categories	Entire Dataset	Train Subset	Test Images
NWPU-RESISC45	0.2–3 m	256 × 256	45	31,500	25,200	6300
AID	0.5–8 m	600 × 600	30	10,000	8000	2000
UC Merced Land-Use	0.3 m	256 × 256	21	2100	1680	420

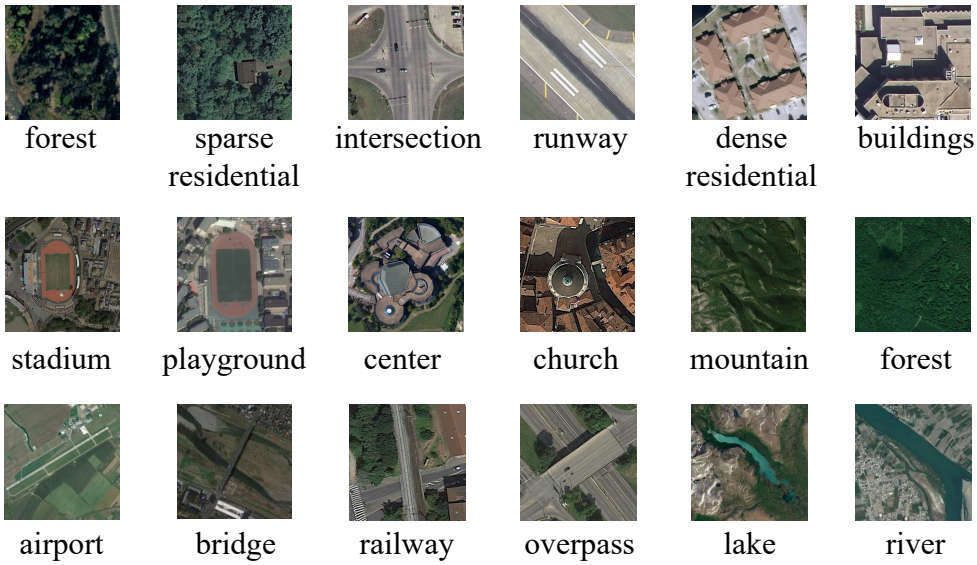

Figure 3. We present some example samples from three RSI scene classification datasets. The images in each row are UC Merced Land-Use, AID and NWPU-RESISC45, respectively. From these examples, we could observe that some categories have very high between-class similarities. For example, the forest and sparse residential area on the first line look very similar.

4.2. Experimental Settings.

Network architecture. In the following experiments, we employ the ResNet [42] as our base architecture for the teacher and student networks. The ResNet series model achieves the state-of-the-art performance by stacking the basic residual blocks. We employ ResNet-101 as the teacher network and ResNet-34 as the student network. Note that, the residual block is two layers deep in ResNet-34 and three layers deep in ResNet-101. The residual block could greatly increase the depth and easily gain accuracy in remote sensing image scene classification. For the NWPU-RESISC45 dataset, we employ the ResNet-101 as the teacher network and the ResNet as the student network. For the other two datasets, we adopt the ResNet-34/ResNet-18 as the teacher/student model pair.

Implementation details. We first conduct experiments on the NWPU-RESISC45 dataset, which has 256 × 256 RGB images. For the training phase, we adopt data augmentations such as horizontal flips and random crops. The original testing data are used for the testing phase. Due to the high number of scene images in the NWPU-RESISC45 dataset, we employ ResNet-101 and ResNet-34 as the teacher model and student model, respectively. We set the mini-batch as 32. We set the training epochs as 200 and initial learning rate as 0.1, which is reduced by a factor of 0.1 on epoch 60, 120, and 160, respectively. We adopt the stochastic gradient descent (SGD) with momentum 0.9. We set the hyper-parameter as follows: $\alpha = 0.1$, $\beta = 1$, $T = 4$.

For the AID and UC Merced Land-Use datasets, we use the same settings as above. The difference is that ResNet-34 and ResNet-18 are adopted as the teacher model and student model, respectively.

5. Discussion

Evaluation Criteria. Overall accuracy (OA) aims to evaluate the performance of the classifiers on the testing dataset. It is formulated as the number of correctly classified images divided by the whole number of testing images. OA is the most used criterion for measuring the effectivity of the methods for RSI scene classification. Thus, we also use the OA as our evaluation criteria for all experiments. In addition, we also adopt the confusion matrix table in the ablation study to show the detailed classification results, which visualizes the accuracy of each class.

Performance Comparison. We verify the effectiveness of our proposed pair-wise similarity knowledge distillation approach on three RSI scene classification datasets. For the baselines, we employ several popular state-of-the-art methods for knowledge distillation. To be specific, we compare our pair-wise similarity KD method with knowledge distillation (KD) [8], attention transfer (AT) [33], similarity-preserving (SP) [36], and relational knowledge distillation (RKD) [35] for all RSI scene classification datasets.

We first evaluate our method on the NWPU-RESISC45 dataset, which is currently the largest remote sensing image scene classification dataset. Thus, we employ ResNet-101 and ResNet-34 as our teacher model and student model, respectively, for this dataset. We prepare the ResNet-101 teacher model by standard back-propagation training, which provides 94.772% classification accuracy with a 70% training ratio. Then, we train ResNet-34 as the student model using the proposed method. Table 2 illustrates the OA of the proposed method compared with several KD methods. Note that the baseline means the student model (ResNet-34) was trained individually, providing 93.619% accuracy. The experimental results show that the proposed approach achieves 95.069% accuracy with a 1.45% improvement over the baseline. Moreover, our method achieves a notable improvement compared with the state-of-the-art KD approaches. For example, the student model of our method achieves a 0.465% improvement compared with the standard KD method. Moreover, the same advantage is also reflected in another experiment with an 80% training ratio. Thus, we conclude that our method could effectively distill the valuable information from the teacher model to the student model and enhance the generalization of the student model.

Table 2. Overall accuracy (%) on NWPU-RESISC45 dataset. The ResNet-101 is adopted as the teacher network and ResNet-34 is employed as the student network. The baseline represents the standard back-propagation for the student.

Dataset	Training Ratio	Baselines	KD [8]	AT [33]	SP [36]	RKD [35]	WRD [43]	Ours	Teacher
NWPU-RESISC45	70%	93.619	94.604	94.127	94.794	94.801	94.813	**95.069**	94.772
	80%	94.021	94.576	94.924	95.313	95.297	95.301	**95.673**	95.394
AID	50%	91.100	91.208	91.302	91.381	91.394	91.328	**91.571**	91.360
	80%	94.350	94.401	94.483	94.494	94.572	94.032	**94.850**	94.550
UC Merced Land-Use	50%	90.195	90.328	90.496	90.805	90.890	90.916	**91.237**	91.062
	80%	93.095	93.409	93.380	94.076	94.340	94.091	**94.903**	94.524

Furthermore, we conduct experiments on the AID and UC Merced Land-Use datasets to verify the effectiveness of our method. We employ the ResNet-34 as the teacher network and the ResNet-18 as the student network for these two datasets. As can be seen from Table 2, our method achieves 94.850% and 94.903% accuracy on the AID and UC Merced Land-Use datasets with an 80% training ratio, respectively. It provides improvements of 0.5% and 1.494% compared with the baselines, which achieve 94.350% and 93.095% on the AID and UC Merced Land-Use datasets, respectively. Interestingly, we find that the proposed approach not only achieves a significant improvement compared with the baselines, but also surpasses the teacher model. The reason is that the proposed distillation loss for RSI scene classification

could efficiently transfer the teacher model's valuable information, which is vital to solve the challenge of large within-class diversity and high between-class similarity.

The training curves and confusion matrix. We present the student model's testing accuracy using different methods. It is trained on the NWPU-RESISC45 dataset with an 80% training ratio. As can be seen from Figure 4, we observe that our method (purple line) achieves a significant improvement compared with the student model, trained individually (blue line). Moreover, the proposed approach outperforms the traditional distillation loss (yellow line), as shown in Figure 4.

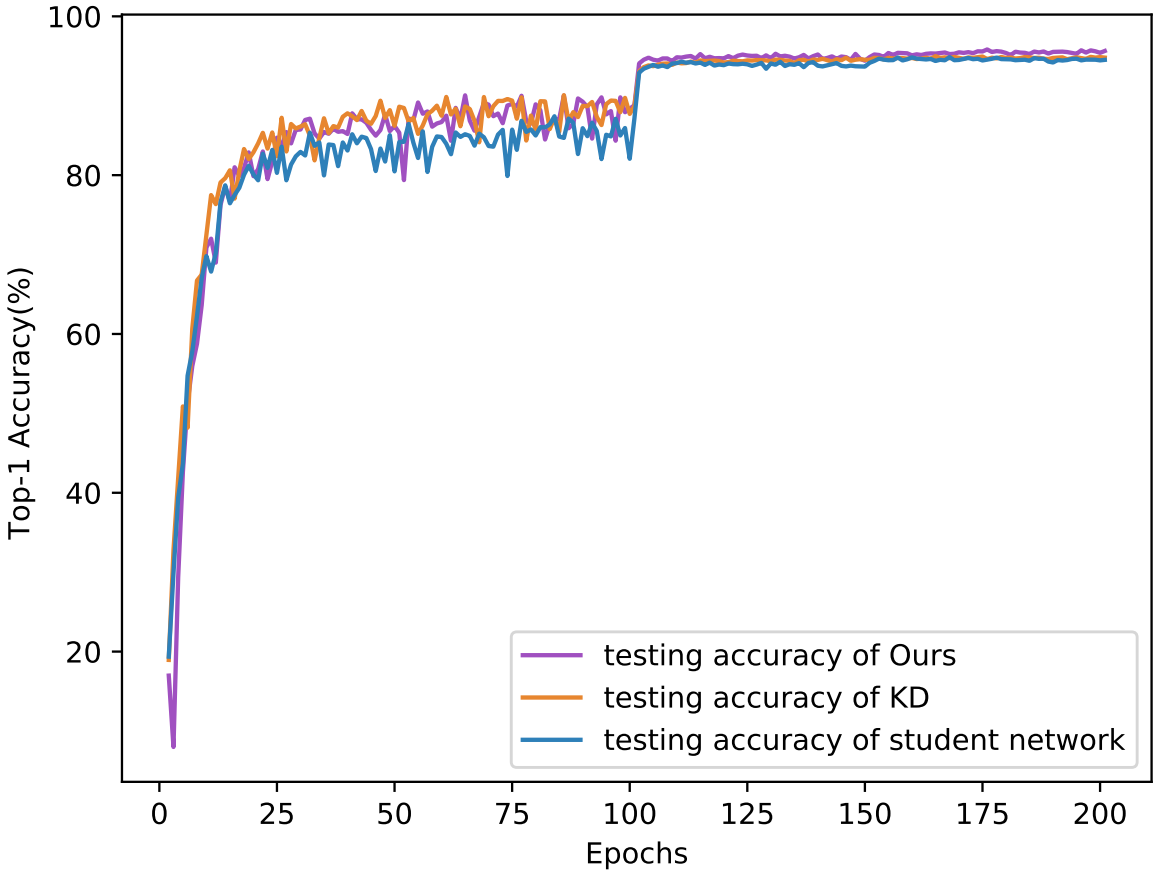

Figure 4. Testing accuracy on the NWPU-RESISC45 dataset, using different approaches. The training ratio is set as 80%.

Then, we illustrate the confusion matrix to find the accuracy of each class. Note that the white spaces stand for elements that are equal to zero. The value of each element (i, j) represents the ratio of testing samples with label i but classified as class j to the total testing images. Figures 5 and 6 show the confusion matrix of the corresponding experiments, given in Table 2. As can be seen from Figure 5, the proposed approach achieves very high accuracy in most categories, except harbor and snowberg. Note that it is also difficult for a human expert to distinguish some of the misclassified images in such classes. Figure 6 also shows the confusion matrix corresponding to the experiment with an 80% training ratio, as given in Table 2. The categories with the lowest classification accuracies are center, resort, and school, due to the fact that the teacher model also finds it difficult to correctly classify these challenging classes.

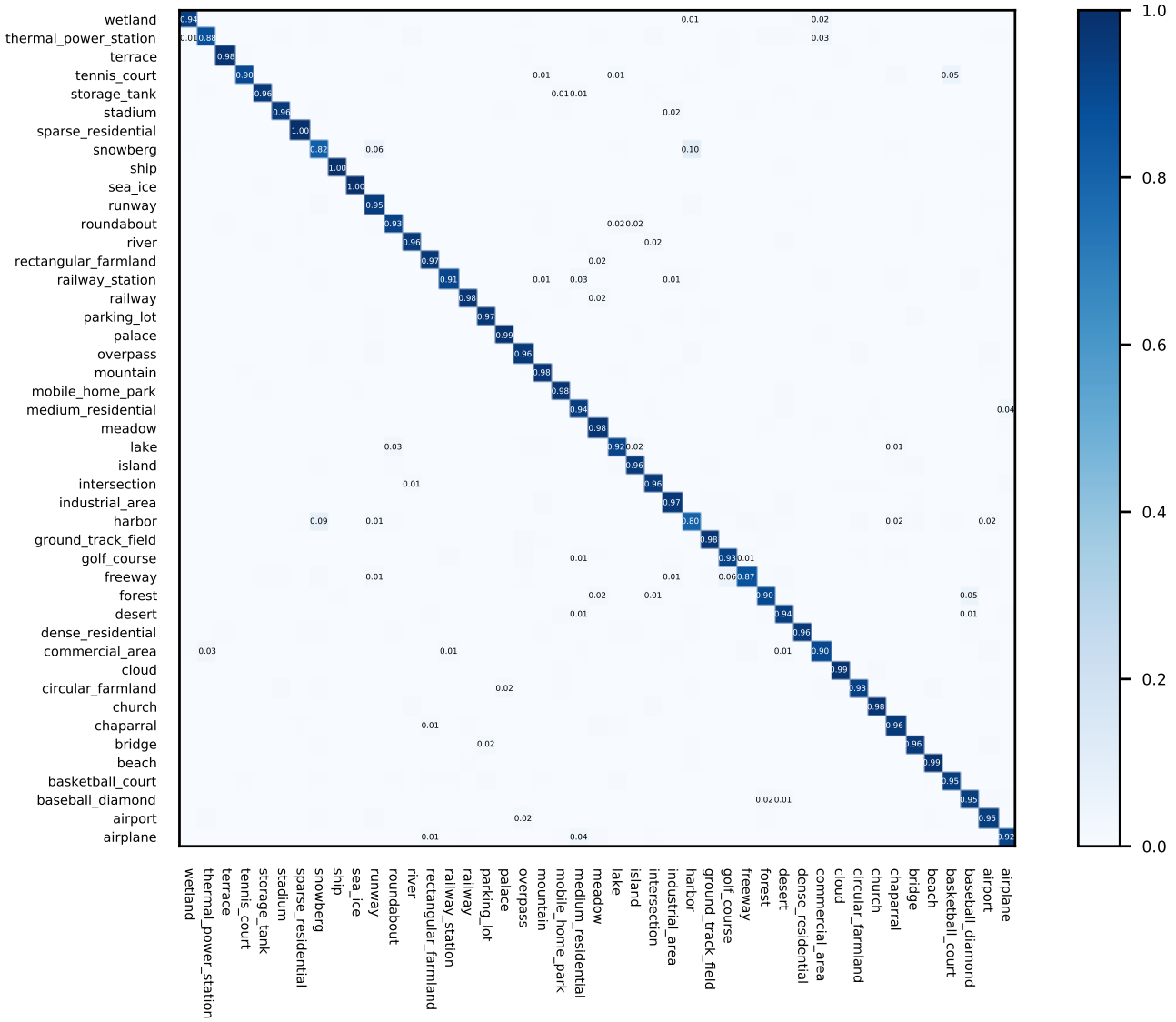

Figure 5. Confusion matrices produced by the proposed approach. We set the training ratio as 80% on the NWPU-RESISC45 dataset.

Ablation Study

Visualization. To demonstrate the effectiveness of the proposed method, we further visualize the features extracted from our method using the t-SNE algorithm [44]. We employ the t-SNE algorithm to embed high-dimensional features into two-dimensional space. It is convenient for us to measure the superiority of our approach by inspecting the derived clusters from the outputs of t-SNE.

To be specific, we randomly sample 11 scenes from the NWPU-RESISC45 dataset and embed them from the high-dimensional space into the two-dimensional space. We visualize the feature representation both with and without the proposed distillation loss function. As can be seen from Figure 7, (a) is the standard back-propagation training manner (ResNet-34 is trained individually), (b) is the method [8] based on the traditional distillation loss, and (c) is our pair-wise similarity KD method. We visualize the high-dimensional features by the t-SNE algorithm to measure whether the large within-class diversity and high between-class similarity are effectively reduced or not. Compared with (a) the student model, trained individually, and (b) the method based on traditional distillation loss, we can observe that in our method the cluster with the same category becomes closer, while clusters with different

categories become much separable. This means that the large within-class diversity and high between-class similarity are effectively reduced.

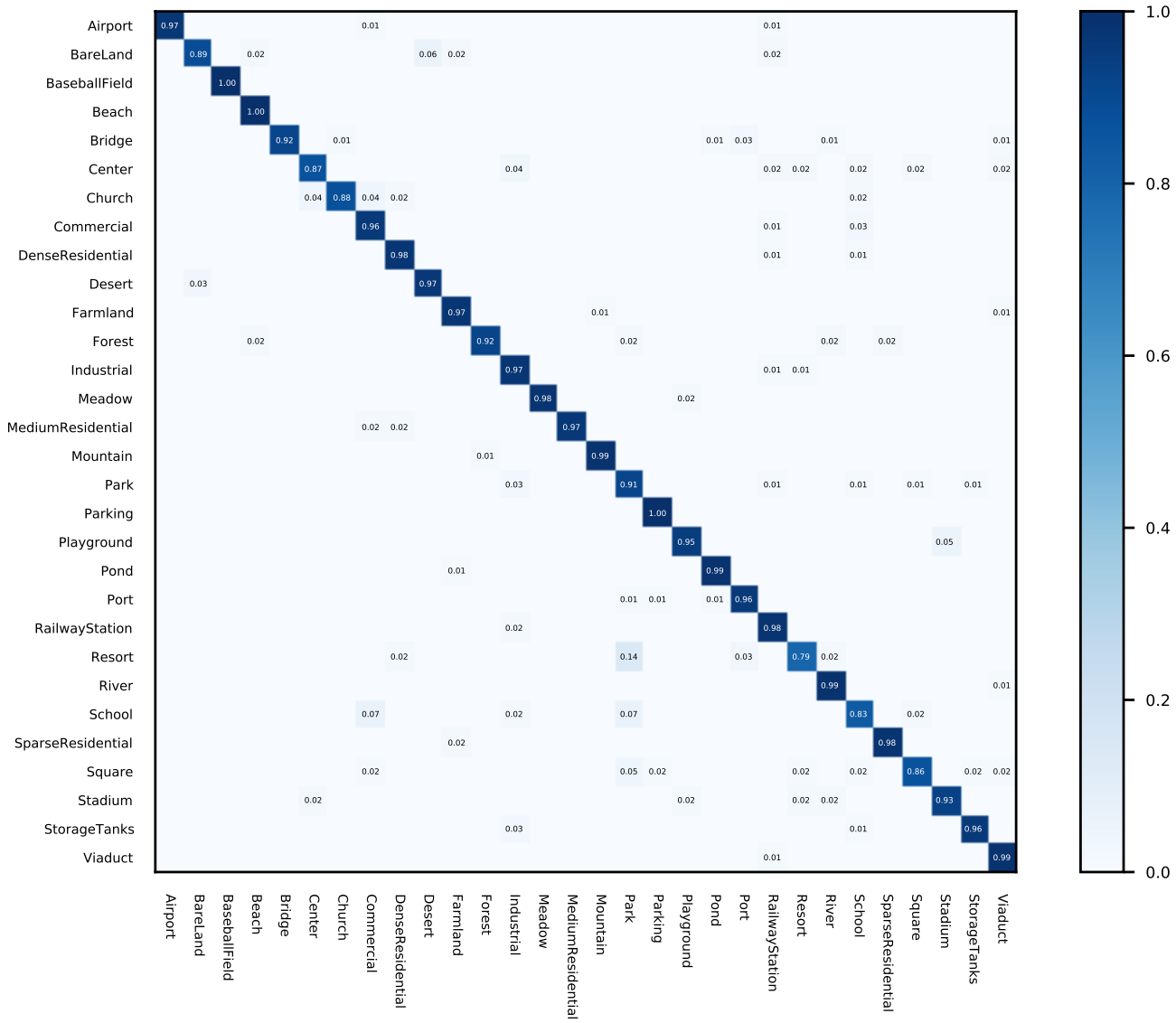

Figure 6. Confusion matrices produced by the proposed approach. We set the training ratio as 80% on the AID dataset.

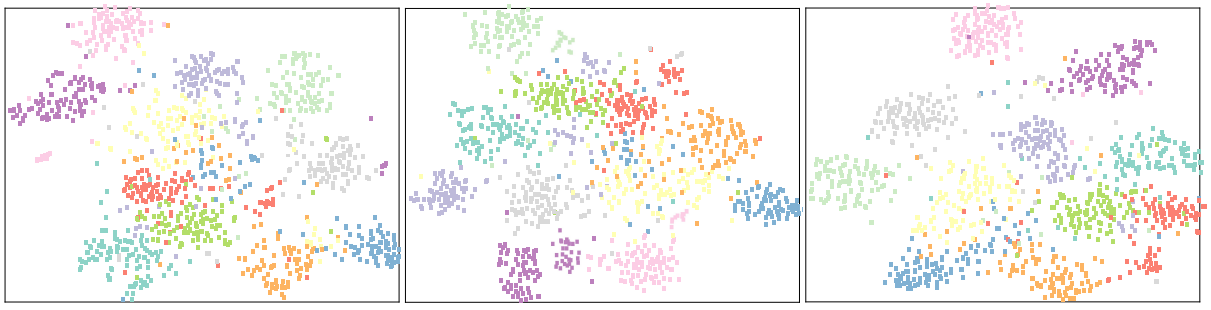

Figure 7. The 2D scatter diagram from the t-SNE algorithm on the NWPU-RESISC45 dataset with an 80% training ratio. (**a**) The student model, trained individually. (**b**) The student model, trained based on traditional distillation loss. (**c**) Our proposed method.

Trade-off between compression ratio and accuracy. We present the FLOPs and Params of the model in Table 3. We find that the proposed approach not only obtains a 0.279% improvement compared with the teacher model, but also achieves a $\times$2.03 compression rate. Thus, we obtain an efficient and compact student model for RSI scene classification.

Table 3. Experiments on the NWPU-RESISC45 dataset with an 80% training ratio. ResNet-101 and ResNet-34 are adopted as the teacher model and student model, respectively. We repeat the experiments with 5 different seeds and compute the median of 5 runs as the overall accuracy (%). The standard deviation over 5 runs is reported. FLOPs means the floating point operations, in units of 10^8 FLOPs. Params means the model size (10^7 M).

Method	Model	FLOPs	Params	NWPU-RESISC45
Student	ResNet-34	36.64	2.18	94.021 $\pm$ 0.27
KD	ResNet-34	36.64	2.18	94.576 $\pm$ 0.32
AT	ResNet-34	36.64	2.18	94.024 $\pm$ 0.16
SP	ResNet-34	36.64	2.18	95.313 $\pm$ 0.22
RKD	ResNet-34	36.64	2.18	95.297 $\pm$ 0.66
Ours	ResNet-34	36.64	2.18	**95.673 $\pm$ 0.28**
Teacher	Resnet-101	78.01	4.44	95.394 $\pm$ 0.21

Multiple model series for comparison. Table 4 summarizes the teacher and student networks' configuration. In addition to the ResNet series models, we use multiple models for comparison to further illustrate the generalization ability of the method. For the NWPU-RESISC45 dataset, we employ the Wide Residual Network (WRN) as the network architecture. Furthermore, WRN-40-1/WRN-16-1 is used for the teacher/student combination. For the UC Merced Land-Use dataset, we employ the VGG as the network architecture. VGG-13/VGG-8 is used for the teacher/student combination. As can be seen from Table 5, the proposed approach achieves 96.05% and 94.86% performances on these two datasets, respectively. It can also verify the generalization ability of the method.

Table 4. The network configurations for the teacher and student models in our experiments.

Layer Type	Output Size	ResNet-34	ResNet-101
Conv1	112 $\times$ 112	7 $\times$ 7, 64, stride 2	7 $\times$ 7, 64, stride 2
Conv2	56 $\times$ 56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
Conv3	28 $\times$ 28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$
Conv4	14 $\times$ 14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$
Conv5	7 $\times$ 7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1 $\times$ 1	average pool, 1000-d fc, softmax	

Table 5. Overall accuracy (%) on the NWPU-RESISC45 and UC Merced Land-Use datasets. We employ different model series for comparison. WRN-40-1/WRN-16-1 is used for the teacher/student combination on the NWPU-RESISC45 dataset. Furthermore, VGG-13/VGG-8 is used for the teacher/student combination on the UC Merced Land-Use dataset. The most recent methods are compared in our experiments. The best result in each experiment is shown in bold.

Dataset	Model (S/T)	Training Ratio	Baseline	KD [8]	AT [33]	SP [36]	RKD [35]	Ours	Teacher
NWPU-RESISC45	WRN-16-1 WRN-40-1	80%	94.41	94.81	94.97	94.91	95.03	**96.05**	95.67
UC Merced Land-Use	VGG-8 VGG-13	80%	93.58	93.84	93.93	94.05	93.98	**94.86**	94.45

Impact of hyper-parameters on the training. Table 6 illustrates the impact of hyper-parameters α, β, and γ on the training process. We conduct the experiments on the NWPU-RESISC45 dataset. We employ ResNet-101 as the teacher network and ResNet-34 as the student network. The default values of α, β, and γ are 0.1, 1, and 0.8, respectively. In case of $\alpha \in (0.01, 0.3)$, the other two hyper-parameters maintain their default values. We observed that the large value of α causes the overall accuracy to deteriorate rapidly. Overall, the OA is insensitive when the hyper-parameters change within an appropriate range.

Table 6. Impact of hyper-parameters α, β, and γ on the NWPU-RESISC45 dataset. We adopt ResNet-101 as the teacher network and ResNet-34 as the student network.

Hyper-parameter	0.01	0.05	α 0.1	0.2	0.3
Overall Accuracy (%)	93.962	94.187	95.673	93.143	90.067
Hyper-parameter	0.1	0.5	β 1	1.25	1.5
Overall Accuracy (%)	91.087	93.269	95.673	93.641	92.013
Hyper-parameter	0.1	0.5	γ 0.8	1	1.5
Overall Accuracy (%)	90.187	92.732	95.673	92.450	91.813

6. Conclusions

In this work, we introduce a pair-wise similarity knowledge distillation approach for RSI scene classification. The purpose is to obtain an efficient and compact model, which can be easily deployed on edge-computing devices for RSI scene classification. The proposed distillation loss could address the challenge of large intra-class variations and high inter-class similarities for a lightweight network by distilling more discriminative knowledge from the teacher network. Specifically, the proposed distillation loss involves two main terms, i.e., (1) training the student to produce similar incorrect outputs by self-distillation to reduce the intra-class variations, and (2) restraining the inter-class similarities by matching the outputs between the teacher and student model. We conduct comprehensive experiments on three popular RSI scene classification datasets, and compare our approach with the state-of-the-art KD approaches. The experimental results clearly verify the effectiveness of the proposed approach.

Author Contributions: Conceptualization, H.Z. and X.S.; methodology, H.Z.; software, H.Z.; validation, H.Z., X.S. and J.D.; formal analysis, H.Z.; investigation, H.Z. and F.G.; resources, H.Z.; data curation, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z.; visualization, H.Z.; supervision, X.S. and F.G.; project administration, J.D. and F.G.; funding acquisition, J.D.

Furthermore, J.D. and X.S. contributed to the manuscript equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 61971388, U1706218 and 41576011, Alexander von Humboldt Foundation and the Key Natural Science Foundation of Shandong Province (grant number ZR2018ZB0852).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from the corresponding authors upon request.

Acknowledgments: We acknowledge the support of the National Natural Science Foundation of China (grant numbers 61971388, U1706218, 41576011), Alexander von Humboldt Foundation and the Key Natural Science Foundation of Shandong Province (grant number ZR2018ZB0852).

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

References

1. Ghazouani, F.; Farah, I.R.; Solaiman, B. A Multi-Level Semantic Scene Interpretation Strategy for Change Interpretation in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8775–8795. [CrossRef]
2. Longbotham, N.; Chaapel, C.; Bleiler, L.; Padwick, C.; Emery, W.J.; Pacifici, F. Very High Resolution Multiangle Urban Classification Analysis. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1155–1170. [CrossRef]
3. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
4. Zhang, F.; Du, B.; Zhang, L. Scene Classification via a Gradient Boosting Random Convolutional Network Framework. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1793–1802. [CrossRef]
5. Minetto, R.; Pamplona Segundo, M.; Sarkar, S. Hydra: An Ensemble of Convolutional Neural Networks for Geospatial Land Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6530–6541. [CrossRef]
6. Han, S.; Pool, J.; Tran, J.; Dally, W.J. Learning both Weights and Connections for Efficient Neural Network. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 1135–1143.
7. Novikov, A.; Podoprikhin, D.; Osokin, A.; Vetrov, D.P. Tensorizing Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 442–450.
8. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *Comput. Sci.* **2015**, *14*, 38–39.
9. Buciluǎ, C.; Caruana, R.; Niculescu-Mizil, A. Model Compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 535–541.
10. Lei, J.B.; Caruana, R. Do Deep Nets Really Need to be Deep? *Adv. Neural Inf. Process. Syst.* **2013**, *27*, 2654–2662.
11. Tian, L.; Wang, Z.; He, B.; He, C.; Wang, D.; Li, D. Knowledge Distillation of Grassmann Manifold Network for Remote Sensing Scene Classification. *Remote Sens.* **2021**, *13*, 4537. [CrossRef]
12. Zhang, R.; Chen, Z.; Zhang, S.; Song, F.; Zhang, G.; Zhou, Q.; Lei, T. Remote Sensing Image Scene Classification with Noisy Label Distillation. *Remote Sens.* **2020**, *12*, 2376. [CrossRef]
13. Liu, B.Y.; Chen, H.X.; Huang, Z.; Liu, X.; Yang, Y.Z. ZoomInNet: A Novel Small Object Detector in Drone Images with Cross-Scale Knowledge Distillation. *Remote Sens.* **2021**, *13*, 1198. [CrossRef]
14. Chai, Y.; Fu, K.; Sun, X.; Diao, W.; Yan, Z.; Feng, Y.; Wang, L. Compact Cloud Detection with Bidirectional Self-Attention Knowledge Distillation. *Remote Sens.* **2020**, *12*, 2770. [CrossRef]
15. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
16. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
17. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the Sigspatial International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; p. 270.
18. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.S.; Zhang, L. Bag-of-Visual-Words Scene Classifier with Local and Global Features for High Spatial Resolution Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 747–751. [CrossRef]
19. Zhao, L.J.; Tang, P.; Huo, L.Z. Land-Use Scene Classification Using a Concentric Circle-Structured Multiscale Bag-of-Visual-Words Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4620–4631. [CrossRef]
20. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2175–2184. [CrossRef]
21. Lu, X.; Zheng, X.; Yuan, Y. Remote sensing scene classification by unsupervised representation learning. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5148–5157. [CrossRef]

22. Fan, J.; Chen, T.; Lu, S. Unsupervised feature learning for land-use scene recognition. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2250–2261. [CrossRef]
23. Romero, A.; Gatta, C.; Camps-Valls, G. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 1349–1362. [CrossRef]
24. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
25. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
26. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1155–1167. [CrossRef]
27. Hua, Y.; Mou, L.; Zhu, X.X. Relation Network for Multilabel Aerial Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4558–4572. [CrossRef]
28. Chen, G.; Zhang, X.; Tan, X.; Cheng, Y.; Dai, F.; Zhu, K.; Gong, Y.; Wang, Q. Training Small Networks for Scene Classification of Remote Sensing Images via Knowledge Distillation. *Remote Sens.* **2018**, *10*, 719. [CrossRef]
29. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
30. Yang, J.; Shen, X.; Xing, J.; Tian, X.; Li, H.; Deng, B.; Huang, J.; Hua, X. Quantization Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 7308–7316.
31. Ba, J.; Caruana, R. Do Deep Nets Really Need to be Deep? In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2654–2662.
32. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Bengio, Y. FitNets: Hints for Thin Deep Nets. In Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.
33. Zagoruyko, S.; Komodakis, N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, 24–26 April 2017.
34. Zhang, H.; Hu, Z.; Qin, W.; Xu, M.; Wang, M. Adversarial co-distillation learning for image recognition. *Pattern Recognit.* **2021**, *111*, 107659. [CrossRef]
35. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3967–3976.
36. Tung, F.; Mori, G. Similarity-preserving knowledge distillation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1365–1374.
37. Peng, B.; Jin, X.; Li, D.; Zhou, S.; Wu, Y.; Liu, J.; Zhang, Z.; Liu, Y. Correlation Congruence for Knowledge Distillation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; pp. 5006–5015.
38. Zhao, H.; Sun, X.; Dong, J.; Dong, Z.; Li, Q. Knowledge distillation via instance-level sequence learning. *Knowl. Based Syst.* **2021**, *233*, 107519. [CrossRef]
39. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
40. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
41. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
43. Chen, L.; Wang, D.; Gan, Z.; Liu, J.; Henao, R.; Carin, L. Wasserstein Contrastive Representation Distillation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; pp. 16296–16305.
44. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.