remote sensing

MDPI

Article

Masked Image Modeling Auxiliary Pseudo-Label Propagation with a Clustering Central Rectification Strategy for Cross-Scene Classification

Xinyi Zhang, Yin Zhuang *, Tong Zhang, Can Li and He Chen

The National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing, Beijing Institute of Technology, Beijing 100081, China; 3220210698@bit.edu.cn (X.Z.); chenhe@bit.edu.cn (H.C.)
* Correspondence: yzhuang@bit.edu.cn

Abstract: Cross-scene classification focuses on setting up an effective domain adaptation (DA) way to transfer the learnable knowledge from source to target domain, which can be reasonably achieved through the pseudo-label propagation procedure. However, it is hard to bridge the objective existing severe domain discrepancy between source and target domains, and thus, there are several unreliable pseudo-labels generated in target domain and involved into pseudo-label propagation procedure, which would lead to unreliable error accumulation to deteriorate the performance of cross-scene classification. Therefore, in this paper, a novel Masked Image Modeling Auxiliary Pseudo-Label Propagation called MIM-AP^2 with clustering central rectification strategy is proposed to improve the quality of pseudo-label propagation for cross-scene classification. First, in order to gracefully bridge the domain discrepancy and improve DA representation ability in-domain, a supervised class-token contrastive learning is designed to find the more consistent contextual clues to achieve knowledge transfer learning from source to target domain. At the same time, it is also incorporated with a self-supervised MIM mechanism according to a low random masking ratio to capture domain-specific information for improving the discriminability in-domain, which can lay a solid foundation for high-quality pseudo-label generation. Second, aiming to alleviate the impact of unreliable error accumulation, a clustering central rectification strategy is designed to adaptively update robustness clustering central representations to assist in rectifying unreliable pseudo-labels and learning a superior target domain specific classifier for cross-scene classification. Finally, extensive experiments are conducted on six cross-scene classification benchmarks, and the results are superior to other DA methods. The average accuracy reached 95.79%, which represents a 21.87% improvement over the baseline. This demonstrates that the proposed MIM-AP^2 can provide significantly improved performance.

Keywords: cross-scene classification; domain adaptation; masked image modeling; pseudo-label

check for updates

Citation: Zhang, X.; Zhuang, Y.; Zhang, T.; Li, C.; Chen, H. Masked Image Modeling Auxiliary Pseudo-Label Propagation with a Clustering Central Rectification Strategy for Cross-Scene Classification. *Remote Sens.* **2024**, *16*, 1983. https://doi.org/10.3390/rs16111983

Academic Editor: Pedro Melo-Pinto

Received: 25 January 2024
Revised: 16 May 2024
Accepted: 20 May 2024
Published: 31 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remote sensing (RS) scene classification enables effective identification and partitioning of massive remote sensing data categories, widely applied in land cover and land use, natural disaster monitoring, urban planning, and other fields [1–3]. Deep-learning-based methods achieve satisfactory classification performance only when training and testing data adhere to the condition of independent and identically distributed (i.i.d.). However, actual testing data may originate from different sensors and be influenced by factors such as geographical location, lighting conditions, and seasonal variations. For instance, variations in natural environment, climate, economic development levels, and population density across different regions result in differing characteristics of land cover and land use. Moreover, disparities in geological environments between regions may yield distinct features for the same type of natural disaster. Consequently, there would exist significant domain discrepancy between labeled training data (i.e., the source domain) and unlabeled

testing data (i.e., the target domain), leading to out-of-distribution (o.o.d.) scenarios. These domain discrepancies hinder models from adapting to a new target domain, resulting in inferior performance, thus constraining a range of practical applications. Considering the classification performance and generalization ability of models, cross-scene classification aims to establish a more effective domain adaptation (DA) way to finely transfer the learnable knowledge from source to target domain to achieve high-precision scene classification under o.o.d. scenarios. Recently, pseudo-labeling stands out as a widely employed technique in the realm of semi-supervised learning (SSL) for RS scene classification [4–7], and it can reasonably transfer the knowledge from labeled data (i.e., the source domain) to unlabeled data (i.e., the target domain) by pseudo-label propagation. However, due to the severe objective domain discrepancy, when transferring the learnable knowledge from source domain to target domain, unreliable pseudo-labels are easily generated in target domain and inevitably involved in pseudo-label propagation during model training, thus leading to unreliable error accumulation to deteriorate the performance for cross-scene classification. Consequently, how to effectively bridge the domain discrepancy to improve the DA representation ability and further enhance the quality of pseudo-label propagation becomes a critical issue for cross-scene classification.

Recently, there have been a lot of research papers dedicated to exploring the DA representation and pseudo-label propagation for DA [8–15]. First, several methods [8–10] focused on improving the DA representation ability and extracting transferable features in order to achieve high-quality pseudo-label generation in target domain. Zhu et al. [8] introduced attention mechanism and multi-scale strategy to optimize feature extraction, and employed pseudo-labels of the target domain to align conditional distribution. Yang et al. [9] proposed an attention-based dynamic alignment and dynamic distribution adaptation (ADA-DDA) approach that uses target domain pseudo-labels to achieve conditional alignment and dynamically balances the relative importance of marginal and conditional distributions. Besides, some studies focus on improving the discriminability in-domain. For example, in [10], the cross-domain distribution is dynamically aligned from local pixel-level and global image-level, where the alignment of the global semantic can bridge domain discrepancy while the alignment of the local information can maintain its discriminability in-domain. Second, besides improving DA representation ability for pseudo-label generation, other methods [11–15] paid attention to pseudo-label propagation procedure, which aims to select high-quality pseudo-labels to alleviate unreliable error accumulation from pseudo-label propagation. Zhang et al. [11] proposed a feature enhancement network (DFENet) that utilizes a fixed threshold to selectively filter pseudo-labels to cooperate with their designed self-motivated learning, facilitating the potential ambiguous discriminative learning of features in target domain. Huang et al. [12] also utilized a fixed threshold to select pseudo-labels with high confidence for class center clustering in target domain, thus achieving cross-domain sample-class alignment. Liang et al. [13] proposed the self-training adversarial DA (STADA) method to make the generator produce pseudo-labels in target domain and then select high-quality pseudo-labels by discriminator for classifier adaptation, achieving RS semantic segmentation. Zhu et al. [14] designed an entropy-based pseudo-label filtering strategy to obtain high confidence pseudo-labels in the target domain for updating the invariant domain-level prototype information. Sun et al. [15] proposed a gradual domain adaptation framework with pseudo-label denoising for SAR target recognition. This approach used image similarity and k-Nearest Neighbor (KNN) algorithm to eliminate part of the unreliable pseudo-labels, selecting only the reliable ones for fine-tuning the classifier and narrowing the category distribution difference.

Although the above-mentioned studies have shown that improving the DA representation ability and the quality of pseudo-label propagation can effectively achieve the better cross-domain transfer learning for cross-scene classification, there are two critical issues that still remain to be properly addressed. First, unlike natural scene images, RS images contain rich color and spatial texture information, among which key spatial contextual clues are transferable. Regrettably, some methods [12,15] do not consider the acquisition

of spatial contextual information. Several recent studies [8–11] have been realized that spatial contextual information possesses a powerful transferability, and they considered using attention mechanism to capture the spatial contextual information from dual-domain RS scenes. However, due to complicated spatial layouts of RS scenes and very limited receptive fields of convolution neural network (CNN), a lot of irrelevant and redundant information would be inevitably extracted into transferable feature representation. In these studies [8–11], the attention mechanism only can suppress some irrelevant and redundant information and cannot guarantee the effectiveness of information capture. Hence, they yield only marginal improvements. In addition, to enhance the DA representation capability, merely concentrating on transferable features is insufficient to ensure the generation of high-quality pseudo-labels in the target domain. Because of the absence of supervised information in the target domain, a lot of domain-specific information is inevitably lost, thereby diminishing their discriminative capabilities in-domain. This would hinder the generation of high-quality pseudo-labels in target domain and directly constraint further improvement of cross-scene classification performance. Notably, how to establish an effective DA representation with robust transferability and discriminability in-domain is very important for cross-scene classification, which can effectively bridge the domain discrepancy and further improve the representation ability for DA. Thus, there is still a need for further exploration. Second, although [11–15] all proposed effective ways to select the high-quality pseudo-label, they neglected to explore the valuable information contained in unreliable pseudo-labeled data, which can be rectified to further facilitate the target domain specific classifier learning. Therefore, it is necessary to further refine the strategy and fully leverage unlabeled data with low-confidence pseudo-labels to unlock the full potential of pseudo-label propagation for cross-scene classification.

In this paper, a novel Masked Image Modeling Auxiliary Pseudo-Label Propagation called MIM-AP^2 with clustering central rectification strategy is proposed to improve the DA representation ability and the quality of pseudo-label propagation for cross-scene classification. First, aiming to improve the DA representation ability under severe domain discrepancy and considering a powerful spatial contextual modeling ability of vision transformer (ViT), a supervised class-token contrastive learning is designed to enable the proposed MIM-AP^2 to learn more consistent contextual clues. It can gracefully bridge the domain discrepancy to transfer the learnable knowledge from source to target domain. Subsequently, a self-supervised MIM mechanism with a low random masking ratio is developed for unlabeled data learning, and it is combining with the supervised class-token contrastive learning to further improve the DA representation ability in target domain, facilitating high-quality pseudo-label generation, where a low random masking ratio for unlabeled data can be seen as a specific data augmentation to enlarge the dual-domain consistent representation space and encourage the search for more consistent contextual clues. Meanwhile, during the DA representation learning, the MIM mechanism can reveal that when each specific random masking pixels have been well reconstructed, the domain-specific information would be captured and latent in representation to improve the discriminability in-domain. As a result, the class-tokens exhibit both transferability and domain-specific discriminability in target domain, which is a solid foundation for the generation of high-quality pseudo-labels. Second, a clustering central rectification strategy is proposed, which adaptively update the reliable clustering central representations referring to the classifier learned from source domain to effectively rectify the unreliable pseudo-labels. By following the proposed strategy, unlabeled data in the target domain can be fully utilized to enhance the quality of pseudo-label propagation. Naturally, according to a powerful DA representation ability and high-quality pseudo-label propagation, a target domain specific classifier can be well learned to achieve the superior cross-scene classification performance. Finally, extensive experiments are conducted on multiple constructed RS cross-scene classification benchmarks, and the results indicate that the proposed MIM-AP^2 is superior to the state-of-the-art (SOTA) methods.

In summary, the main contributions of our study can be outlined as follows:

(1) A novel DA framework called MIM-AP^2 is proposed for cross-scene classification in the RS domain, which can not only achieve a powerful DA representation ability from source to target domain even if under severe domain discrepancy, but also can provide a high-quality pseudo-label propagation for learning a superior target domain specific classifier to achieve the SOTA cross-scene classification performance.
(2) A new DA representation way is proposed based on a supervised class-token contrastive learning incorporated with self-supervised MIM mechanism, which not only can utilize a low random masking ratio as a specific data augmentation for unlabeled data to encourage the search for more consistent contextual clues between source and target domain, but also adopt random masking pixel reconstruction to reveal the capture of the domain-specific information for further improving the discriminability in-domain. Thus, the DA representation with transferability while maintaining discriminability is set up, which is a solid foundation for high-quality pseudo-label generation.
(3) A novel clustering central rectification strategy is proposed to assist in setting up a specific classifier in target domain based on a powerful DA representation, which can effectively rectify unreliable pseudo-label based on adaptively updating the reliable clustering central representations by referring to the classifier learned from source domain. Thus, by fully excavating valuable information from unlabeled data, a superior target domain specific classifier can be constructed for cross-scene classification.

2. Related Work

2.1. Domain Adaptation

Domain adaptation (DA) mainly solves the problem of knowledge transfer between source and target domains with severe data distribution discrepancy. Its core objective is to align the data distribution of the source domain with that of the target domain through the acquisition of domain-invariant feature representation, so that the model trained on the labeled source domain can be directly migrated to the unlabeled target domain without a substantial performance decline. Presently, the commonly employed DA methods are falling into two categories: (1) discrepancy metric-based approaches and (2) adversarial-based approaches.

The discrepancy metric-based approaches usually design a specific metric way to measure the distribution discrepancy between source and target domains, and then minimize the metric to align the distribution of the two domains. Long et al. [16] proposed the deep adaptation network (DAN) and explored the use of multiple cores maximum mean discrepancy (MK-MMD) [17] to minimize the marginal distribution of two domains. Later, they took the joint probability distribution of features and labels into consideration [18]. In addition, Chen et al. [19] developed a new metric named maximum density divergence (MDD) and introduced it into adversarial DA framework.

The adversarial-based approaches encourage domain confusion through the minimax game between the generator and discriminator to learn the generated information as a domain-invariant representation by confusing the domain discriminator. Ganin et al. [20] took the lead in applying adversarial learning to DA, and they proposed gradient reversal layer (GRL) to train domain adversarial neural networks (DANN) for DA. Adaptive discriminant domain adaptation (ADDA) [21] employed GAN [22] loss to divide the DA optimization process into two independent objectives including the generation and discrimination. Conditional domain adversarial network (CDAN) [23] designs multilinear conditioning and aligned conditional probability distribution to promote DA representation between source and target domains. Moreover, the maximum classifier discrepancy (MCD) [24] introduces a new adversarial paradigm, which makes the features of the generator farther away from the decision boundary through the minimax game between the classifiers and generator on cross-classifier outputs discrepancy.

The aforementioned DA methods for natural images have been introduced and extensively applied to RS images cross-scene classification. Ammour et al. [25] proposed

the additional asymmetric adaptation convolution layer and reduced the domain shift by minimizing the mean-squared error (MSE) loss. Song et al. [26] proposed a subspace alignment layer (SA) and minimized Bregman matrix divergence to align the source and target domains in feature subspace. Zhu et al. [27] designed a multi-linear conditioning strategy based on adversarial learning to capture the multi-modal structures of feature distribution and achieve category alignment. Teng et al. [28], Ma et al. [29], and Zheng et al. [30] devised multiple classifiers and engage in a minimax game between the classifiers and generator, which aims to establish better decision boundaries in the target domain. In the face of complex and diverse RS images, some researchers have focused on the representation learning in DA [8]. Zhu et al. [8], Yang et al. [9], and Zhang et al. [11] used techniques such as channel attention, spatial attention, and multi-scale strategy to extract more robust features and complete information, and then realized domain alignment by discrepancy metric- or adversarial- based approaches. Niu et al. [10] used multiple discriminators to achieve image-level adaptation and pixel-level adaptation to obtain contextual semantic information and local detail features, respectively.

Moreover, owing to a powerful spatial contextual modeling ability of ViT, several studies [31–34] have been conducted on natural scenes based on ViT, leading to impressive DA performance. Yang et al. [31] employed a patch-level domain discriminator to assign weights to each patch, compelling ViT to focus on transferable and discriminative features. Xu et al. [32] and Wang et al. [33] applied self-attention and cross-attention to learn a transferable spatial contextual representation base on ViT for domain alignment. Ma et al. [34] designed a domain-oriented transformer (DOT) that has two individual classification tokens to learn different domain-oriented representations. Consequently, considering the importance of spatial contextual information, our focus is centered on establishing better DA representations based on it for cross-scene classification in the RS field.

2.2. Pseudo-Labeling

One widely used SSL technique is pseudo-labeling [35], a process that iteratively assigns pseudo-labels to unlabeled data in the target domain based on their maximum prediction scores. Subsequently, the network is retrained in a supervised manner by using the pseudo-labeled data. Essentially, DA can be regarded as a specific instance of transductive SSL, wherein labeled and unlabeled data are sampled from distinct distributions. Liang et al. [36] enhanced the quality of pseudo-labels by introducing auxiliary classifiers for target domain data, and developed the nearest centroid classifier and neighborhood aggregation. Subsequently, they utilized the nearest centroid classifier to assign pseudo-labels, achieving DA without access to the source data [37]. Gu et al. [38] designed an adversarial DA method in the spherical feature space and weighted the pseudo-labels in the target domain through the evaluation of a Gaussian-uniform mixture model. Zhang et al. [39] combined domain collaborative and adversarial learning, and adopted a self-paced strategy to select target domain pseudo-labels from easy to hard. Xu et al. [32] used cross-attention in ViT for domain alignment, and proposed a two-way center-aware labeling method to select high-quality training pairs. In the semi-supervised domain adaptation (SSDA) setting with partially annotated target domain, there are also some studies that leverage pseudo-labels. For example, Lu et al. [40] selected trusted pseudo-labels according to the maximum correlation coefficient and the 1-nearest neighbor classifier. Yu et al. [41] treated the source data as a noisy version of the ideal target data and utilize the pseudo-center of the target domain to correct the source label. The complex nature of RS images and large domain discrepancies present a formidable challenge in acquiring high-quality pseudo-labels in DA. Kwak et al. [42] and Liang et al. [13] combined adversarial training with self-training for crop type mapping and RS image segmentation, respectively. The former utilized contextual information and past crop boundary maps as additional constraints to obtain reliable pseudo-labels of the target domain, while the latter employed a label discriminator for selection. Gao et al. [43] improved the quality of pseudo-labels through contextual information to support learning multiple prototypes for each class, achieving

better alignment. Our research also focuses on pseudo-label optimization strategies to improve the quality of pseudo-label propagation for cross-scene classification.

2.3. Masked Image Modeling

Initially introduced in natural language processing, masked language modeling (MLM) involves a language modeling task wherein random percentages of tokens within the text are masked and then reconstructed from the encoding results of the remaining text [44]. Subsequent research has extended this concept from natural language processing to computer vision, incorporating the masking of diverse proportions of image patches for recovery. With the rapidly development of ViT, MIM [45–51] demonstrates robust potential in representation learning. In the wake of BERT's [44] advancements in natural language processing, Bao et al. [45] introduced BEiT, aiming to reconstruct the original visual tokens by leveraging corrupted image patches. SimMIM [46] has a moderately large masked patch size and predicts the original pixels by direct regression task. MAE [47] employs only the visible patches as input for the encoder, introducing mask tokens between the encoder and decoder. This intentional asymmetry in design notably alleviates the computational burden on the encoder. Subsequently, MultiMAE [48] harnesses the efficiency of the MAE, expanding its application to multimodal and multitask settings. MR-MAE [49] inherits the reconstruction loss from MAE and leverages advanced semantics extracted from off-the-shelf models to supervise the features of visible tokens in MAE. MixMIM [50] substitutes the masked tokens of one image with the visible tokens of another image, and reconstruct the original two images from the combined input, leading to a notable enhancement in efficiency. MaskFeat [51] takes a handmade feature descriptor, histograms of oriented gradient (HOG), as the target of reconstruction. MIM mechanism applied for ViT can make model capture the intrinsic pattern, structures, relationships and semantics by reconstructing randomly masked content, so as to establish more effective and robust feature representation. Several studies [52–54] have shown that MIM-based self-supervised learning on large-scale RS dataset can effectively enhance model representation capabilities and adapt to RS tasks. Therefore, in this article, the MIM mechanism is also explored to further improve the DA representation for cross-scene classification.

3. Methodology

3.1. Overview

First, let us denote $X_s = \left\{(x_s^i, y_s^i)\right\}_{i=1}^{n_s}$ as a source domain with n_s labeled images and $X_t = \left\{(x_t^j)\right\}_{j=1}^{n_t}$ as a target domain with n_t unlabeled images. The source and target domains are sampled from different probability distributions P and Q ($P \neq Q$), respectively, but they have the K common categories. Our objective is to train a scene classification model based on labeled and unlabeled data in source and target domains, conquering domain discrepancies to achieve high performance on unlabeled data in target domain. The proposed MIM-AP^2 for more effective cross-scene classification is shown in Figure 1. The shared weight encoder $E(\cdot)$, which consists of L stacked transformer blocks, performs feature extraction on the input images that are divided into several regular non-overlapping patches. The patches are first converted to vectors in the latent space:

$$\mathbf{z}_0 = [c; x_{in}\mathbf{E}] + \mathbf{E}_{pos} \tag{1}$$

where x_{in} stands for the input patches. A linear projection $\mathbf{E}$ is used for patch embedding, making the image patches project as visual tokens. Subsequently, a series of visual tokens are concatenated with a learnable class token c and then adding with a position embedding $\mathbf{E}_{pos}$. Each transformer block is including multi-head self-attention (MSA) and multiple layer perception (MLP), which can be written as:

$$\mathbf{z}_\ell^{'} = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \ \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \ \ell = 1 \ldots L \tag{3}$$

where $\mathrm{LN}(\cdot)$ is the layer normalization and $\mathrm{MSA}(\cdot)$ is an extension of self-attention for capturing the long-range dependencies. The shared weight decoder $D(\cdot)$ is also made up of a series of transformer blocks. First, a supervised class-token contrastive learning is designed to find more consistent contextual clues. Subsequently, a self-supervised MIM mechanism with a low random masking ratio is developed to randomly masked and removed a few patches in the target domain and reconstruct the image for unlabeled data learning. It is incorporated with the supervised class-token contrastive learning to further improve the discriminability in-domain. Second, corresponding to the proposed MIM-AP^2, a clustering central rectification strategy is also proposed to improve the quality of pseudo-label propagation, which can combine with the set up powerful DA representation to train a superior target domain specific classifier for cross-scene classification. In the following section, a more effective DA representation learning for pseudo-label generation, a clustering central rectification strategy for pseudo-label propagation and corresponding joint loss, is elaborated.

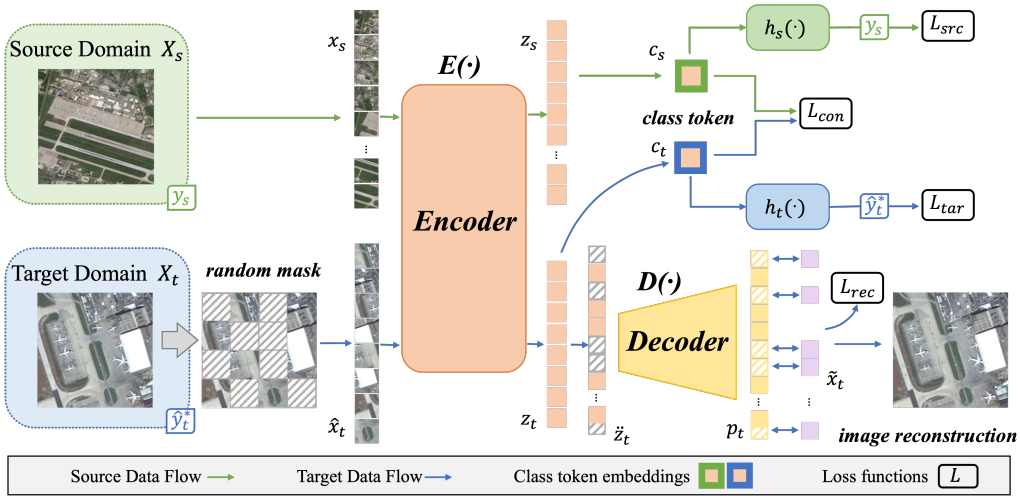


Figure 1. Overview of the proposed MIM-AP^2 framework in training phase.

3.2. Domain Adaption Representation Learning

When facing the objective domain discrepancy, how to effectively transfer the learnable knowledge and encourage the discriminative learning in the target domain is very important. In this context, ViT has garnered attention due to its powerful spatial contextual modeling capabilities. Given the considerable transferability of consistent contextual clues to bridge the domain discrepancy, ViT is considered as the dual-domain feature extractor for capturing more consistent contextual clues between labeled data X_s in the source domain and unlabeled data X_t in the target domain. The class-tokens within the encoded features aggregate global information that can be used for classification because of effective contextual clue description. Inspired by contrastive learning [55], a supervised class-token contrastive learning is designed to achieve the conditional distribution alignment under the supervision of labels in source domain and pseudo-labels in target domain, which narrows the feature distance of the same category and increases the feature distance of different categories between source and target domains. Therefore, as shown in Figure 1, class tokens c_s and c_t of the source and target domains are taken out from z_s and z_t, respectively. In this way, consistent contextual clues of the same category both in source and target domains can be found to bridge domain discrepancy and achieve knowledge transfer. In addition, due to the absence of definite supervision signal in the target domain, a lot of domain-specific information would be easily ignored, resulting in in-domain under-representation. Therefore, a self-supervised MIM mechanism with a low random masking ratio is developed for unlabeled data learning. As shown in Figure 1, z_s and z_t represent the encoded feature

from dual-domain sharing weight encoder $E(\cdot)$ according to the source domain image x_s and random masked image $\hat{x}_t$ in the target domain, respectively:

$$z_s = E(x_s), z_t = E(\hat{x}_t) \tag{4}$$

where different from MIM mechanism in MAE [47] to set up an effective representation by reconstructing whole image pixels according to a small portion of information, a low random masking ratio can be seen as a weak augmentation for unlabeled data to generate abundant random masking patterns in target domain. This helps maintain original scene information to some extent and enlarges the representation space z_t for unlabeled data, assisting the designed supervised class-token contrastive learning in seeking more consistent contextual clues between c_s and c_t. Meanwhile, the random masking pixel reconstruction at the decoder $D(\cdot)$ can be written as:

$$p_t = D(\ddot{z}_t) \tag{5}$$

It can also be seen as an evident signal, because when these specific random masking pixels are well resumed, it reveals that the domain-specific information in target domain has been well captured and latent in the z_t. Consequently, the low random masking ratio of MIM combining with the supervised class-token contrastive learning in the proposed MIM-AP^2 can set up a powerful DA representation, which makes c_t both transferability and discriminability for the generation of high-quality pseudo-labels in the target domain.

3.3. Clustering Central Rectification Strategy

Although setting up a powerful DA representation c_t in target domain can ensure the high-quality pseudo-label generation to a certain extent, there are inevitably a few unreliable pseudo-labels produced in the process of pseudo-label propagation. This can result in the accumulation of unreliable errors during model training, which would deteriorate the specific classifier learning in target domain, and thus, produce unreliable classification decision boundary for cross-scene classification. Consequently, in addition to setting up a powerful DA representation, a more reliable pseudo-label propagation strategy is also a crucial aspect. Corresponding to the proposed MIM-AP^2 as shown in Figure 1, a clustering central rectification strategy is proposed for continuously optimizing the generated pseudo-labels during pseudo-label propagation, as illustrated in Figure 2.

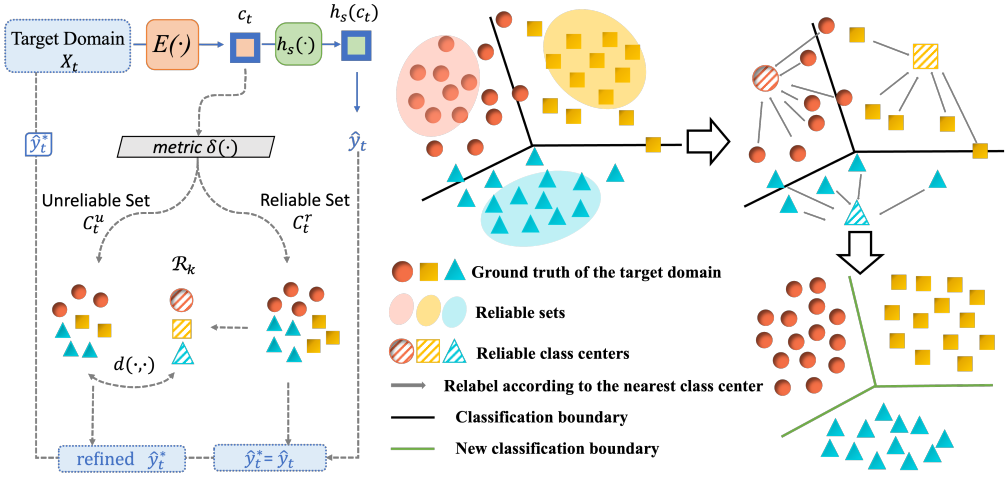


Figure 2. Overview of clustering central rectification strategy.

To prevent the dual-domain sharing classifier from leading to the unreliable scene classification in target domain, two classifiers, i.e., $h_s(\cdot)$ and $h_t(\cdot)$, are designed for source and target domains, respectively. Considering that the source domain classifier is not supervised by pseudo-labels in target domain, it is not affected by the previous unreliable

pseudo-label noise. Therefore, as shown on the left in Figure 2, the encoded target class tokens c_t are fed into the source domain classifier $h_s(\cdot)$ to obtain the prediction scores $h_s(c_t)$ for reliability evaluation, which is aiming to avoid the effect of noisy iterations. The initial predictions are $\hat{y}_t = \underset{k=1,\ldots,K}{\arg\max}(h_s(c_t))^{(k)}$. Let $c_{t,k}$ denotes the corresponding class-tokens of all samples in target domain with prediction $\hat{y}_t = k$, and the corresponding set is called $C_{t,k}$. C_t represents the sets of class-tokens corresponding to the encoded features of all target samples. Next, C_t is divided into a reliable subset C_t^r and an unreliable subset C_t^u. Most of the previous methods are relying on softmax-based scores to select high confidence pseudo-labels, while in out-of-distribution detection [56–59], energy scores are widely used and have been shown to be better distinguish in- and out-of-distribution samples than softmax-based scores. We believe that reliable and unreliable pseudo-labels can be analogous to in- and out-of-distribution samples in the problem of out-of-distribution detection, so a metric $\delta(\cdot)$ based on energy scores is leveraged to evaluate the reliability of all target samples by Equation (6):

$$\delta(c_t) = log \sum_k \exp(h_s(c_t)^{(k)}) \tag{6}$$

Then, reliable set $C_t^r = \cup_k C_{t,k}^r$ is constructed, where $C_{t,k}^r = \left\{ c_t | c_t \in C_{t,k}, \delta(c_t) \geq \overline{\delta(c_{t,k})} \right\}$ while the rest is selected in unreliable set C_t^u.

As shown on the left in Figure 2, for the samples in C_t^r, their predictions can be trusted and $\hat{y}_t$ are adopted directly as pseudo-labels $\hat{y}_t^*$. The reliable subset contains high quality pseudo-labels, so the centroids of reliable representations can be considered as more stable anchor points, intended for the correction of unreliable pseudo-labels. Consequently, the reliable class centers $\mathcal{R}_k$ can be derived by Equation (7). For the samples in C_t^u, they are relabeled according to their nearest class centers from reliable subsets according to Equation (8):

$$\mathcal{R}_k = \frac{1}{\left| C_{t,k}^r \right|} \sum_{c_{t,k} \in C_{t,k}^r} c_{t,k} \tag{7}$$

$$\hat{y}_t^* = \begin{cases} \hat{y}_t, & c_t \in C_t^r \\ \arg\min_k d(c_t, \mathcal{R}_k), & c_t \in C_t^u \end{cases} \tag{8}$$

where $|\cdot|$ represents the number of elements in the set and $d(\cdot, \cdot)$ denotes the cosine distance measure. The schematic diagram on the right side of Figure 2 illustrates the process of rectifying unreliable pseudo-labels through the feature clustering of reliable pseudo-labels. The introduced clustering central rectification strategy judiciously leverages trusted pseudo-labels, diminishing noise labels within the unreliable subset after label reassignment. This process enhances the overall quality of the pseudo-labels.

The clustering central rectification strategy is executed every certain number of steps during the training process. In Algorithm 1, the procedure of the proposed clustering central rectification strategy was presented. Through continuous iterative training and the updating of pseudo-labels, the proportion of unreliable subsets would diminish gradually, leading to a steady improvement in accuracy.

Algorithm 1 Algorithm of Clustering Central Rectification Strategy

Input: Class tokens c_t of the target domain; source domain classifier $h_s(\cdot)$; metric $\delta(\cdot)$; category number K

Output: Final pseudo-labels $\hat{y}_t^*$ of the target domain

Acquire initial predictions $\hat{y}_t = \arg\max_{k=1,\ldots,K}(h_s(c_t))^{(k)}$; Calculate $\delta(c_t)$ by Equation (6);

for $i = 1$ to K **do**

Calculate the average of $\delta(c_t)$ which has initial predictions $\hat{y}_t = k$ to obtain $\overline{\delta(c_{t,k})}$;

Pick $c_{t,k}$ which $\delta(c_{t,k}) \geq \overline{\delta(c_{t,k})}$ into reliable set $C_{t,k}^r$;

Compute the reliable class center $\mathcal{R}_k$ by Equation (7);

end for

Put the rest c_t in unreliable set C_t^u;

Obtain final pseudo-labels $\hat{y}_t^*$ by Equation (8).

3.4. Total Loss

Corresponding to the proposed MIM-AP^2 with clustering central rectification strategy, the objective jointly loss includes four components as illustrated in Equation (9):

$$L = L_{src} + L_{tar} + \lambda_1 L_{con} + \lambda_2 L_{rec} \tag{9}$$

where λ_1 and λ_2 are hyperparameters. The detailed calculation of each loss is described below.

(1) L_{src}: L_{src} is the cross-entropy loss of classifier $h_s(\cdot)$ supervised by source labels y_s:

$$L_{src} = \frac{1}{n_s} \sum_{i=0}^{n_s} H(h_s(c_s^i), y_s^i) \tag{10}$$

where $H(\cdot, \cdot)$ is the cross-entropy loss function.

(2) L_{tar}: L_{tar} is the cross-entropy loss of classifier $h_t(\cdot)$ supervised by target pseudo-labels $\hat{y}_t^*$:

$$L_{tar} = \frac{1}{n_t} \sum_{j=0}^{n_t} H(h_t(c_t^j), \hat{y}_t^{j*}) \tag{11}$$

(3) L_{con}: The contrastive learning-based alignment loss L_{con} can be expressed as follows:

$$L_{con} = \sum_{c_t^j \in C_t} -\frac{1}{|P_j|} \sum_{c_s^+ \in P_j} \left[log \frac{\exp(c_t^j \cdot c_s^+ / \tau)}{\sum_{c_s \in C_s} \exp(c_t^j \cdot c_s / \tau)} \right] \tag{12}$$

where C_s represents the sets of class tokens corresponding to the encoded features of all source samples and P_j is the set of class tokens c_s^+ of positive samples x_s^+ in source domain for a target sample x_t^j, i.e., $y_s = \hat{y}_t^{j*}$.

(4) L_{rec}: The reconstruction loss L_{rec} in the target domain is calculated as the mean square error (MSE) between the reconstructed image and the normalized original pixels, and this calculation is only performed on the masked patches:

$$L_{rec} = \frac{1}{|X_t|} \sum_{x_t^j \in X_t} \frac{1}{|M|} \parallel \left(p_t^j - \tilde{x}_t^j \right)_M \parallel_2^2 \tag{13}$$

where M denotes the set of masked pixels and $\tilde{x}_t$ is the corresponding normalized pixel values of x_t.

4. Experiments and Results

4.1. Datasets

In this section, three distinct open-source RS scene classification datasets, i.e., NWPU-RESISC45 [60], AID [61], and UC Merced Land-Use [62], were chosen to construct cross-scene classification tasks for assessing the performance of the proposed MIM-AP^2 framework. The specifics of each dataset are elaborated upon below:

- The NWPU-RESISC45 dataset stands as a large-scale open-source benchmark in the realm of RS scene classification, meticulously crafted by Northwestern Polytechnical University. Sourced from the Google Earth service, this dataset encompasses 31,500 images that span 45 distinct classes of RS scenes. Spatial resolutions across the dataset range from 0.2 m to 30 m. Notably, each class is represented by 700 images, each measuring 256 $\times$ 256 pixels in size.
- The AID dataset is derived from the Google Earth service, comprising 10,000 aerial images spanning 30 scene classes. The quantity of sample images varies, ranging from 220 to 420 across different aerial scene categories. Each aerial image is sized at 600 $\times$ 600 pixels, with a spatial resolution ranging from 0.5 m to 8 m.
- The UC Merced Land-Use dataset is composed of 21 diverse land use image types extracted from aerial orthoimagery, boasting a spatial resolution of 0.3 m. Originating from the United States Geological Survey (USGS) National Map, the original images spanning 20 distinct regions across the United States. Following the download, these images were cropped into 256 $\times$ 256 pixels. Each class is represented by a set of 100 images, culminating in a dataset totaling 2100 images.

Utilizing the aforementioned datasets, i.e., NWPU-RESISC45 (N) [60], AID (A) [61], and UC Merced Land-Use (U) [62], we formulate six cross-scene tasks by identifying shared categories between the datasets, which are termed N→A, A→N, N→U, U→N, A→U, and U→A, representing the knowledge transfer from the source domain to the target domain. Samples of the cross-scene classification tasks are shown in Figure 3. Table 1 details the image counts for both source and target domains, along with the number of common categories in each task.

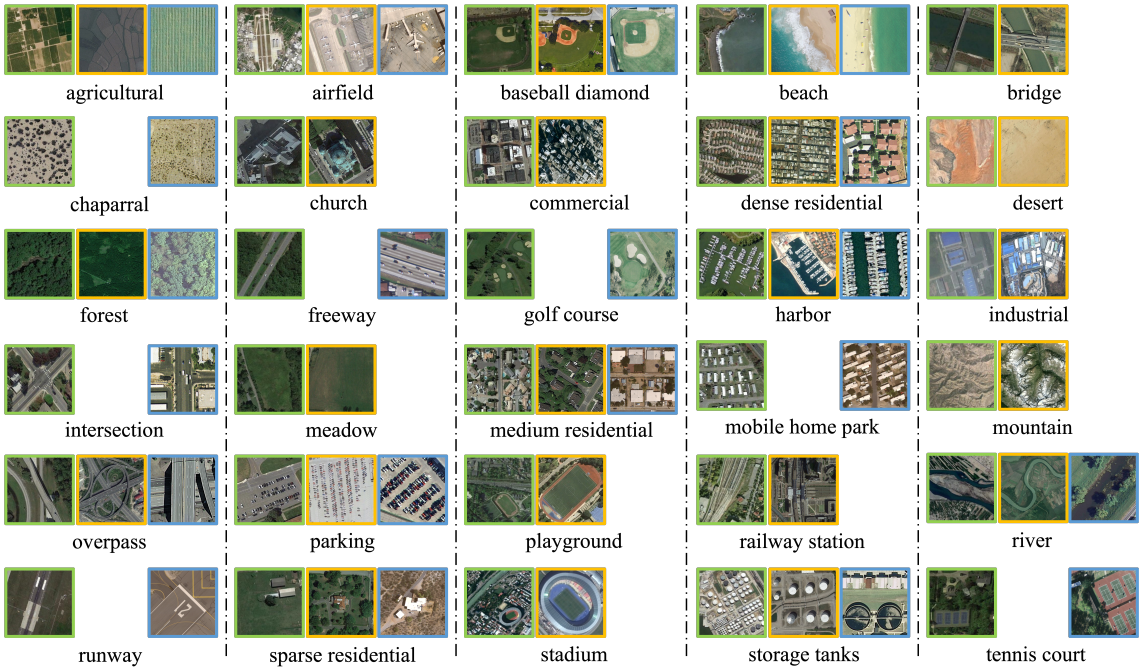


Figure 3. Sample images extracted from the NWPU (with green border), AID (with yellow border), and UCM (with blue border) datasets of our task.

Table 1. Labeled and unlabeled images used for each cross-scene classification task.

Task	Labeled Source	Unlabeled Target	Common Classes
N→A	16,100	7740	23
A→N	7740	16,100	23
N→U	14,000	2000	20
U→N	2000	14,000	20
A→U	4560	1300	13
U→A	1300	4560	13

4.2. Experimental Implementation Details

We implement our algorithm through the PyTorch framework, and all experiments are carried out on NVIDIA Tesla V100 GPUs. The encoder and decoder in our proposed framework load the parameters from the self-supervised pretraining CSPT in [54] as the domain-level knowledge model. We follow the default parameters of MAE [47] and conduct 400 epochs of self-supervised pretraining with target domain data for each cross-scene classification task. Then, in the DA training, we use the stochastic gradient descent (SGD) with momentum 0.9 as the optimizer, and the learning rate is adjusted as: $\eta_p = \frac{\eta_0}{(1+\alpha p)^\beta}$, where p is the training progress linearly changing from 0 to 1, $\alpha = 10$ and $\beta = 0.75$. Given that the classification layer is trained from scratch, we set its learning rate to 10 times that of the backbone. The initial learning rate η_0 of the backbone is set to 1×10^{-4} except for U→N, which is set to 2×10^{-4}. MIM-AP^2 utilizes a masking ratio of 10%, and the hyperparameter $\lambda_1 = 1.5$, $\lambda_2 = 10$.

4.3. Comparison Analysis

We evaluate the proposed MIM-AP^2 and compared it with other SOTA DA methods. Table 2 summarizes the average accuracy according to six cross-scene classification tasks and computational complexity of each method, and MIM-AP^2 attains the highest average accuracy of 95.79%. It can be seen the proposed MIM-AP^2 achieved the highest accuracy among the four tasks (A→N, N→U, U→N, and U→A), with the rest (N→A and A→U) achieving the second-highest accuracy. For source only methods without any DA techniques, ViT-B* outperforms ResNet-50* by about 4% in average accuracy due to the superior transferability of spatial contextual modeling. Notably, since DAN [16], JAN [18], ADDA [21], and CDAN [23] tend to neglect the representation learning of spatial context clues, they are prone to suboptimal solutions. Although AMRAN [8] employs attention mechanism to mine spatial information, the attention mechanism can only suppress some irrelevant redundant noise caused by the limited receptive field and bring marginal improvement. Methods such as CDTrans [32] and DOT [34], built upon ViT, can bring greater improvement because of a powerful spatial contextual modeling ability. Besides, they both designed pseudo-label optimization strategies for assisting in learning a superior classifier in target domain. However, CDTrans [32] overlooks domain-specific information learning and lacks sufficient feature mining in the target domain. On the other hand, while DOT [34] adopts two class tokens to maintain domain-specific information, it lacks an explicit signal indicating the capture of effective specific information. Our method places greater emphasis on adequate domain-specific information extraction and employs a more effective pseudo-label refinement strategy, ultimately culminating in superior outcomes. As listed in Table 2, despite ViT-based methods (CDTrans [32], DOT [34], and the proposed MIM-AP^2) often have a higher computational complexity than CNN-based methods (DAN [16], JAN [18], ADDA [21], CDAN [23], and AMRAN [8]), this increase is acceptable compared to the performance improvement. Furthermore, the detailed results of six cross-scene classification tasks are shown in Figures 4–9.

Table 2. Average accuracy (%) of different method on six cross-scene classification tasks. * indicates the baseline approaches that solely employ the classification loss in source domain without any DA method. The results displaying the utmost and second highest accuracy are accentuated in bold and underlined, respectively.

Method	Public	N→A	A→N	N→U	U→N	A→U	U→A	Avg	FLOPs(G)
ResNet-50 * [63]	CVPR 2016	83.14	71.74	71.18	58.24	72.74	60.89	69.66	4.13
ViT-B * [54]	RS 2022	89.88	83.22	77.75	65.42	70.69	56.57	73.92	16.86
DAN [16]	ICML 2015	85.16	76.68	82.35	64.64	83.54	72.87	77.54	4.13
JAN [18]	ICML 2017	87.61	82.04	89.18	79.89	87.74	89.88	86.06	4.13
ADDA [21]	CVPR 2017	86.51	81.53	88.37	81.64	91.13	87.71	86.15	4.13
CDAN [23]	NIPS 2018	93.32	86.86	92.18	85.21	93.69	94.17	90.91	4.13
AMRAN [8]	TGRS 2021	92.43	86.06	94.17	78.14	93.36	94.09	89.71	8.32
CDTrans [32]	ICLR 2022	93.23	85.93	94.27	83.89	91.74	95.53	90.76	16.98
DOT [34]	ACMMM 2022	**98.13**	91.52	95.25	92.44	**95.10**	95.77	94.70	16.95
MIM-AP^2	ours	97.92	**94.08**	**96.80**	**94.02**	93.92	**97.97**	**95.79**	16.86

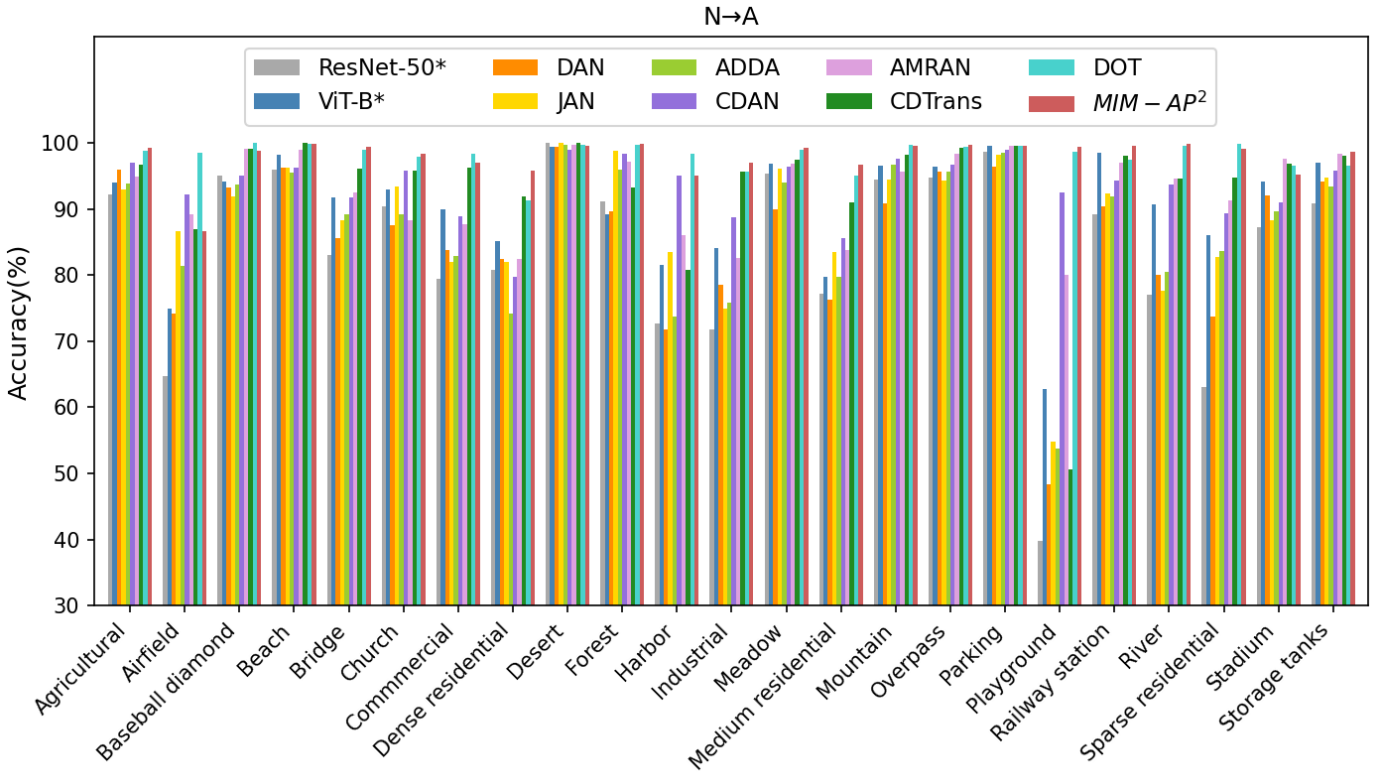


Figure 4. Accuracy (%) comparisons of different methods on the N→A task. * indicates the baseline approaches that solely employ the classification loss in source domain without any DA method.

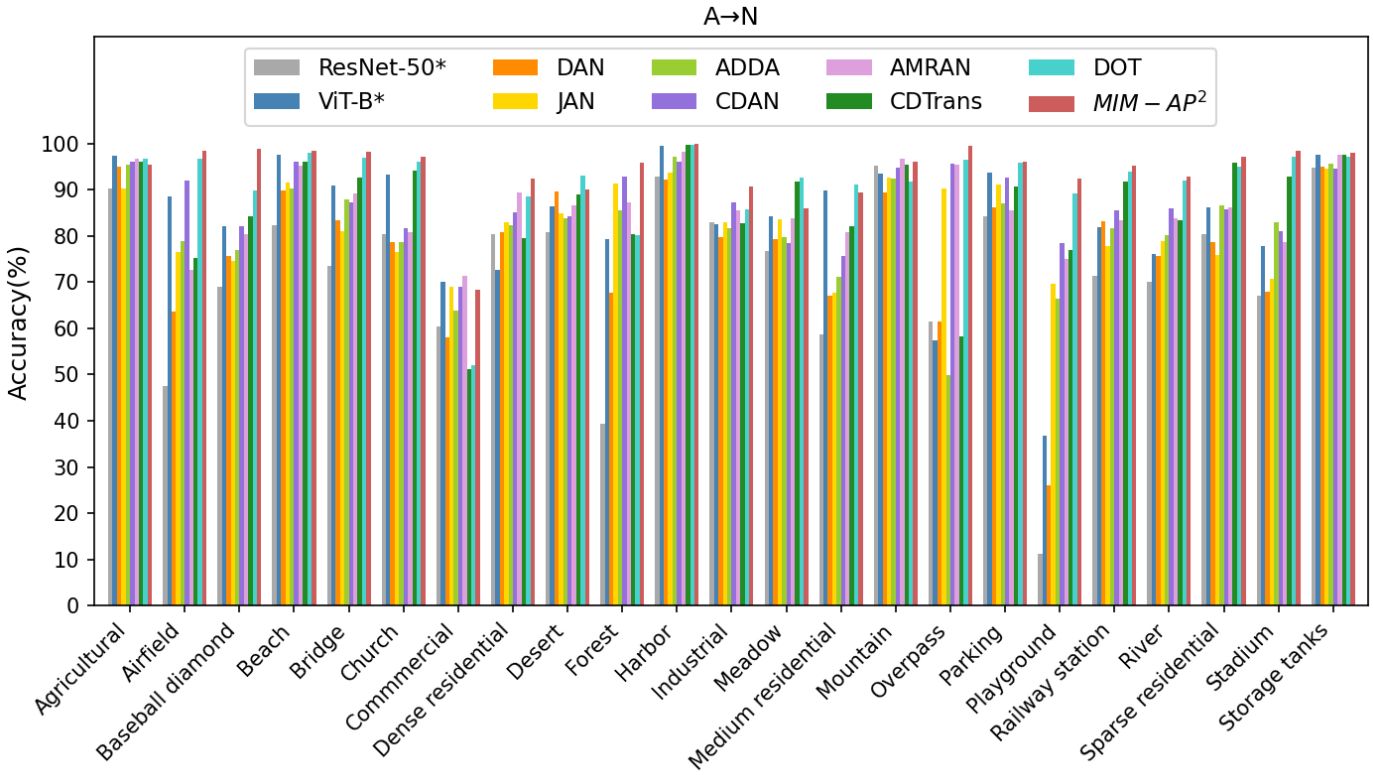


Figure 5. Accuracy (%) comparisons of different methods on the A→N task. * indicates the baseline approaches that solely employ the classification loss in source domain without any DA method.

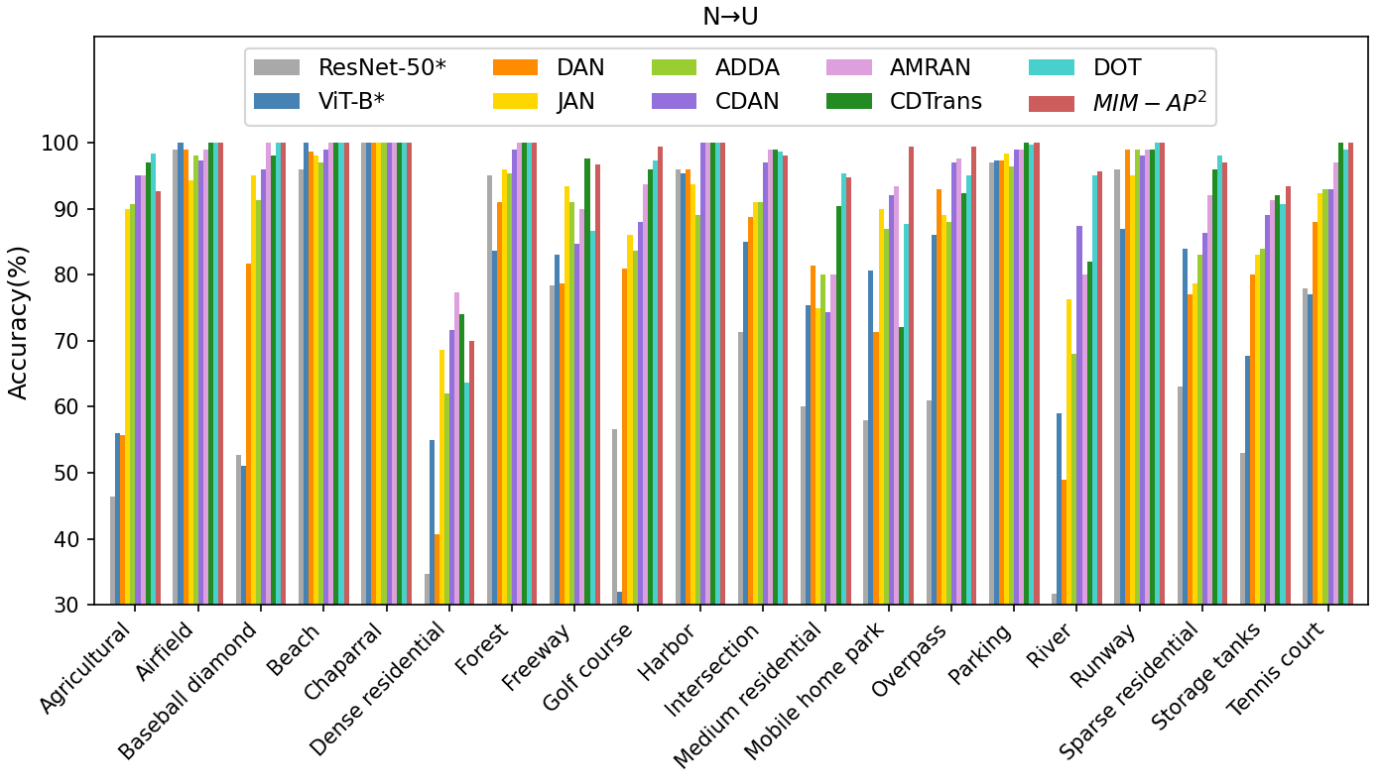


Figure 6. Accuracy (%) comparisons of different methods on the N→U task. * indicates the baseline approaches that solely employ the classification loss in source domain without any DA method.

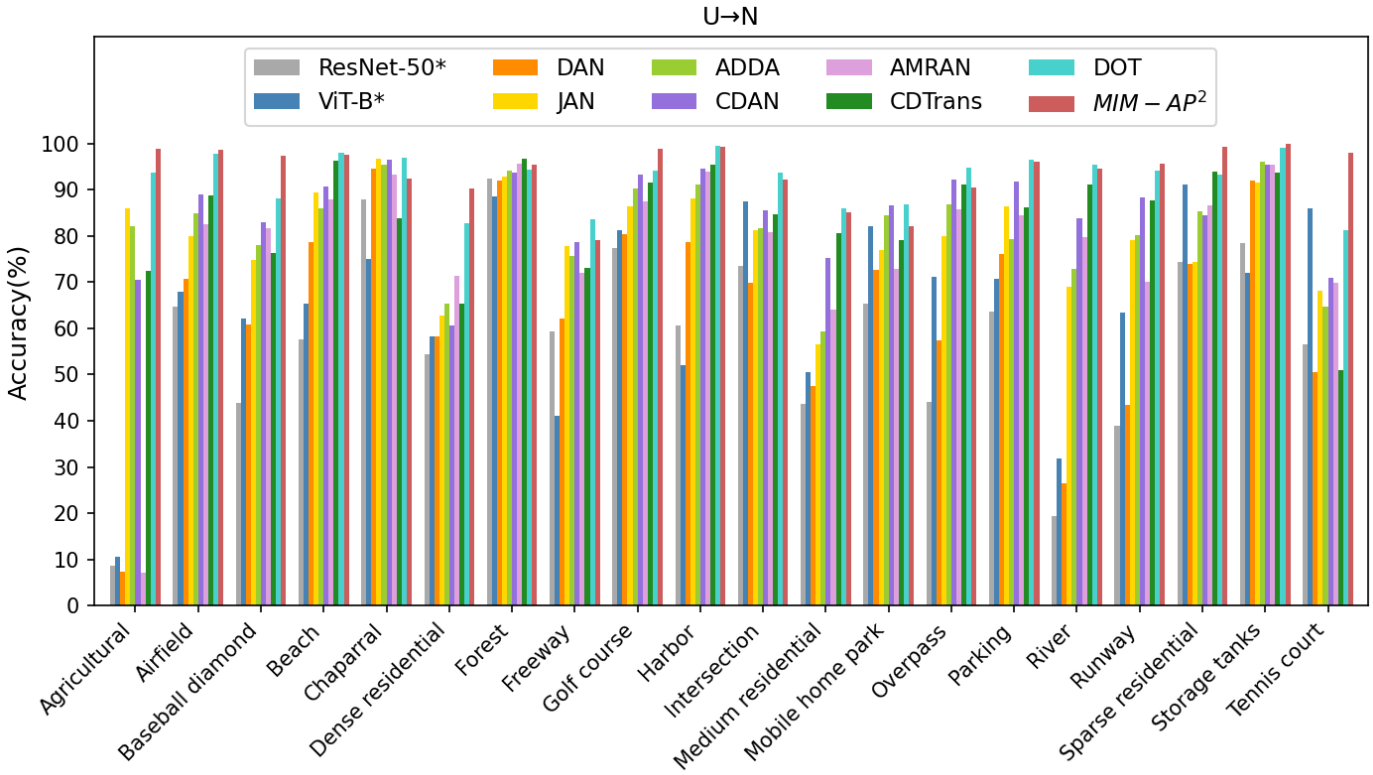


Figure 7. Accuracy (%) comparisons of different methods on the U→N task. * indicates the baseline approaches that solely employ the classification loss in source domain without any DA method.

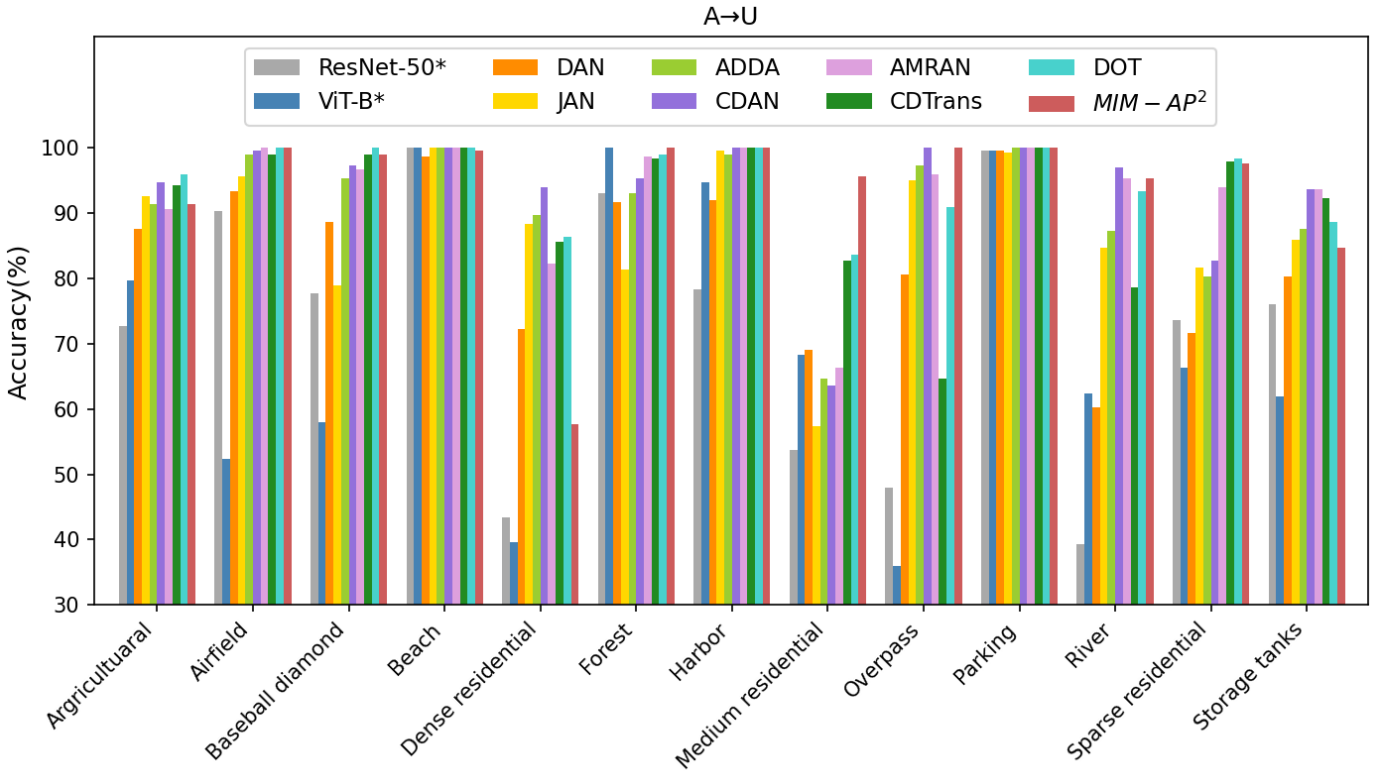


Figure 8. Accuracy (%) comparisons of different methods on the A→U task. * indicates the baseline approaches that solely employ the classification loss in source domain without any DA method.

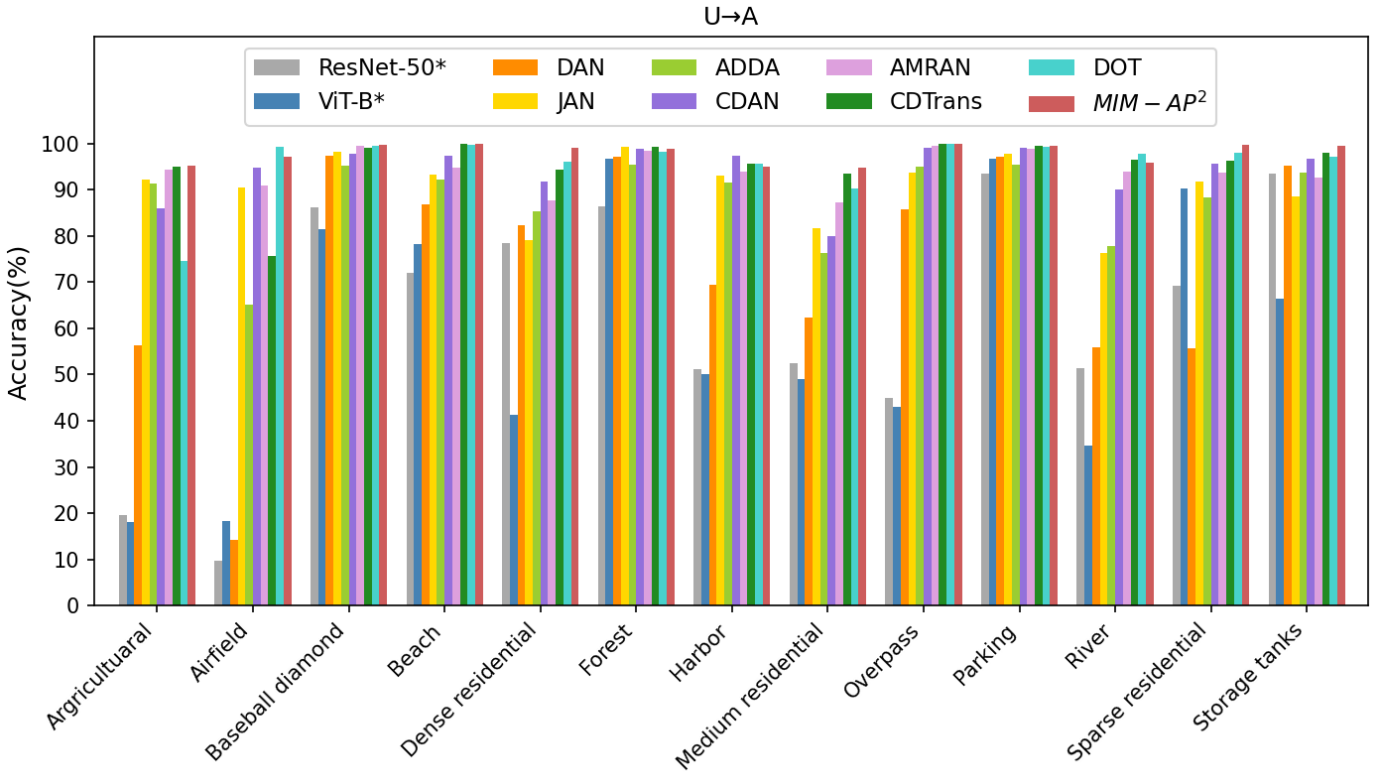


Figure 9. Accuracy (%) comparisons of different methods on the U→A task. * indicates the baseline approaches that solely employ the classification loss in source domain without any DA method.

4.4. Ablation Study

An array of ablation experiments were meticulously conducted, and the average accuracy of the six tasks is shown in Table 3. The first line represents the classification results after self-supervised pretraining on the target domain, which only use the classification loss in the source domain without any DA methods. In the second line, the contrastive learning-based alignment loss is added, where the category with the highest classifier prediction score is directly used as the pseudo-label for the target domain. The employment of supervised class-token contrastive learning aids in finding consistent contextual clues in the source and target domains, facilitating knowledge transfer and resulting in a 2.74% accuracy enhancement. The third line introduces our designed clustering central rectification strategy aimed at rectifying unreliable pseudo-labels, resulting in a substantial 9.85% improvement and an overall approach improvement of 12.59%. The fourth line further integrates self-supervised MIM mechanism in the target domain, forming our complete MIM-AP^2. This final integration achieves a remarkable 13.74% improvement. Notably, random masking and reconstruction can assist model capturing domain-specific information, further enhancing in-domain discrimination and yielding an additional accuracy boost of 1.15%.

Table 3. Ablation study. Pre denotes the self-supervised pretraining on the target domain. Contrastive Learning indicates that the L_{con} is added. Strategy denotes the utilization of the proposed clustering central rectification strategy. MIM denotes MIM for the target domain image. ↑ shows the improved average accuracy compared to the self-supervised pretraining model.

Pre	Contrastive Learning	Strategy	MIM	Avg (%)
✓				82.05
✓	✓			84.79 (↑ 2.74%)
✓	✓	✓		94.64 (↑ 12.59%)
✓	✓	✓	✓	95.79 (↑ 13.74%)

For a more intuitive presentation of the effects stemming from distinct components of the proposed MIM-AP^2, t-stochastic neighbor embedding (t-SNE) [64] is used to visualize the embedding feature distributions within the more challenge U→N task, as shown in Figures 10 and 11. The four images in Figure 10 or Figure 11 each correspond to four rows in Table 3, respectively. Figure 10 showcases the feature distributions between the source and target domains, where blue points and red points represent the source and target domain data, respectively. Observing Figure 10a, it is evident that serious domain discrepancies exist between the source and target domains. Progressing from Figure 10a–d, the disparities in distribution between source and target domains gradually diminish, with the feature distributions gradually aligning. In Figure 11, the focus shifts to the feature distribution within the target domain, with distinct colors denoting various categories. In Figure 11a, a notable confusion is observed among the target domain categories. Progressing from Figure 11a–d, the in-domain discrimination gradually improves, leading to clearer classification boundaries within the target domain.

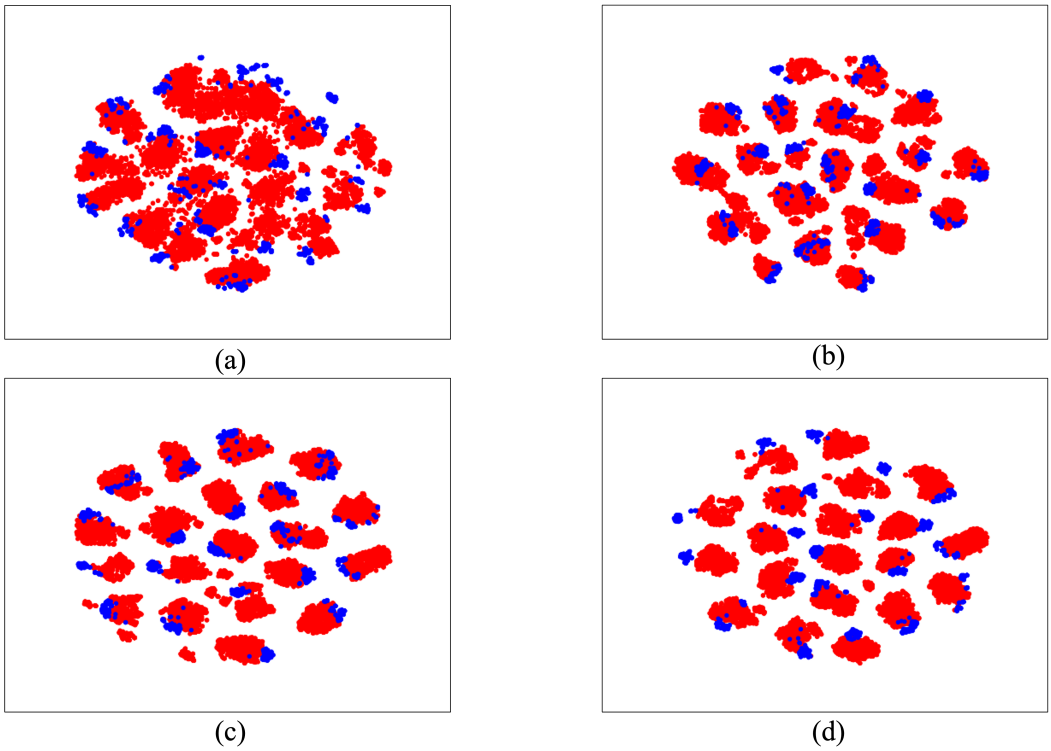


Figure 10. The t-SNE visualization of source and target domains features on U→N, where blue points are source domain data and red points are target domain data. (**a**) Features learned by the self-supervised pretraining model; (**b**) Features learned by contrastive learning with the pretrained model; (**c**) Features learned by contrast learning and clustering central rectification strategy with the pretrained model. (**d**) Features learned by MIM-AP^2.

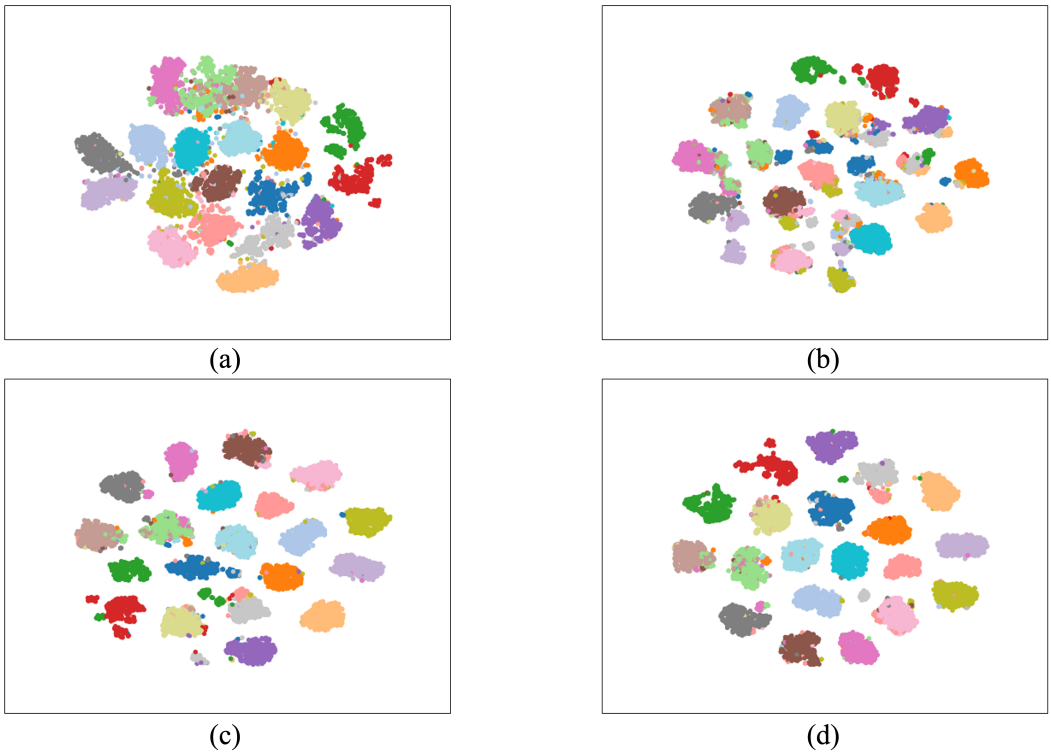


Figure 11. The t-SNE visualization of the target domain features on U→N, where different colored points represent different categories of the target domain. (**a**) Features learned by the self-supervised pretraining model; (**b**) Features learned by contrastive learning with the pretrained model; (**c**) Features learned by contrast learning and clustering central rectification strategy with the pretrained model. (**d**) Features learned by MIM-AP^2.

4.5. Parameter Discussion

The impact of the masking ratio is discussed in Figure 12. The yellow line illustrates the average accuracy of the proposed MIM-AP^2 on six cross-scene classification tasks, which are evaluated under four distinct masking ratios: 1%, 10%, 20%, 30%, and 75%. Notably, the results underscore that employing lower masking ratios yields superior classification performance, with the highest accuracy achieved at 10%. Where the green horizontal line represents MIM-AP^2 without MIM in target domain. As the masking ratio escalates, there is a gradual decline in accuracy, ultimately leading to negative optimization at the 75% masking ratio. To further verify the impact of random masking on the proposed MIM-AP^2, we excluded the decoder component and illustrated the results of only masking without reconstruction, depicted by the blue line. It is evident that the enhancement in accuracy at lower random masking ratios of 1%, 10%, and 20% substantiates that this data augmentation enriches potential feature representations. It facilitates the extraction of more consistent contextual information, thereby improving the model's transferability. Given the intricate and dense nature of information in RS images, a higher random masking ratio would result in the loss of significant information, affecting the learning of feature representation and leads to the inaccurate derivation of pseudo-labels. Additionally, under different random masking ratios, the trend of the yellow line is similar to that of the green line, but the accuracy is higher. The decoder's reconstruction capability enables the capture of specific information from the target domain and inject it into the class token, thereby establishing a sturdy foundation for generating high-quality pseudo-labels in the target domain.

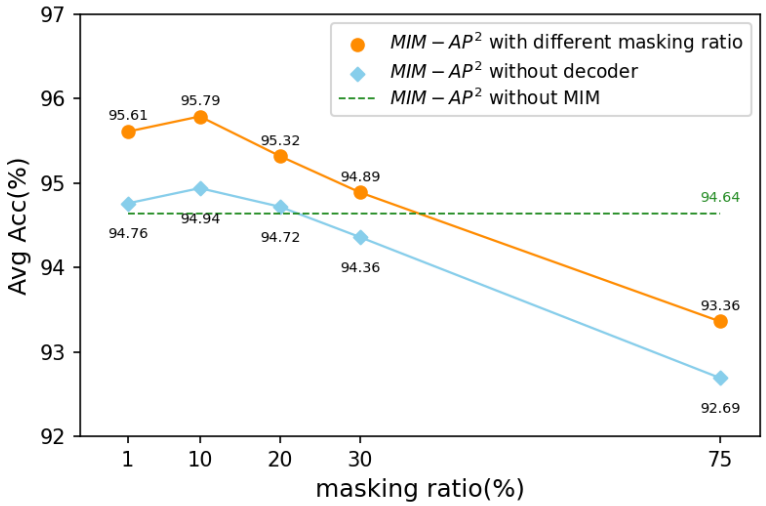


Figure 12. Masking ratio analysis.

5. Discussion

RS images contain abundant land cover information and exhibit complex spatial layouts, resulting in large intra-class variations and high inter-class similarities. The commonly used DA methods are based on discrepancy metric-based approaches and adversarial-based approaches. The discrepancy metric-based approaches typically seek domain-invariant features by narrowing the gap between the source and target domains, often overlooking domain-specific information within each domain. Particularly for unlabeled target domains, in the absence of supervised information, minimizing discrepancy metrics may only yield domain-invariant information similar to the source domain, and the domain-specific information that contributes to enhancing feature discriminability cannot be obtained. In adversarial training mechanisms, it is often challenging to assess and ensure the level of the opposing parties. Multi-classifier adversarial learning for cross-scene classification has garnered widespread attention, which is believed to enhance discriminability in the target domain. However, the discriminative capability of classifiers is also correlated with the quality of feature extraction. These methods often focus on adversarial DA on the relevant distribution of classifier outputs, neglecting how to better capture discriminative information during feature extraction. Leveraging pseudo-label propagation techniques in DA can unearth semantic information from unlabeled target domain samples, and achieves finer distribution alignment by using pseudo-label of the target domain. While ensuring the transferability and discriminability of extracted features, designing sensible pseudo-label optimization strategies holds significant potential in RS image cross-scene classification and should be promoted.

6. Conclusions

In this paper, a novel DA framework called MIM-AP^2 is proposed for remote sensing cross-scene classification. First, a new DA representation learning is introduced, which combines supervised class token contrastive learning with a self-supervised MIM mechanism. This way, not only achieves knowledge transfer by encouraging consistency in contextual clues between the source and target domains, but also leverages random masking pixel reconstruction in target domain to capture domain-specific information, thereby improving discriminability in-domain. Second, a novel clustering central rectification strategy is designed to reduce noise labels, so as to further bolster the precision of the pseudo-labels, facilitating the construction of a target domain specific classifier. In summary, the MIM-AP^2 framework, with its innovative DA representation way and clustering central rectification strategy, achieve the SOTA cross-scene classification performance in extensive experiments.

It can improve the ability of the model to process the actual application scenario data, which is very different from the training data. In applications such as land resource survey and natural resources monitoring, the proposed DA method can label newly acquired data and iterate the model continuously, reducing additional manual data labeling and lengthy model retraining process. In the future, we plan to further explore the MIM mechanism and better pseudo-label refinement methods to further improve the performance of our model.

Author Contributions: Conceptualization, X.Z. and Y.Z.; data curation, X.Z. and T.Z.; software and validation, X.Z.; formal analysis, X.Z. and Y.Z.; writing—original draft preparation, X.Z. and Y.Z.; writing—review and editing, X.Z., Y.Z., T.Z., C.L. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the General Program of National Natural Science Foundation of China under grant number 62371048, in part by the National Science Foundation for Young Scientists of China under Grant number 62101046, and in part by the multisource satellite data hardware acceleration computing method with low energy consumption under Grant number 2021YFA0715204.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Al-Kindi, K.M.; Alqurashi, A.F.; Al-Ghafri, A.; Power, D. Assessing the Impact of Land Use and Land Cover Changes on Aflaj Systems over a 36-Year Period. *Remote Sens.* **2023**, *15*, 1787. [CrossRef]
2. Fernandez, L.; Ruiz-de Azua, J.A.; Calveras, A.; Camps, A. On-Demand Satellite Payload Execution Strategy for Natural Disasters Monitoring Using LoRa: Observation Requirements and Optimum Medium Access Layer Mechanisms. *Remote Sens.* **2021**, *13*, 4014. [CrossRef]
3. Bai, H.; Li, Z.; Guo, H.; Chen, H.; Luo, P. Urban Green Space Planning Based on Remote Sensing and Geographic Information Systems. *Remote Sens.* **2022**, *14*, 4213. [CrossRef]
4. Liu, Q.; He, M.; Kuang, Y.; Wu, L.; Yue, J.; Fang, L. A Multi-Level Label-Aware Semi-Supervised Framework for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5616112. [CrossRef]
5. Tian, Y.; Dong, Y.; Yin, G. Early Labeled and Small Loss Selection Semi-Supervised Learning Method for Remote Sensing Image Scene Classification. *Remote Sens.* **2021**, *13*, 4039. [CrossRef]
6. Miao, W.; Geng, J.; Jiang, W. Semi-Supervised Remote-Sensing Image Scene Classification Using Representation Consistency Siamese Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5616614. [CrossRef]
7. Li, J.; Liao, Y.; Zhang, J.; Zeng, D.; Qian, X. Semi-Supervised DEGAN for Optical High-Resolution Remote Sensing Image Scene Classification. *Remote Sens.* **2022**, *14*, 4418. [CrossRef]
8. Zhu, S.; Du, B.; Zhang, L.; Li, X. Attention-Based Multiscale Residual Adaptation Network for Cross-Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5400715. [CrossRef]
9. Yang, C.; Dong, Y.; Du, B.; Zhang, L. Attention-Based Dynamic Alignment and Dynamic Distribution Adaptation for Remote Sensing Cross-Domain Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5634713. [CrossRef]
10. Niu, B.; Pan, Z.; Wu, J.; Hu, Y.; Lei, B. Multi-Representation Dynamic Adaptation Network for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5633119. [CrossRef]
11. Zhang, X.; Yao, X.; Feng, X.; Cheng, G.; Han, J. DFENet for Domain Adaptation-Based Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5611611. [CrossRef]
12. Huang, W.; Shi, Y.; Xiong, Z.; Wang, Q.; Zhu, X.X. Semi-supervised bidirectional alignment for Remote Sensing cross-domain scene classification. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 192–203. [CrossRef]
13. Liang, C.; Cheng, B.; Xiao, B.; Dong, Y. Unsupervised Domain Adaptation for Remote Sensing Image Segmentation Based on Adversarial Learning and Self-Training. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6006005. [CrossRef]
14. Zhu, J.; Guo, Y.; Sun, G.; Yang, L.; Deng, M.; Chen, J. Unsupervised Domain Adaptation Semantic Segmentation of High-Resolution Remote Sensing Imagery With Invariant Domain-Level Prototype Memory. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5603518. [CrossRef]
15. Sun, Y.; Wang, Y.; Liu, H.; Hu, L.; Zhang, C.; Wang, S. Gradual Domain Adaptation with Pseudo-Label Denoising for SAR Target Recognition When Using Only Synthetic Data for Training. *Remote Sens.* **2023**, *15*, 708. [CrossRef]
16. Long, M.; Cao, Y.; Wang, J.; Jordan, M. Learning Transferable Features with Deep Adaptation Networks. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Bach, F., Blei, D., Eds.; Volume 37, pp. 97–105.
17. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A Kernel Two-Sample Test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.

18. Long, M.; Zhu, H.; Wang, J.; Jordan, M.I. Deep Transfer Learning with Joint Adaptation Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Precup, D., Teh, Y.W., Eds.; PMLR: Birmingham, UK; Volume 70, pp. 2208–2217.
19. Li, J.; Chen, E.; Ding, Z.; Zhu, L.; Lu, K.; Shen, H.T. Maximum Density Divergence for Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3918–3930. [CrossRef]
20. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
21. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial Discriminative Domain Adaptation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2962–2971. [CrossRef]
22. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; MIT Press: Cambridge, MA, USA, 2014; Volume 27.
23. Long, M.; Cao, Z.; Wang, J.; Jordan, M.I. Conditional Adversarial Domain Adaptation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; MIT Press: Cambridge, MA, USA, 2018; Volume 31.
24. Saito, K.; Watanabe, K.; Ushiku, Y.; Harada, T. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
25. Ammour, N.; Bashmal, L.; Bazi, Y.; Al Rahhal, M.M.; Zuair, M. Asymmetric Adaptation of Deep Features for Cross-Domain Classification in Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 597–601. [CrossRef]
26. Song, S.; Yu, H.; Miao, Z.; Zhang, Q.; Lin, Y.; Wang, S. Domain Adaptation for Convolutional Neural Networks-Based Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1324–1328. [CrossRef]
27. Zhu, S.; Luo, F.; Du, B.; Zhang, L. Adversarial Fine-Grained Adaptation Network for Cross-Scene Classification. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2369–2372. [CrossRef]
28. Teng, W.; Wang, N.; Shi, H.; Liu, Y.; Wang, J. Classifier-Constrained Deep Adversarial Domain Adaptation for Cross-Domain Semisupervised Classification in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 789–793. [CrossRef]
29. Ma, C.; Sha, D.; Mu, X. Unsupervised Adversarial Domain Adaptation with Error-Correcting Boundaries and Feature Adaption Metric for Remote-Sensing Scene Classification. *Remote Sens.* **2021**, *13*, 1270. [CrossRef]
30. Zheng, Z.; Zhong, Y.; Su, Y.; Ma, A. Domain Adaptation via a Task-Specific Classifier Framework for Remote Sensing Cross-Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620513. [CrossRef]
31. Yang, J.; Liu, J.; Xu, N.; Huang, J. TVT: Transferable Vision Transformer for Unsupervised Domain Adaptation. *arXiv* **2021**, arXiv:2108.05988.
32. Xu, T.; Chen, W.; WANG, P.; Wang, F.; Li, H.; Jin, R. CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
33. Wang, X.; Guo, P.; Zhang, Y. Domain Adaptation via Bidirectional Cross-Attention Transformer. *arXiv* **2022**, arXiv:2201.05887.
34. Ma, W.; Zhang, J.; Li, S.; Liu, C.H.; Wang, Y.; Li, W. Making The Best of Both Worlds: A Domain-Oriented Transformer for Unsupervised Domain Adaptation. In Proceedings of the 30th ACM International Conference on Multimedia (MM'22), Lisboa, Portugal, 10–14 October 2022; pp. 5620–5629. [CrossRef]
35. Lee, D.H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Proceedings of the Workshop on Challenges in Representation Learning, ICML, Atlanta, GA, USA, 16–21 June 2013; Volume 3, p. 896.
36. Liang, J.; Hu, D.; Feng, J. Domain Adaptation with Auxiliary Target Domain-Oriented Classifier. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16632–16642.
37. Liang, J.; Hu, D.; Feng, J. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 6028–6039.
38. Gu, X.; Sun, J.; Xu, Z. Unsupervised and Semi-supervised Robust Spherical Space Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *46*, 1757–1774. [CrossRef] [PubMed]
39. Zhang, W.; Xu, D.; Ouyang, W.; Li, W. Self-Paced Collaborative and Adversarial Network for Unsupervised Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2047–2061. [CrossRef]
40. Lu, Y.; Wong, W.K.; Zeng, B.; Lai, Z.; Li, X. Guided Discrimination and Correlation Subspace Learning for Domain Adaptation. *IEEE Trans. Image Process.* **2023**, *32*, 2017–2032. [CrossRef]
41. Yu, Y.C.; Lin, H.T. Semi-Supervised Domain Adaptation with Source Label Adaptation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 24100–24109. [CrossRef]
42. Kwak, G.H.; Park, N.W. Unsupervised domain adaptation with adversarial self-training for crop classification using remote sensing images. *Remote Sens.* **2022**, *14*, 4639. [CrossRef]
43. Gao, K.; Yu, A.; You, X.; Qiu, C.; Liu, B. Prototype and Context Enhanced Learning for Unsupervised Domain Adaptation Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5608316. [CrossRef]

44. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the naacL-HLT, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, p. 2.
45. Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT Pre-Training of Image Transformers. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
46. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. SimMIM: A Simple Framework for Masked Image Modeling. *arXiv* **2021**, arXiv:2111.09886.
47. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988. [CrossRef]
48. Bachmann, R.; Mizrahi, D.; Atanov, A.; Zamir, A. MultiMAE: Multi-modal Multi-task Masked Autoencoders. *arXiv* **2022**, arXiv:2204.01678.
49. Gao, P.; Lin, Z.; Zhang, R.; Fang, R.; Li, H.; Li, H.; Qiao, Y. Mimic before reconstruct: Enhancing masked autoencoders with feature mimicking. *Int. J. Comput. Vis.* **2023**, *132*, 1546–1556. [CrossRef]
50. Liu, J.; Huang, X.; Yoshie, O.; Liu, Y.; Li, H. MixMIM: Mixed and Masked Image Modeling for Efficient Visual Representation Learning. *arXiv* **2022**, arXiv:2205.13137.
51. Wei, C.; Fan, H.; Xie, S.; Wu, C.Y.; Yuille, A.; Feichtenhofer, C. Masked feature prediction for self-supervised visual pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14668–14678.
52. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A Remote Sensing Foundation Model with Masked Image Modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5612822. [CrossRef]
53. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model. *arXiv* **2022**, arXiv:2208.03987.
54. Zhang, T.; Gao, P.; Dong, H.; Zhuang, Y.; Wang, G.; Zhang, W.; Chen, H. Consecutive Pre-Training: A Knowledge Transfer Learning Strategy with Relevant Unlabeled Data for Remote Sensing Domain. *Remote Sens.* **2022**, *14*, 5675. [CrossRef]
55. Wang, R.; Wu, Z.; Weng, Z.; Chen, J.; Qi, G.J.; Jiang, Y.G. Cross-Domain Contrastive Learning for Unsupervised Domain Adaptation. *IEEE Trans. Multimed.* **2023**, *25*, 1665–1673. [CrossRef]
56. Liu, W.; Wang, X.; Owens, J.; Li, Y. Energy-based out-of-distribution detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21464–21475.
57. Wang, H.; Liu, W.; Bocchieri, A.; Li, Y. Energy-based Out-of-distribution Detection for Multi-label Classification. In Proceedings of the International Conference on Learning Representations, ICLR 2021, Vienna, Austria, 4 May 2021.
58. Lin, Z.; Roy, S.D.; Li, Y. MOOD: Multi-Level Out-of-Distribution Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15313–15323.
59. Choi, H.; Jeong, H.; Choi, J.Y. Balanced Energy Regularization Loss for Out-of-distribution Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 15691–15700.
60. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
61. 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]
62. Yang, Y.; Newsam, S. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS'10), San Jose, CA, USA, 2–5 November 2010; pp. 270–279. [CrossRef]
63. 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), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
64. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

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