Robust Representation Learning Based on Deep Mutual Information for Scene Classification Against Adversarial Perturbations

Linjuan Li, Gang Xie, *Member, IEEE*, Haoxue Zhang, Xinlin Xie, and Heng Li

***Abstract*—Remote sensing scene classification enables data-driven decisions for various applications, such as environmental monitoring, urban planning, and disaster management. However, deep learning models used for scene classification are highly vulnerable to adversarial samples, resulting in incorrect predictions and posing significant risks. While most current methods focus on improving adversarial robustness, they face a trade-off that compromises accuracy on clean, unperturbed images. To address this challenge, we utilized information theory by incorporating a mutual information (MI) representation module, which allows the model to capture high-quality, robust features. Furthermore, a domain adversarial training strategy is applied to promote the learning of domain-invariant features, reducing the effect of distribution differences between clean images and adversarial samples. We propose a novel algorithm that accurately differentiates between clean and adversarial scenes by introducing the MI and domain adaptation-guided network. Extensive experiments demonstrate the effectiveness of our approach against adversarial attacks, revealing a positive correlation between adversarial perturbations and image information entropy, and a negative correlation with robust accuracy.**

***Index Terms*—Adversarial examples, deep mutual information (MI), deep neural networks (DNNs), remote sensing (RS) images, scene classification, unsupervised domain adaptation (UDA).**

I. Introduction

With the rapid expansion of remote sensing (RS) data and the advancement of computational capabilities, RS scene classification has become increasingly important for various applications, including land and resource surveys [1], [2], disaster assessment [4], [5], and urban development planning [3]. Nevertheless, in open environments, the presence of unseen noise [6], unknown anomalies [7], and imperceptible perturbations [8] poses significant challenges to the robustness of scene classification models, which is a primary concern for both academia and industry.

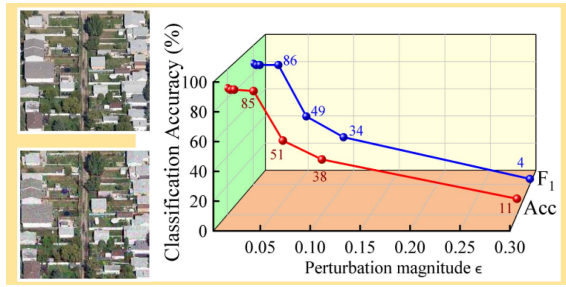

Fig. 1. Impact of adversarial samples on DNN classifier performance. The left shows a pair of clean images alongside their corresponding adversarial example, while the right depicts the change in Acc of the DNN classifier in response to perturbations. As the magnitude of adversarial perturbation increases, the model's performance significantly degrades.

These issues are further intensified by the limitations inherent in deep neural network (DNN)-based algorithms. Although DNNs have achieved notable success in scene classification [9], [10], [11], they remain highly susceptible to adversarial samples–inputs subtly altered to mislead the model, thus undermining both accuracy (Acc) and robustness [12], [13], [14]. As shown in Fig. 1, despite clean and adversarial samples appearing nearly identical, increasing the perturbation magnitude ϵ causes a steep drop in performance on the RSSCN7 dataset. For $\epsilon = 0.03$, Acc and F1 scores reach 85% and 86%, respectively, but with $\epsilon = 0.06$, they decrease sharply to 51% and 49%, indicating a nearly 40% reduction. Higher perturbation magnitudes render the model ineffective in recognizing common landmarks and structures.

The potential risks posed by adversarial perturbations are particularly concerning in critical areas, such as account security [15], [16], privacy protection [17], [18], and military operations [19], [20]. For example, adversarial images placed on building rooftops could obscure crucial infrastructure from surveillance systems or, in a military setting, mislead drones into misclassifying civilian structures as military targets, leading to disastrous consequences. This underscores the urgent need for robust methods to defend against adversarial threats.

Various strategies have been developed to enhance the robustness of deep scene classification models. Initially, data

Received 26 January 2025; revised 7 March 2025 and 30 March 2025; accepted 17 April 2025. Date of publication 28 April 2025; date of current version 16 May 2025. This work was supported in part by the Fundamental Research Program of Shanxi Province under Grant 202303021212222 and Grant 202303021221141, in part by the Industry-University-Research Innovation Fund for Chinese Universities under Grant 2021ZYA11005, and in part by the Key Research and Development Plan of Shanxi Province under Grant 202202010101005. *(Corresponding author: Gang Xie.)*

Linjuan Li, Haoxue Zhang, and Xinlin Xie are with the School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China (e-mail: linjuanli@tyust.edu.cn; B202115310014@stu.tyust.edu.cn).

Gang Xie is with the Shanxi Key Laboratory of Advanced Control and Equipment Intelligence, Taiyuan 030024, China (e-mail: xiegang@tyust.edu.cn).

Heng Li is with the Department of Building and Real Estate, The Hong Kong Polytechnic University,, Hong Kong (e-mail: bshengli@polyu.edu.hk).

Digital Object Identifier 10.1109/JSTARS.2025.3564376

augmentation techniques introduced diversity to training data, improving model adaptability to varying inputs [21], [22], [23]. Model architecture enhancements, such as multiscale feature fusion [24], [25] and attention mechanisms [26], [27], [28], have further bolstered feature selection robustness. Another crucial aspect for robust scene representations is the use of multiple instance learning (MIL), which mitigates the impact of permutation and ensures more stable feature learning. Key works in this area, such as [29], [30], and [31] have employed MIL in RS, demonstrating its ability to enhance scene classification robustness by better handling local features and permutation effects. Moreover, robust loss functions and regularization approaches help mitigate outliers and reduce overfitting [32]. However, These methods may fall short against well-crafted adversarial samples.

In recent years, adversarial defense techniques, including adversarial training (AdvT), purification, and detection, have garnered significant attention. Among these, AdvT has emerged as one of the most widely adopted methods for enhancing model robustness against adversarial attacks. This approach integrates adversarial samples, generated during backpropagation using gradient information, into the training process. Popular methods for generating adversarial samples include the fast gradient sign method (FGSM) [33], the basic iterative method [34], projected gradient descent (PGD) [35], and the DeepFool attack [36]. Despite its popularity, AdvT has notable limitations: it relies on a fully supervised framework requiring labeled adversarial samples, which are costly and labor-intensive to obtain, incurs high computational overhead from iterative sample generation, and degrades performance on clean samples, particularly with diverse data distributions.

Adversarial purification encompasses methods aimed at reducing or eliminating perturbations or noises in images before classification. For instance, Nie et al. [37] employed diffusion models to achieve this using a reverse diffusion process to restore clean samples from adversarial perturbated images. Similarly, Wu et al. [38] utilized diffusion models to incrementally diminish noise, enabling a return of adversarial samples to a condition closer to their original state. While these approaches can effectively handle simpler perturbations, their iterative nature results in substantial computational overhead and diminished effectiveness when addressing intricate adversarial perturbations.

Adversarial detection [39], [40], [41] focuses on identifying adversarial samples before their processing by DNNs. This approach aims to maintain standard model Acc while enhancing robustness against adversarial attacks. Typically, this involves the use of specialized detectors, such as auxiliary classifiers [42], [43] or tailored statistical measures [44], [45]. However, these detectors often struggle to generalize across various tasks and scenarios. The solutions above frequently overlook the differences in data distribution between adversarial and clean images, particularly in the context of RS. RS images often feature more complex backgrounds, multiscale land objects, and greater detail than natural images, which can lead to poor generalization of models when confronted with unknown perturbations.

Building on this analysis, our study seeks to address the challenge of adversarial samples that lead to high-confidence misclassifications in scene classification models by reevaluating data distribution, ensuring consistent performance and resilience even in adversarial contexts.

Unlike the aforementioned methods, which rely on traditional supervised approaches, we propose an unsupervised domain adaptation (UDA) framework to enhance model robustness, particularly for cross-domain tasks without semantic annotation and unknown adversarial distributions. To this end, we introduce the MI and domain adaptation-guided network (MIDANet) for RS scene classification, which incorporates the following key components: 1) mutual information representation (MIR) module, drawing inspiration from [46], the MIR module learns robust representations at both local and global scales using information theory. This dual-scale approach ensures resilience to adversarial perturbations by capturing discriminative and domain-invariant features. 2) domain AdvT strategy, to address significant distribution shifts between clean (source domain) and adversarial (target domain) samples, we integrate a domain AdvT strategy, which aligns feature distributions, enabling the learning of domain-invariant features. Our approach overcomes the challenges associated with complex iterative processes and the reliance on additional structures, achieving a balance between computational efficiency and versatility. This ensures robust performance across diverse adversarial distributions, making it highly suitable for a wide range of RS tasks.

Furthermore, to thoroughly assess robustness, we generated a diverse array of adversarial samples with varying types and levels of perturbation using FGSM and PGD techniques. Our extensive experiments reveal that adversarial perturbations considerably influence uncertainty and complexity. Our findings indicate a positive correlation between adversarial perturbations and image information entropy, alongside a negative correlation with robust Acc. The key contributions of this study include the following.

1) *MIDANet framework:* We introduce a robust framework for RS scene classification, designed to withstand adversarial perturbations by employing MIR and domain adaptation techniques.
2) *MIR module:* We have integrated the MIR module to facilitate robust representation learning, maximizing both local and global MI between input images and deep-learned features.
3) *Domain AdvT strategy:* We implement a domain AdvT strategy to address the discrepancies between clean and adversarial samples, reducing representation gaps and promoting domain-invariant feature learning to enhance generalization across various adversarial distributions.
4) *Adversarial correlation:* A series of experiments confirm that adversarial perturbations elevate the information entropy of images, resulting in an increase in the model's predictions. This finding suggests a positive correlation between adversarial perturbations and information entropy, as well as a negative correlation with robust Acc.

The rest of this article is organized as follows. Section II introduces the foundational concepts relevant to the description and generation of adversarial samples. In Section III, the proposed network architecture and its key components are thoroughly

explained. Section IV outlines a series of experiments and provides an in-depth analysis of the results. Finally, Section VI concludes this article and highlights the key contributions of the study.

II. Preliminaries

This section outlines the concept of adversarial examples, followed by a brief overview of the two primary methods used for generating adversarial examples: the FGSM, and PGD method. Besides, MIR learning is described in detail, highlighting the differences between our designed MIR module and previous works.

A. Adversarial Examples Description

Consider a well-trained DNN model for RS scene classification, denoted as $f(\cdot) : \mathbf{x} \to y$. Let $\mathbf{x}$ represent the original image, and $\mathbf{x}^{\text{adv}}$ represent its corresponding adversarial image. The true label of $\mathbf{x}$ is y, while y^* is the prediction generated by the DNN model. Adversarial samples can be formulated as a minimum optimization problem, as follows:

$$\begin{aligned} &\min_{\eta} \left\|\mathbf{x}^{\text{adv}} - \mathbf{x}\right\|_p \\ &\text{s.t. } \left\|\mathbf{x}^{\text{adv}} - \mathbf{x}\right\|_p \leqslant \eta \\ &f(\mathbf{x}) = y, f\left(\mathbf{x}^{\text{adv}}\right) = y^*, y \neq y^* \end{aligned} \tag{1}$$

where $\|\cdot\|_p$ represents the p-norm, which measures the distance between the original image $\mathbf{x}$ and its adversarial counterpart $\mathbf{x}^{\text{adv}}$. In this formulation, $\mathbf{x}$ is a high-dimensional matrix, specifically an image with dimensions $h \times w \times c$, where h and w are the height and width, and c is the number of channels (e.g., $c = 3$ for a color image). The p-norm is computed element-wise across the matrix, as follows: $\|\mathbf{x}^{adv} - \mathbf{x}\|_p = (\sum_{i,j,c} |\mathbf{x}^{\text{adv}}_{i,j,c} - \mathbf{x}_{i,j,c}|^p)^{1/p}$, where i and j are the spatial coordinates and c is the channel index. η denotes the adversarial perturbation. Equation (1) illustrates the existence of a minimum perturbation, where adding η to the RS input $\mathbf{x}$ causes the DNN model to produce an incorrect prediction y^*, which differs from the true label y of $\mathbf{x}$.

B. Adversarial RS Image Generation

In generating adversarial examples for RS images, we employ two widely-used methods: FGSM and PGD. These methods generate key adversarial perturbations, $\mathbf{x}^{\text{adv}}$, that challenge the robustness of models in the RS domain.

1) FGSM: This technique works by identifying a perturbation in the direction of the gradient ascend to increase the model's loss function. The algorithm calculates the gradient of the loss function with respect to the input and applies a small ℓ_∞ perturbation to generate adversarial data

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \mathbf{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{x}, y\right)\right) \tag{2}$$

where $\mathcal{L}(\mathbf{x}, y)$ is the model's loss function, $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, y)$ represents the first-order derivative of the loss with respect to input $\mathbf{x}$, and $\mathbf{sign}(\cdot)$ refers to the sign function. The adversarial perturbation is given by $\eta = \epsilon \mathbf{sign}(\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, y))$. FGSM aims to lead the model to produce incorrect predictions via a one-step gradient update.

Due to its simplicity and efficiency, FGSM is a fast and effective method for evaluating the model's vulnerability under minimal perturbations. By using FGSM-generated adversarial samples, we can quickly identify potential weaknesses in the model and explore strategies for improving its robustness.

2) PGD method: PGD is a more robust adversarial approach, extending FGSM by performing multiple iterative steps with smaller perturbations at each step. While FGSM uses a single-step approach, PGD performs multiple iterative steps, taking smaller steps at each iteration. In each step, the perturbation is clipped to stay within a specified range. The process is defined as

$$\mathbf{x}^{t+1} = \mathbf{Clip}_{\mathbf{x},\epsilon}\left(\mathbf{x}^t + \alpha \mathbf{sign}\left(\nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{x}, y\right)\right)\right) \tag{3}$$

where t denotes the iteration number, α is the step size, and $\mathbf{Clip}_{\mathbf{x},\epsilon}$ ensures that $\mathbf{x}$ remains within ϵ. Although PGD is more time-consuming than FGSM, it is recognized as one of the most effective gradient-based adversarial attacks.

In our work, PGD is crucial for evaluating the model's robustness against more complex and iterative adversarial attacks. Its iterative nature helps uncover deeper vulnerabilities and assess resilience under progressively stronger perturbations. PGD also aids in evaluating defense strategies by revealing weaknesses not exposed by simpler attacks, such as FGSM, making it an essential tool for enhancing model robustness against advanced adversarial threats.

In our experiments, n our experiments, we utilized both FGSM and PGD to generate adversarial examples with varying perturbation magnitudes, specifically $\epsilon = 0.01, 0.03, 0.06, 0.1, 0.3$. These perturbation levels were systematically adjusted to evaluate the model's performance under different adversarial conditions. The adversarial samples with varying perturbation magnitudes were then used to assess the robustness of our model in RS tasks.

C. MIR Learning

Mutual information (MI) is crucial for understanding the relationship between input data and learned representations, as it quantifies the shared information between them. Recently, several methods have been developed to use MI to improve a model's ability to capture meaningful features and enhance its representational power.

One of the foundational works in this area is mutual information neural estimation (MINE), proposed by Belghazi et al. [47]. MINE estimates MI using a neural network and the Donsker–Varadhan (DV) bound to optimize a lower bound on Kullback–Leibler (KL) divergence. However, MINE relies on a single discriminator for MI estimation, and the asymmetry of KL divergence can cause training instability, especially in high-dimensional or complex data. Advancements in MI-based representation learning were proposed by Hjelm et al. [46], who introduced MI maximization and a more stable, symmetric MI estimation method. While this reduces instability, challenges remain, including inefficiencies in handling high-dimensional

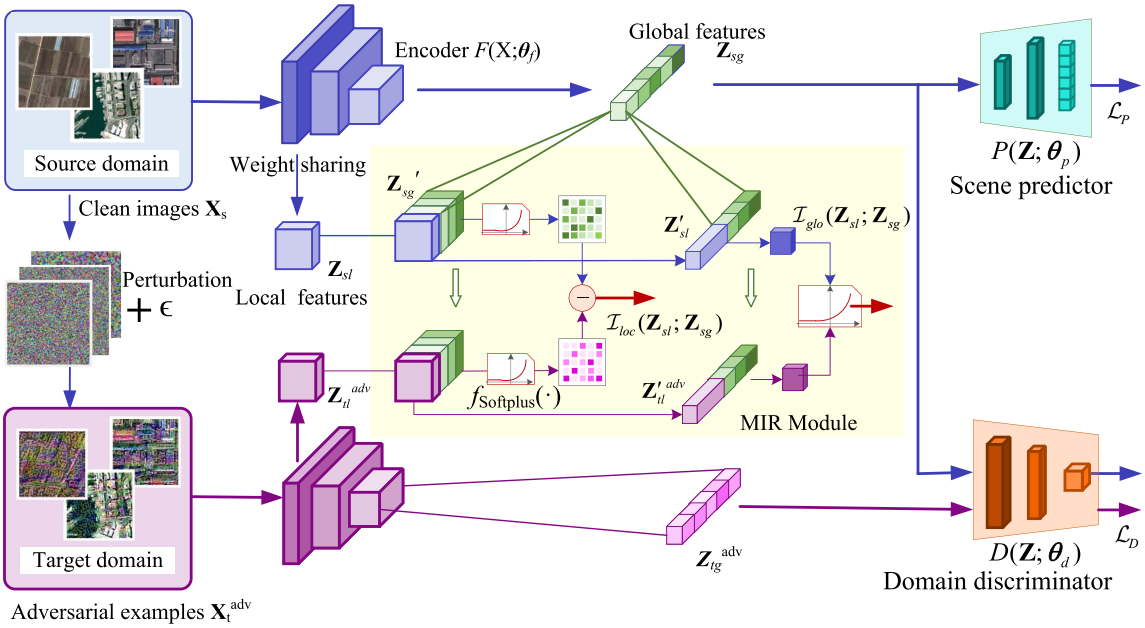

Fig. 2. Overview of the MIDANet structure. The MIDANet framework consists of four primary components: a feature encoder $F(\mathbf{X};\theta_f)$, a domain discriminator $D(\mathbf{Z};\theta_d)$, a scene predictor $P(\mathbf{Z};\theta_p)$, and the MIR module.

data, difficulties in domain adaptation, and the need for more robust MI representations under adversarial perturbations.

Our work builds on these methods but introduces several key differences. First, we propose a MIR module that uses dual-scale MI estimation with both local and global discriminators, allowing us to capture both fine-grained and high-level features. Second, we replace the KL divergence with Jensen–Shannon divergence (JSD), which is more stable and symmetric, addressing training instability. Third, we integrate the MIR module within a generative adversarial framework, enabling domain-invariant feature learning. In addition, the MIR module is designed to be robust to adversarial perturbations in RS data.

In summary, our approach improves on existing methods by combining dual-scale MI estimation, stable divergence, and adversarial robustness, making it more effective for RS tasks with domain adaptation and adversarial challenges.

III. Method

This section first presents the problem description and the proposed solution. Subsequently, a detailed overview of the proposed MIDANet's architecture is provided. Each network component is then described with its specific functionality. Finally, the loss function employed during the network training process is elaborated.

A. Problem Description

Let $\mathcal{X}$ represent a set of clean RS images with corresponding labels $\mathcal{Y}$, sampled from a distribution $\mathcal{S}$, i.e., $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}|_{i=1}^{k}$, where $(\mathbf{x}_i, y_i) \in (\mathcal{X}, \mathcal{Y})$, and k is the number of clean images. The set of unlabeled adversarial samples, denoted as $\mathcal{X}^{\text{adv}}$, is from a different distribution $\mathcal{T}$, such that $\mathcal{T} = \{(\mathbf{x}^{\text{adv}})\}|_{i=1}^{k}$, where $\mathbf{x}^{\text{adv}} \in \mathcal{X}^{\text{adv}}$. It is important to note that $\mathcal{S} \neq \mathcal{T}$. We treat the clean image set as the source domain and the adversarial samples set as the target domain.

The goal is to learn a mapping function $\mathcal{F}_S : \mathcal{X} \to \mathcal{Y}$ from the source domain that can be transferred to the target domain, i.e., $\mathcal{F}_S : \mathcal{X}^{\text{adv}} \to \hat{\mathcal{Y}}^{\text{adv}}$. This learned mapping should correctly predict the label $\hat{y}^{\text{adv}}$ for an unknown adversarial sample, effectively mitigating the adverse impact of adversarial samples on DNN. To achieve this, we utilize a generative adversarial framework focused on learning domain-invariant representations.

B. Network Overview

The architecture of the proposed MIDANet, operating as a generative adversarial network (GAN), is illustrated in Fig. 2. The architecture consists of four key components: the feature encoder $F(\mathbf{X};\theta_f)$, the domain discriminator $D(\mathbf{Z};\theta_d)$, the scene predictor $P(\mathbf{Z};\theta_p)$, and the MIR module, where $\mathbf{X}$ denotes the input sample, $\mathbf{Z}$ represents the encoded features, and θ_f, θ_d, and θ_p are the parameters associated with each respective network component.

The feature encoder extracts relevant features from the input data, regardless of whether the data comes from the source or target domain, mapping the data into a lower dimensional feature space $\mathcal{Z}$. The domain discriminator determines whether the encoded features originate from the source or target domain, helping the encoder learn features transferable across both domains. The discriminator additionally functions as an estimator of deep MI, proficiently quantifying MI within high-dimensional spaces. The scene predictor produces classification results based on the encoded features. The MIR module aims to maximize the MI between the input data $\mathbf{X}$ and its representation $\mathbf{Z}$. This ensures that the extracted features thoroughly and effectively capture the key characteristics of the input samples, thereby

facilitating consistent representations across both the source and target domains.

C. Feature Mapping

The feature encoder is designed to extract features from all input samples from the source or target domain, completing the mapping from input samples, whether sourced from the source or target domain, thereby mapping the input data to the feature space as represented by $F(\mathbf{X};\theta_f) : \mathbf{X} \to \mathrm{Z}$. We implement an encoder compromising four stacked convolutional layers followed by a global pooling layer. The channel configuration for the convolutional layer is $c = [64, 126, 256, 512]$, with a convolutional kernel size of $k = 3 \times 3$ and a stride of $s = 2$. This simple architecture effectively captures domain-invariant features while minimizing computational complexity and the risk of overfitting. When a clean image $\mathbf{X}$ from the source domain is processed through the feature encoder, it produces feature maps at various levels. We select the local features $\mathbf{Z}_{sl} \in \mathbb{R}^{c1\times m\times m}$ from the first layer and the global features $\mathbf{Z}_{sg} \in \mathbb{R}^{c2\times 1\times 1}$ from the fourth layer as inputs for the MI model. Similarly, when an adversarial sample $\mathbf{X}^{\text{adv}}$ is processed, it generates local features $\mathbf{Z}_{tl}^{\text{adv}} \in \mathbb{R}^{c1\times m\times m}$ and global features $\mathbf{Z}_{tg}^{\text{adv}} \in \mathbb{R}^{c2\times 1\times 1}$.

D. MI Maximization

In information theory, MI quantifies the degree of dependence between two random variables [49], [50]. For two discrete random variables X and Y, MI $\mathcal{I}(X;Y)$ is defined as

$$\mathcal{I}(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{4}$$

where $p(x,y)$ represents the joint probability distribution of X and Y, while $p(x)$ and $p(y)$ denote their marginal probability distributions. A larger MI value indicates a stronger association between X and Y.

Despite its significance, accurately computing MI has historically posed a challenge. However, various algorithms have been developed to provide effective estimations. Notably, Donsker, and Varadhan [51] established a lower bound for MI based on the DV representation of the KL divergence

$$\mathcal{I}(X;Y) = \mathcal{D}_{\text{KL}}(J||M) \geq \hat{\mathcal{I}}^{(\text{DV})}(X;Y) \tag{5}$$

where $\hat{\mathcal{I}}^{(\text{DV})}$ denotes the estimated MI. This insight reformulates the objective of maximizing MI as the problem of maximizing this lower bound.

In the context of deep learning, GANs have proven to be a powerful means for learning and approximating data distributions while effectively capturing dependencies between variables. By leveraging GANs, MI can be estimated with greater efficacy, transforming the task of maximizing MI into a generative AdvT challenge.

Drawing from principles in information theory, we have incorporated an MI module into our proposed MIDANet framework. RS images often exhibit high interclass similarity and significant intraclass variability, which complicates the extraction of distinguishable features, particularly in the presence of adversarial perturbations. To enhance the model's feature encoding capability, we specifically design the MI representation module. This module quantifies the amount of information $\mathcal{I}(\mathbf{X};F(\mathbf{X};\theta_f))$ obtained about the encoder's output $F(\mathbf{X};\theta_f)$ through the input images $\mathbf{X}$. Following the methodology of [52], we employ the JSD to estimate the lower bound of MI. Consequently, our objective can be articulated as follows:

$$(\hat{\omega},\ \hat{\theta}_f) = \arg\max_{\omega,\,\theta_f} \mathcal{I}_\omega(\mathbf{X};\ F(\mathbf{X};\ \theta_f)) \tag{6}$$

$$\mathcal{I}_\omega(\mathbf{X};\ F(\mathbf{X};\ \theta_f)) \geq \hat{\mathcal{I}}_\omega^{(\text{JSD})}(\mathbf{X};\ F(\mathbf{X};\ \theta_f)) \tag{7}$$

$$\begin{aligned}\hat{\mathcal{I}}_\omega^{(\text{JSD})}(\mathbf{X};\ F(\mathbf{X};\ \theta_f)) :=\ & \mathbb{E}_{\mathcal{S}}[-f_{\text{sp}}(-D_\omega(\mathbf{X};\ F(\mathbf{X};\ \theta_f)))] \\ & -\mathbb{E}_{\mathcal{S}\times\mathcal{T}}[f(D_\omega(\mathbf{X}';\ F(\mathbf{X};\ \theta_f)))]\end{aligned} \tag{8}$$

where $\hat{\mathcal{I}}_\omega^{(\text{JSD})}$ denotes the estimated MI based on the JSD. D_ω represents the discriminator with parameters ω, while $\mathbf{X}$ and $\mathbf{X}'$ are inputs sampled from the source and target data distributions $\mathcal{S}$ and $\mathcal{T}$, respectively. The function $f_{\text{sp}}(.)$ refers to the SoftPlus activation function. The primary objective of the MIR module is to facilitate the network's learning of a meaningful and informative representation of the original data.

Fig. 3 illustrates the architecture of the MIR module, which consists of a pair of discriminators: the local discriminator $\text{LD}((\mathbf{Z}_J;\mathbf{Z}_M)\ ;\ w_1)$ and the global discriminator $\text{GD}((\mathbf{Z}_J;\mathbf{Z}_M)\ ;\ w_2)$. These discriminators are utilized to estimate MI between the features and the observed data from both local and global perspectives. In this context, $\mathbf{Z}_J$ and $\mathbf{Z}_M$ represent the joint and marginal probability distributions of the features, respectively, while w_1 and w_2 are the parameters associated with the local and global discriminators. The global discriminator comprises two convolutional layers, three fully connected layers, and an activation layer, whereas the local discriminator is composed of three convolutional layers and an activation layer. The activation function used is the SoftPlus function, which yields a probability distribution. The MIR module enforces constraints on both local and global MI on the encoded output, thereby facilitating the generation of high-quality image representations.

It is well recognized that the information contained in low-level feature maps is closely related to the input, while global feature vectors encapsulate the output from the encoder. Therefore, MI $\mathcal{I}(\mathbf{X};F)$ can be expressed as $\mathcal{I}(\mathbf{Z}_{sl};\mathbf{Z}_{sg})$, which can be decomposed into local MI and global MI components.

For local MI, the global features $\mathbf{Z}_{sg}$ from the source domain are initially expanded along the channel dimension to align with the dimensions of the local features, resulting in $\mathbf{Z}'_{sg} \in \mathbb{R}^{c1\times m\times m}$. These expanded global features are then fused with the local features $\mathbf{Z}_{sl}$ to form a joint probability distribution that is subsequently processed by the local discriminator. Simultaneously, $\mathbf{Z}'_{sg}$ is concatenated with local features $\mathbf{Z}_{sl}^{\text{adv}}$ from the target domain, thereby creating the marginal probability distribution for the local discriminator. The LD outputs joint probabilities p_{lj} and marginal probabilities p_{mj} based on these inputs, which are then employed to compute the local MI. The loss function $\mathcal{L}_{\text{LD}}$ for the local discriminator can be

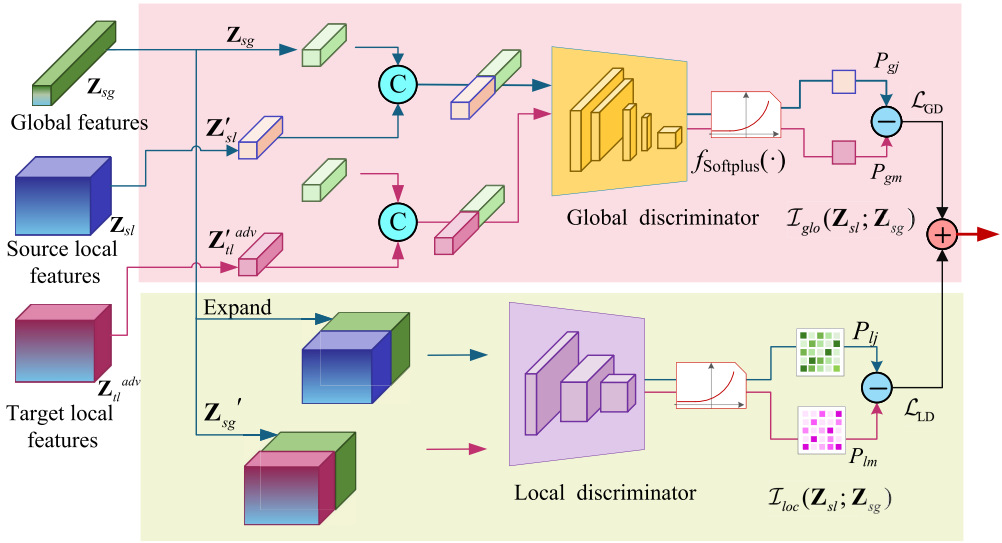

Fig. 3. Structure of the MIR module. The MIR module comprises local and global discriminators that maximize MI between the input image and the encoder output, enhancing the model's representational capacity.

formulated as

$$\mathcal{L}_{\text{LD}} = -(\mathbb{E}_{\mathcal{S}}[-f_{\text{sp}}(-LD_{\omega 1}(\mathbf{Z}_{sl},\ \mathbf{Z'}_{sg}))] - \mathbb{E}_{\mathcal{S}\times\mathcal{T}}[f_{\text{sp}}(LD_{\omega 1}(\mathbf{Z}_{sl}{}^{adv},\ \mathbf{Z'}_{sg}))]). \tag{9}$$

For global MI, the local features from the source domain, denoted as $\mathbf{Z}_{sl}$, undergo convolution and flattening to yield $\mathbf{Z}'_{sl} \in \mathbb{R}^{c2\times 1\times 1}$, thereby aligning with the global features. These transformed local features are then combined with $\mathbf{Z}_{sg}$ to create a joint distribution, which is processed by the global discriminator to produce the joint probability p_{gj}. Similarly, in the target domain, the local features $\mathbf{Z}_{tg}^{\text{adv}}$ are transformed into $\mathbf{Z'}_{tg}^{\text{adv}}$ and concatenated with $\mathbf{Z}_{sg}$ to form the marginal distribution. The global discriminator subsequently outputs the marginal probability p_{gm}. The loss function for the GD is formulated as follows:

$$\mathcal{L}_{\text{GD}} = -(\mathbb{E}_{\mathcal{S}}[-f_{\text{sp}}(-\text{GD}_{\omega 2}(\mathbf{Z'}_{sl},\ \mathbf{Z}_{sg}))] - \mathbb{E}_{\mathcal{S}\times\mathcal{T}}[f_{\text{sp}}(\text{GD}_{\omega 2}(\mathbf{Z'}_{sl}{}^{adv},\ \mathbf{Z}_{sg}))]). \tag{10}$$

The MIR loss $\mathcal{L}_{\text{MIR}}$ is defined as follows:

$$\mathcal{L}_{\text{MIR}} = \alpha\mathcal{L}_{\text{LD}} + \beta\mathcal{L}_{\text{GD}} \tag{11}$$

where the hyperparameters α and β correspond to the weights assigned to the local and global MI losses, respectively. These weights can be tailored based on the specific task requirements of the task at hand in practical applications.

E. Representation Discrimination

The scene predictor, $P(\mathbf{Z};\theta_p)$, is tasked with classifying the feature vectors generated by the encoder, with the objective of correctly predicting the scene label y. The predictor consists of three fully connected layers, each followed by an activation function. The dimensions of the fully connected layers are (512, 256), (256, 128), and (128, n), where n represents the number of scene categories. During training, the scene predictor minimizes the classification loss using the cross-entropy loss function. This process enhances the model's classification of Acc within the source domain

$$\mathcal{L}_P = -\mathbb{E}_{\mathbf{X}\sim\mathcal{S}}[\log P(\mathbf{Z}_{sg};\ \theta_p)]. \tag{12}$$

The domain discriminator, $D(\mathbf{Z};\theta_d)$, serves as a binary classifier designed to determine whether the feature vectors produced by the encoder originate from the source domain or target domain. It classifies the domain of each feature vector by assigning one of two labels, d. The discriminator is structured with two fully connected layers and an activation layer. The first fully connected layer reduces the input feature size from 128 to 64, followed by a normalization layer and a ReLU activation function. The second fully connected layer further reduces the feature size from 64 to 2, corresponding to the binary domain labels. The primary goal of the domain discriminator is to optimize its parameters, θ_d, by minimizing the domain classification loss, which is defined as

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{X}\sim\mathcal{S}}[\log\ D(\mathbf{Z}_{sg};\ \theta_d) - \mathbb{E}_{\mathbf{X}^{\text{adv}}\sim\mathcal{T}}[\log\ (1 - D(\mathbf{Z}_{tg}^{\text{adv}};\ \theta_d)). \tag{13}$$

The domain discriminator maximizes the loss, $\mathcal{L}_D$, to effectively differentiate between source and target domain features, whereas the encoder minimizes this loss to generate domain-invariant features.

F. Overall Loss Function

The overall loss function of the proposed framework incorporates three components: MI loss, scene predictor loss, and domain discriminator loss. The total loss can be formulated as follows:

$$\mathcal{L}_{\text{loss}} = \mathcal{L}_{\text{MI}} + \mathcal{L}_P + \gamma\mathcal{L}_D \tag{14}$$

where γ represents the weight coefficient of $\mathcal{L}_D$ that adjusts the proportion of the loss in the total loss.

IV. Experiments and Result Analysis

A. Data Description

1) RSSCN7 dataset [53]: The RSSCN7 dataset, created by Wuhan University, consists of images collected from Google Earth at four different scales (1:700, 1:1300, 1:2600, 1:5200). Each scale includes 100 images, resulting in a total of 2800 images covering seven scene categories: grassland, fields, industrial areas, rivers, forests, residential areas, and parking lots. Each image has a resolution of 400×400 pixels. This dataset is particularly challenging due to the diverse range of scenarios, due to the diverse range of scenarios, which includes variations in weather, seasons, and image scales.

2) SIRI-WHU dataset [54]: This RS scene classification dataset was developed by Wuhan University, with images predominantly extracted from urban areas in China using Google Earth imagery. The dataset contains 2400 images, each with a resolution of 200×200 pixels and a spatial resolution of 2 m. It includes 12 scene categories: agriculture, commercial areas, harbors, idle land, industrial areas, meadows, overpasses, parks, ponds, residential areas, rivers, and water bodies, with 200 images per category.

3) UC Merced Land-Use dataset [55]: The UC Merced Land-Use dataset, published by the University of California, is designed for urban land-use classification studies. The images were sourced from the United States Geological Survey National Map, with a resolution of 1 ft per pixel. This dataset includes 2100 images, divided into 21 categories, such as including agriculture, buildings, parking lots, sparse residential areas, storage tanks, tennis courts, and freeways. Each category consists of 100 images, with each image having a size of 256×256 pixels.

B. Experimental Settings

1) Data Augmentation: To enhance the diversity of the training data, various data augmentation techniques were applied, aimed at improving the model's generalization capabilities. The augmentation methods included horizontal flipping ($p = 0.5$) and vertical flipping ($p = 0.5$), where p represents the probability of flipping. A random rotation within the range of $(-30°, 30°)$ was also implemented. These transformations help reduce the model's sensitivity to object positioning. In addition, the brightness and contrast of the image were randomly adjusted between 60% and 140% of their original values, with γ representing the adjustment ratio. This augmentation strategy enhances the color tone and diversity of the RS images.

2) Training and inference: The proposed algorithm was implemented using the PyTorch deep learning framework and run on an Ubuntu 20.04 operating system. The training was conducted on a single-card NVIDIA RTX A4000 GPU with 16 GB of memory. The batch size was set to 16, and training was conducted over 100 epochs. The Adam optimizer was used with an initial learning rate of 0.001, a first-order decay rate of β_1 set at 0.9, and a second-order decay rate of β_2 at 0.999. Model performance was validated every five epochs. During inference, a multiscale inference strategy was employed, utilizing scales of [0.5, 0.75, 1, 1.25, 1.5, 1.75], which effectively enhanced scene classification Acc.

TABLE I
Results of Ablation Experiment

Dataset	Baseline	Baseline+MIR	SAcc	RAcc
RSSCN7 [53]	✓		82.7	80.8
		✓	86.2	83.6
SIRI-WHU [54]	✓		83.9	82.6
		✓	86.1	85.4
UC Merced Land-Use [55]	✓		76.6	75.8
		✓	80.9	80.3

TABLE II
Selection of Hyperparameters α and β

α	β	SAcc	RAcc
0.0	0.0	82.7	80.8
1.0	0.0	85.2	82.4
0.0	1.0	83.1	81.6
0.5	0.5	84.5	82.1
1.0	0.5	86.2	83.6

3) Evaluation metrics. For comprehensive model evaluation, four key metrics were employed: Acc, precision (P), recall (R), and F1 score. These metrics are widely used in scene classification tasks and offer a quantitative measure of model effectiveness, with higher values indicating better classification performance and greater robustness. In particular, Acc is assessed using two measures: standard accuracy (SAcc) and robust accuracy (RAcc). SAcc refers to the model's performance on clean, unperturbed images, reflecting its ability to correctly classify data under normal, noise-free conditions.In contrast, RAcc refers to the model's performance on adversarial samples, indicating how well the model can maintain classification Acc when faced with perturbations designed to deceive it.

C. Ablation Study

1) Impact of the MIR module: To assess the effect of the MIR module, a critical element of our network, ablation studies were performed on three datasets. By removing the MIR module, a baseline for comparison was established. As shown in Table I, integrating the MIR module resulted in significant performance improvements, with a 3.5% increase in SAcc and a 2.8% increase in RAcc on the RSSCN7 dataset, along with the largest gains observed on UC Merced Land-Use dataset for both metrics. The MIR module consistently enhanced performance across all datasets, highlighting its positive contribution to both robustness and Acc in RS scene classification.

2) Impact of global and local MI: In our work, we introduce the hyperparameters α and β, which respectively represent the contributions of local MI and global mutual information within the overall MIR module. To evaluate their impact on model performance, we conducted experiments with various values for these parameters, as detailed in Table II.

When both α and β are set to 0.0, excluding MI, the model shows the lowest performance (SAcc $=$ 82.7, RAcc $=$ 80.8), confirming the necessity of incorporating MI. leads to a significant performance improvement, with SAcc increasing by 2.5% and RAcc increasing by 1.6%, indicating its critical role in enhancing feature representation. Similarly, global MI alone

TABLE III
SELECTION OF HYPERPARAMETER γ

γ	SAcc	RAcc
0.2	85.2	81.3
0.4	85.8	81.7
0.5	85.5	82.5
0.6	84.9	83.2
0.8	83.5	83.7
1.0	82.7	82.8

improves results, with SAcc increasing by 0.4% and RAcc increasing by 0.8%, but is less effective than local MI. Combining the two equally yields balanced gains SAcc 84.5%, RAcc 82.1%, showcasing their complementary nature. The best performance is observed with $\alpha = 1.0$ and $\beta = 0.5$, underscoring the dominant role of local MI while leveraging global MI to enhance robustness and domain invariance. These results highlight the importance of jointly optimizing local and global MI, with a stronger emphasis on local MI, to achieve optimal performance on both clean and adversarial samples.

3) Impact of hyperparameter γ: We conducted experiments on the RSSCN7 dataset to evaluate the impact of the hyperparameter γ on model performance. The results in Table III show that as γ increases, robust Acc improves, while standard Acc gradually declines. At $\gamma = 0.2$, SAcc is high (85.2%), but RAcc is relatively low (81.3%), indicating limited adaptation. As γ increases to 0.5 and 0.6, RAcc improves to 82.5% and 83.2%, respectively, while maintaining stable SAcc. The highest RAcc (83.7%) occurs at 0.8, but SAcc declines more significantly, suggesting excessive AdvT negatively impacts source domain classification. Overall, $\gamma = 0.5$ strikes a good balance, effectively improving adversarial robustness while maintaining competitive classification Acc.

D. Generation of Adversarial Samples

To evaluate the robustness of the RS scene interpretation model under different adversarial conditions, we generate adversarial samples using the FGSM and PGD algorithms. For FGSM, the perturbation magnitude ϵ is set at 0.01, 0.03, 0.06, 0.1, and 0.3. For PGD, the iteration step is set to 1, with a perturbation range of (-0.5, 0.5), and perturbation magnitudes ϵ of 0.03, 0.06, 0.1, and 0.3. Adversarial samples with $\epsilon = 0.03$ are used for model training, while samples with other perturbation levels are employed to evaluate the model's adversarial robustness.

Fig. 4 presents examples of adversarial samples, illustrating how adversarial perturbations affect clean images. These clean samples are derived from a RS image scene classification task based on the UC Merced Land-Use dataset, which consists of 2100 images categorized into 21 land-use classes. Specifically, the first image in Fig. 4 belongs to the airplane category, while the second image is from the parking lot category.

As the perturbation magnitude increases, the disruption to the clean images becomes more pronounced, although the core semantics of the images remain visually discernible. This is important because even when adversarial perturbations are visually subtle, they can still lead to significant classification errors.

In comparison to FGSM, adversarial samples generated using PGD appear more deceptive and aggressive. Despite exhibiting minimal visual differences from the original images, PGD-based adversarial samples are particularly effective at misleading deep learning models, which highlights their severity in terms of model vulnerability.

To further emphasize the robustness of our model, Fig. 4 visualizes different types of adversarial samples, as well as samples with varying perturbation magnitudes. This illustrates that we did not rely on a single sample for evaluation, but instead tested the model's performance on a wide range of adversarial examples. By doing so, we demonstrate the model's ability to generalize and maintain robustness across diverse adversarial conditions.

Fig. 5 presents the frequency distributions of grayscale values across the three image channels (RGB). Although the samples in the upper and lower rows exhibit minimal visual differences—both belonging to the building scene category—a detailed statistical analysis of the grayscale values in the RGB channels reveals significant variations in data frequency. These variations indicate that, despite the visual similarity between the adversarial and clean images, their underlying data distributions are quite different.

This discrepancy in data distributions plays a crucial role in how deep learning models perceive and classify these images. Although the images may look nearly identical to the human eye, the statistical differences in pixel intensities and color distributions can cause the model to produce divergent prediction outputs, potentially leading to misclassifications. This observation underscores the importance of addressing the distributional differences between adversarial examples and clean samples, which is the primary challenge in adversarial example classification.

This insight provides the foundation for applying domain adaptation techniques, which are incorporated into our method to better recognize and classify adversarial samples by adapting to these subtle distributional shifts.

E. Quantitative Comparison

1) Results on the RSSCN7 dataset: The effectiveness of the MIDANet framework, which incorporates domain AdvT, is compared with standard training (StaT) and AdvT approaches. In the StaT method, 70% of clean images are used for training, while the remaining 30%, are reserved for validation, with adversarial samples being used solely for testing. The AdvT method, by contrast, integrates adversarial samples into the training process to adjust model parameters. Our approach involves using clean images as the source domain and adversarial samples as the target domain for joint training.

Table IV presents the classification results for different scene categories on the RSSCN7 dataset for these training strategies. Our method achieved average SAcc and RAcc scores of 86.2% and 83.6%, respectively, reflecting improvements of 6.4% and 8.2% over AdvT. While StaT demonstrates superior performance on clean samples relative to both adversarial and domain adaptation training, it exhibits significant weaknesses

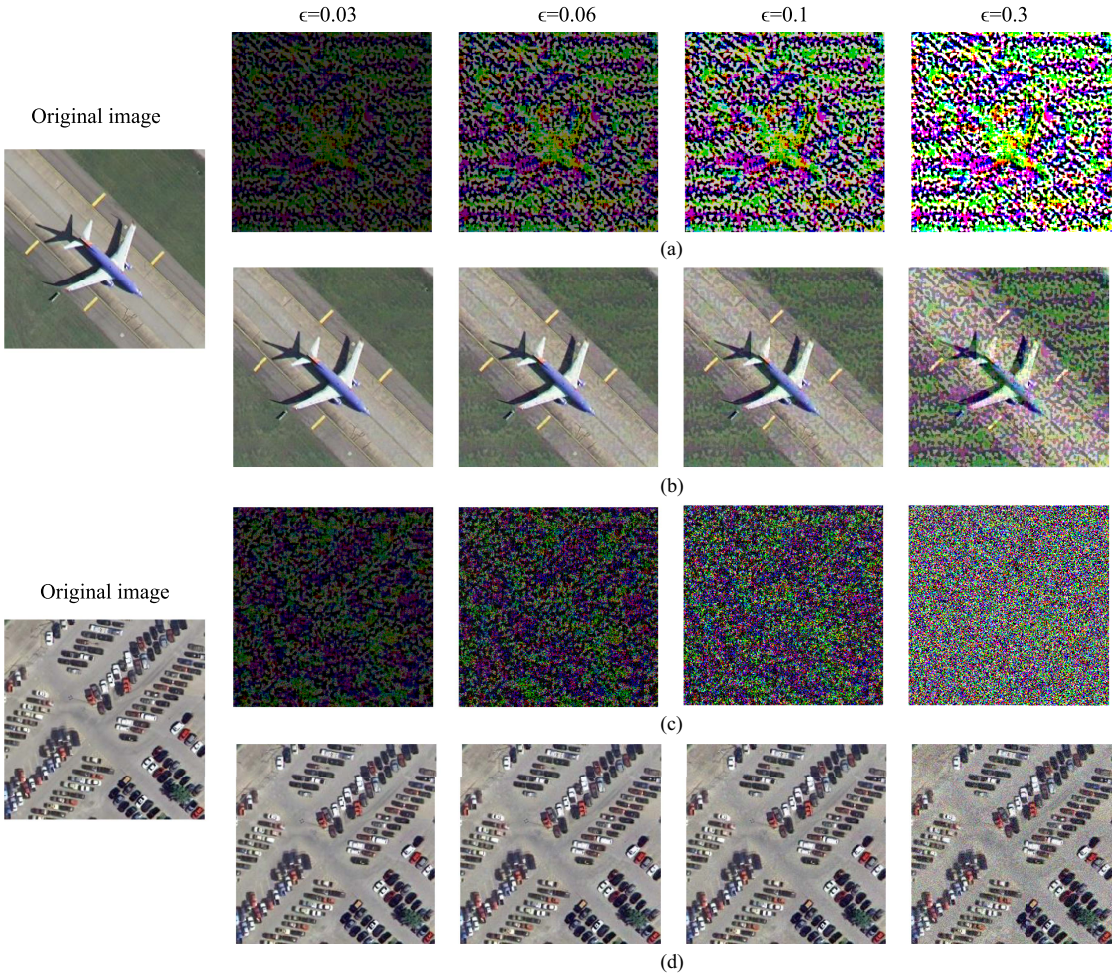

Fig. 4. Examples of adversarial samples at various perturbations to the original sample create an adversarial sample. In comparison to FGSM adversarial samples, PGD adversarial samples exhibit greater deception and intensity. (a) Amplified perturbation (magnification = 20). (b) FGSM adversarial samples. (c) Amplified perturbation (magnification = 20). (d) PGD adversarial samples.

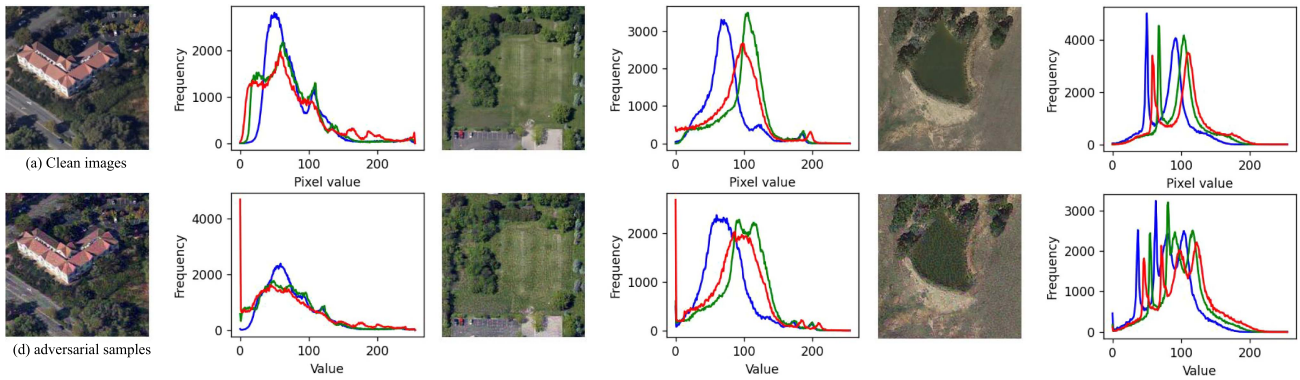

Fig. 5. Data frequency distributions of grayscale values. (a) Frequency distributions of RGB channels (red—R channel, green—G channel, and blue—B channel) for clean images. (b) Frequency distributions for adversarial samples. While the visual differences between (a) and (b) are subtle, the frequency distributions of grayscale values show significant variation.

when applied to adversarial samples, achieving a low RAcc of 74.3%. Conversely, AdvT provides greater stability on adversarial samples, particularly excelling in the forest and industrial categories, though it results in a comparatively low average SAcc of 79.8%. Benefiting from MIR, our method not only improves the robustness of the model but also improves strong classification performance on clean samples, successfully balancing adversarial robustness with SAcc during training. In particular, for the residential category, both SAcc and RAcc reach 95%.

TABLE IV
SCENE CLASSIFICATION RESULTS ON THE RSSCN7 DATASET

Scene	StaT		AdvT		MIDANet	
	SAcc	RAcc	SAcc	RAcc	SAcc	RAcc
Grass	85.0	70.0	74.2	70.0	79.2	78.3
Field	85.8	79.2	83.3	80.0	90.1	83.3
Industry	77.5	62.5	85.0	83.3	79.2	73.3
RiverLake	83.3	38.3	69.2	49.2	83.3	78.3
Forest	96.6	97.5	99.2	99.2	94.1	95.8
Resident	92.5	86.7	91.7	89.2	95.8	95.8
Parking	89.2	85.8	55.8	52.5	80.8	80.0
Mean	87.4	74.3	79.8	74.8	86.2	83.6

TABLE V
COMPARISON OF CLASSIFICATION PERFORMANCE FOR EACH CATEGORY IN THE RSSCN7 DATASET

Scene	Clean image			Adversarial sample		
	P	R	F1	P	R	F1
Grass	89.6	79.2	84.1	84.7	78.3	81.4
Field	80.2	90.8	85.2	78.7	83.3	81.0
Industry	81.2	79.2	80.2	78.6	73.3	75.9
RiverLake	92.6	83.3	87.7	92.2	78.3	84.7
Forest	92.6	94.1	93.4	90.6	95.8	93.1
Resident	85.8	95.8	90.6	82.7	95.8	88.8
Parking	82.9	80.8	81.9	78.7	80.0	79.3

TABLE VI
SCENE CLASSIFICATION RESULTS ON THE SIRI-WHU DATASET (%)

Scene	StaT		AdvT		MIDANet	
	SAcc	RAcc	SAcc	RAcc	SAcc	RAcc
Agriculture	88.3	48.3	100.0	100.0	96.7	96.7
Commercial	96.7	80.0	90.0	90.0	90.0	90.0
Harbor	83.3	35.0	71.6	70.0	88.3	88.3
Idle land	86.7	83.3	78.3	78.0	93.3	93.3
Industrial	93.3	85.0	91.7	91.7	93.3	91.7
Meadow	81.7	70.0	40.0	39.3	75.0	73.3
Overpass	85.0	75.0	71.7	71.7	71.7	71.7
Park	78.3	88.3	48.3	50.0	83.3	83.3
Pond	86.7	45.0	88.3	88.0	78.3	76.7
Residential	91.7	86.7	88.3	87.9	90.0	86.7
River	71.7	30.0	65.0	63.3	75.0	75.0
Water	100.0	30.0	100.0	100.0	98.3	98.3
Mean	86.9	63.1	77.8	77.5	86.1	85.4

Table V details the classification performance of the proposed algorithm on both clean and adversarial samples across various scene categories within the RSSCN7 dataset. For clean samples, the P, R, and F1 scores are robust across all categories, with particularly strong results in the forest category, achieving scores of 92.6%, 94.1%, and 93.4%, respectively. Although there is a slight decline in performance on adversarial samples, the decrease is minimal, with the metrics remaining comparable to those for clean samples. The elevated scores in both the forest and residential categories underscore the robustness of the algorithm. Overall, the proposed algorithm demonstrates significant resilience to adversarial perturbations while maintaining high classification Acc across both clean and adversarial samples, particularly in critical categories.

2) Results on the SIRI-WHU dataset: As shown in Table VI, the StaT method achieved an average SAcc of 86.9%, indicating strong classification performance; however, the model's robustness against adversarial samples was notably compromised, resulting in an average RAcc of only 63.1%. This underscores the model's susceptibility to adversarial attacks under StaT conditions.

By contrast, AdvT significantly enhanced the robustness of the model, increasing the average RAcc to 77.5%, which represents a marked improvement over StaT. Particularly, categories, such as farmland, ponds, and residential areas, exhibited RAcc values approaching or even reaching 100.0%, demonstrating the effectiveness of AdvT in bolstering robustness for these categories. Nevertheless, categories, such as grassland and overpasses, recorded lower RAcc values of 39.3% and 71.7%, respectively.

Our domain AdvT method performed admirably across all categories, achieving an average SAcc of 86.1% and RAcc of 85.4%, closely aligning with the SAcc of StaT. This indicates that our approach not only sustains high classification performance on clean samples but also markedly improves robustness against adversarial samples.

TABLE VII
COMPARISON OF CLASSIFICATION PERFORMANCE FOR EACH CATEGORY IN THE SIRI-WHU DATASET

Scene	Clean image			Adversarial sample		
	P	R	F1	P	R	F1
Agriculture	89.2	96.7	92.8	89.2	96.7	92.8
Commercial	90.0	90.0	90.0	88.5	90.0	89.3
Harbor	80.3	88.3	84.1	80.3	88.3	84.1
Idle land	87.5	93.3	90.3	86.2	93.3	89.6
Industrial	87.5	93.3	90.3	87.3	91.7	89.4
Meadow	93.8	75.0	83.3	93.6	73.3	82.2
Overpass	74.1	71.7	72.9	72.9	71.7	72.3
Park	89.3	83.3	86.2	86.2	83.3	84.8
Pond	72.3	78.3	75.2	71.8	76.7	74.2
Residential	90.0	90.0	90.0	89.7	86.7	88.1
River	86.5	75.0	80.4	86.5	75.0	80.4
Water	95.2	98.3	96.7	95.2	98.3	96.7

Our domain AdvT method performed admirably across all categories, achieving an average SAcc of 86.1% and RAcc of 85.4%, closely aligning with the SAcc of StaT. This indicates that our approach not only sustains high classification performance on clean samples but also markedly improves robustness against adversarial samples.

Table VII presents the classification metrics for both clean and adversarial samples obtained using the proposed method on the SIRI-WHU dataset. Overall, the classification performance for clean and adversarial samples is notably similar. Most categories, such as farmland, commercial areas, ports, open spaces, industrial areas, and water bodies, demonstrate P, R, and F1 scores for adversarial samples that closely match those for clean samples, indicating a high degree of robustness. For example, in the water category, the P, R, and F1 scores for both clean and adversarial images are 95.2%, 98.3%, and 96.7%, respectively. The minimal variations in P, R, and F1 scores between clean and adversarial samples across the majority of scene categories suggest that our framework effectively preserves model stability in the presence of adversarial samples.

3) Results on the UC Merced Land-Use dataset: In the complex scene classification task comprising 21 classes, as illustrated in Table VIII, our proposed approach achieved average SAcc and RAcc of 80.9% and 80.3%, respectively, reflecting improvements of 4.8% and 5.3% over AdvT. In comparison to StaT, the domain adaptation strategy exhibited a 1.9% reduction in SAcc but a notable 7.6% enhancement in RAcc.

TABLE VIII
SCENE CLASSIFICATION RESULTS ON THE UC MERCED LAND-USE DATASET (%)

Scene	StaT		AdvT		MIDANet	
	SAcc	RAcc	SAcc	RAcc	SAcc	RAcc
Agriculture	76.7	76.7	96.7	96.7	83.3	90.0
Airplane	83.3	63.3	70.0	66.7	76.7	70.0
Baseball diamond	86.7	63.3	56.7	56.7	86.7	80.0
Beach	96.7	86.7	100.0	96.7	100.0	93.3
Buildings	80.0	63.3	86.7	83.3	76.7	73.3
Chaparral	96.7	96.7	100.0	100.0	100.0	100.0
Dense residential	86.7	66.7	90.0	86.7	76.7	73.3
Forest	96.7	96.7	86.7	86.7	100.0	100.0
Freeway	73.3	66.7	53.3	56.7	76.7	80.0
Golf course	76.7	46.7	73.3	66.7	93.3	90.0
Harbor	100.0	96.7	100.0	96.7	96.7	96.7
Intersection	93.3	60.0	43.3	43.3	63.3	63.3
Medium residential	60.0	50.0	53.3	53.3	73.3	76.7
Home park	86.7	86.7	80.0	80.0	73.3	73.3
Overpass	70.0	63.3	90.0	90.0	60.0	66.7
Parking lot	93.3	90.0	100.0	100.0	93.3	93.3
River	86.7	66.7	83.3	86.7	80.0	76.7
Runway	96.7	93.3	96.7	96.7	96.7	96.7
Sparse residential	70.0	90.0	33.3	33.3	60.0	63.3
Storage tanks	66.7	33.3	46.7	40.0	70.0	66.7
Tennis court	62.5	70.8	58.3	58.3	62.5	62.5
Mean	82.8	72.7	76.1	75.0	80.9	80.3

While the StaT method demonstrates superior performance on clean samples, its robustness against adversarial samples remains comparatively inadequate, yielding an average RAcc of 72.7%. AdvT does enhance classification performance for adversarial samples; however, declines in Acc are observed across some scene categories. The chaparral and forest categories consistently achieved high performance across all three training strategies, particularly under the domain AdvT framework, where the RAcc reached 100%. This outcome suggests that the model employing MI exhibits robust feature extraction and classification capabilities in these categories, demonstrating resilience to adversarial perturbations.

F. Qualitative Visualization

To further illustrate the advantages of the proposed algorithm, we employ t-SNE to visualize the data distribution of clean images and adversarial samples. Fig. 6 shows the distribution characteristics across the RSSCN7, SIRI-WHU, and UC Merced Land-Use datasets. In these visualizations, the source domain represents the clean samples, while the target domain denotes the adversarial samples with $\epsilon = 0.03$. The left side of the figure highlights a significant discrepancy between the clean images and the adversarial samples, revealing distinct centers of data distribution for each domain. For example, in Fig. 6(a), the centers of clean sample data for the RSSCN7 dataset are positioned higher in the data space and demonstrate a wider spread, while the centers of the adversarial samples are situated lower and appear more compact. Conversely, the right side of the figure illustrates the feature distributions with a relatively tight alignment between the source and target domains. This alignment suggests that the domain AdvT strategy effectively reduces the distribution discrepancy. Importantly, in Fig. 6(c), the UC Merced Land-Use dataset exhibits a nearly complete overlap of feature distributions between clean and adversarial images, highlighting the strategy's effectiveness in addressing complex scenarios.

Fig. 7 presents a t-SNE visualization that contrasts the original data distribution with the feature distribution of adversarial samples ($\epsilon = 0.03$). The left column illustrates the feature distribution of the original data, while the right column shows the feature distribution after applying domain adaptation.

The comparison indicates that the original data distribution exhibits considerable overlap and dispersion among data points from different classes, especially at the class boundaries. This overlap poses a risk to the classification model's Acc. Conversely, our method yields a more compact and distinctly separated feature distribution, characterized by tighter clustering of intraclass data points. This improvement signifies that our approach effectively enhances intraclass consistency, leading to greater robustness and Acc in the classification model.

TABLE IX
PERFORMANCE COMPARISON UNDER VARIOUS PERTURBATIONS ON THE RSSCN7 DATASET

Sample	Perturbation	Image entropy	StaT RAcc	AdvT RAcc	MIDANet RAcc
FGSM	$\epsilon = 0.01$	6.503	86.6	79.5	85.1
	$\epsilon = 0.03$	6.572	76.9	74.8	83.6
	$\epsilon = 0.06$	6.682	74.2	64.6	76.3
	$\epsilon = 0.10$	6.815	57.7	49.6	53.3
	$\epsilon = 0.30$	7.237	25.9	21.1	26.7
PGD	$\epsilon = 0.03$	6.551	85.6	77.1	84.2
	$\epsilon = 0.06$	6.596	84.9	74.2	84.1
	$\epsilon = 0.10$	6.661	82.6	73.3	82.5
	$\epsilon = 0.30$	6.994	56.9	53.3	65.0

G. Correlation Analysis

Adversarial perturbations not only impair model performance but also modify the complexity of the original images, thereby increasing the uncertainty associated with image information. To explore the relationship among adversarial perturbations, information uncertainty, and RAcc, we evaluated the classification performance of various methods on adversarial samples from the RSSCN7 dataset across different perturbation levels.

We employed image information entropy to quantify the uncertainty or complexity of the information, using the following formula: $H(X) = -\sum_{i=1}^{L} p(x_i)\log_2 p(x_i)$, where $H(X)$ denotes the entropy of the image X, L represents the total number of gray levels, and $p(x_i)$ is the probability of pixels exhibiting the gray level x_i in the image.According to Table IX, an increase in the perturbation magnitude ϵ leads to a decline in classification Acc across all training methods. At lower perturbation levels ($\epsilon = 0.01$ or $\epsilon = 0.03$), the models demonstrate relatively stable performance, experiencing minimal effects from adversarial perturbations. However, at a higher perturbation magnitude ($\epsilon = 0.30$), the models exhibit a significant reduction in robustness. Notably, the method proposed in this study slightly outperforms both standard and AdvT approaches under these conditions.

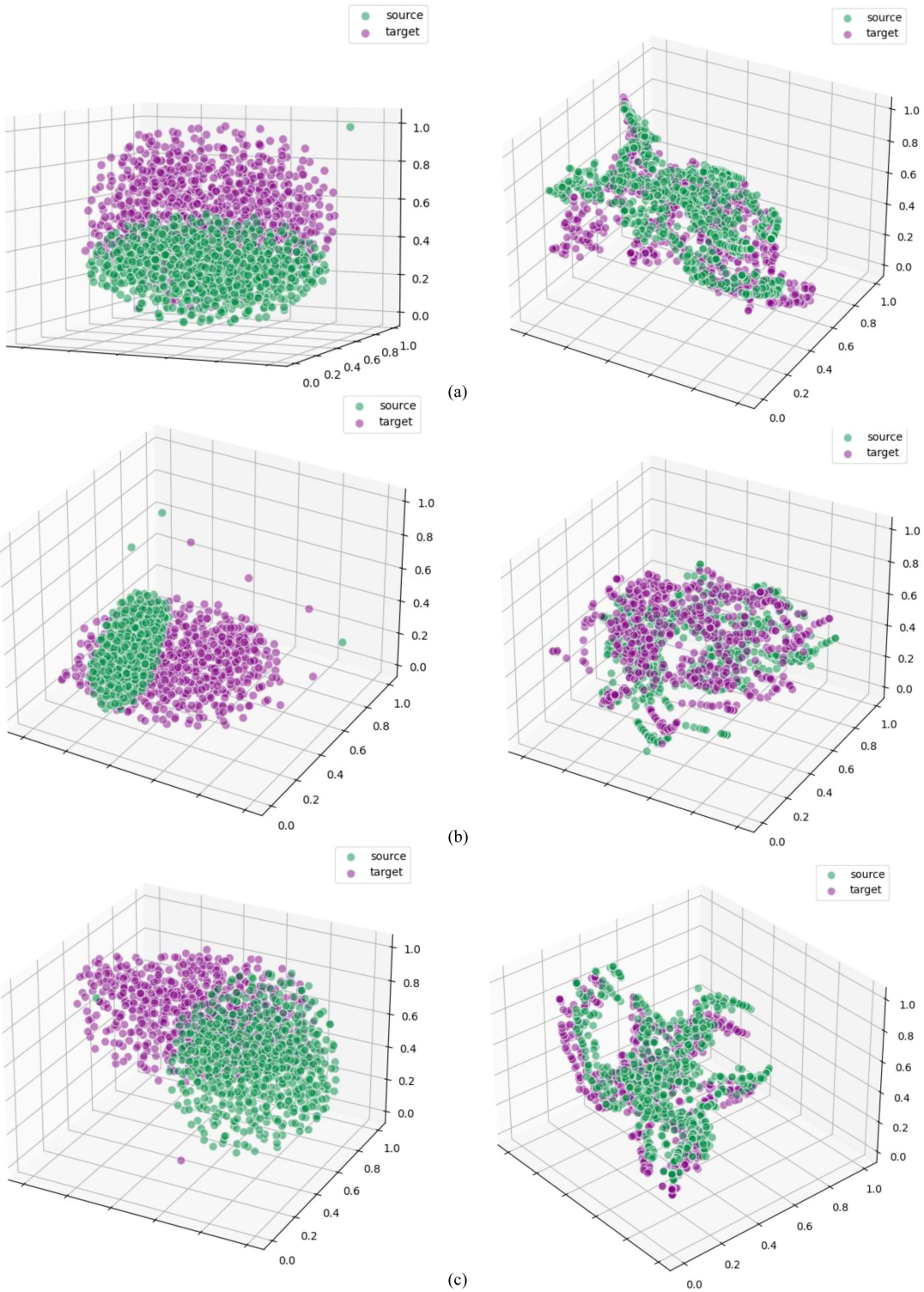

Fig. 6. t-SNE Visualization of data distribution for clean images and adversarial samples. The left panel illustrates the original data distribution, while the right panel shows feature alignment. Despite the distribution discrepancy, clean images (source domain) and adversarial samples with $\epsilon = 0.03$ (target domain) are effectively aligned. (a) RSSCN7 dataset. (b) SIRI-WHU dataset. (c) UC Merced Land-Use dataset.

In addition, image entropy increases progressively with rising perturbation magnitudes ϵ. This trend indicates that adversarial perturbations not only reduce classification Acc but also significantly complicate the images, thereby increasing the uncertainty of the information contained within them. There exists a positive correlation between higher image entropy and greater perturbation magnitudes, while a negative correlation is observed between higher image entropy and lower classification Acc. For example, in the case of FGSM adversarial samples, at $\epsilon = 0.01$, the image entropy measures 6.503, and the domain

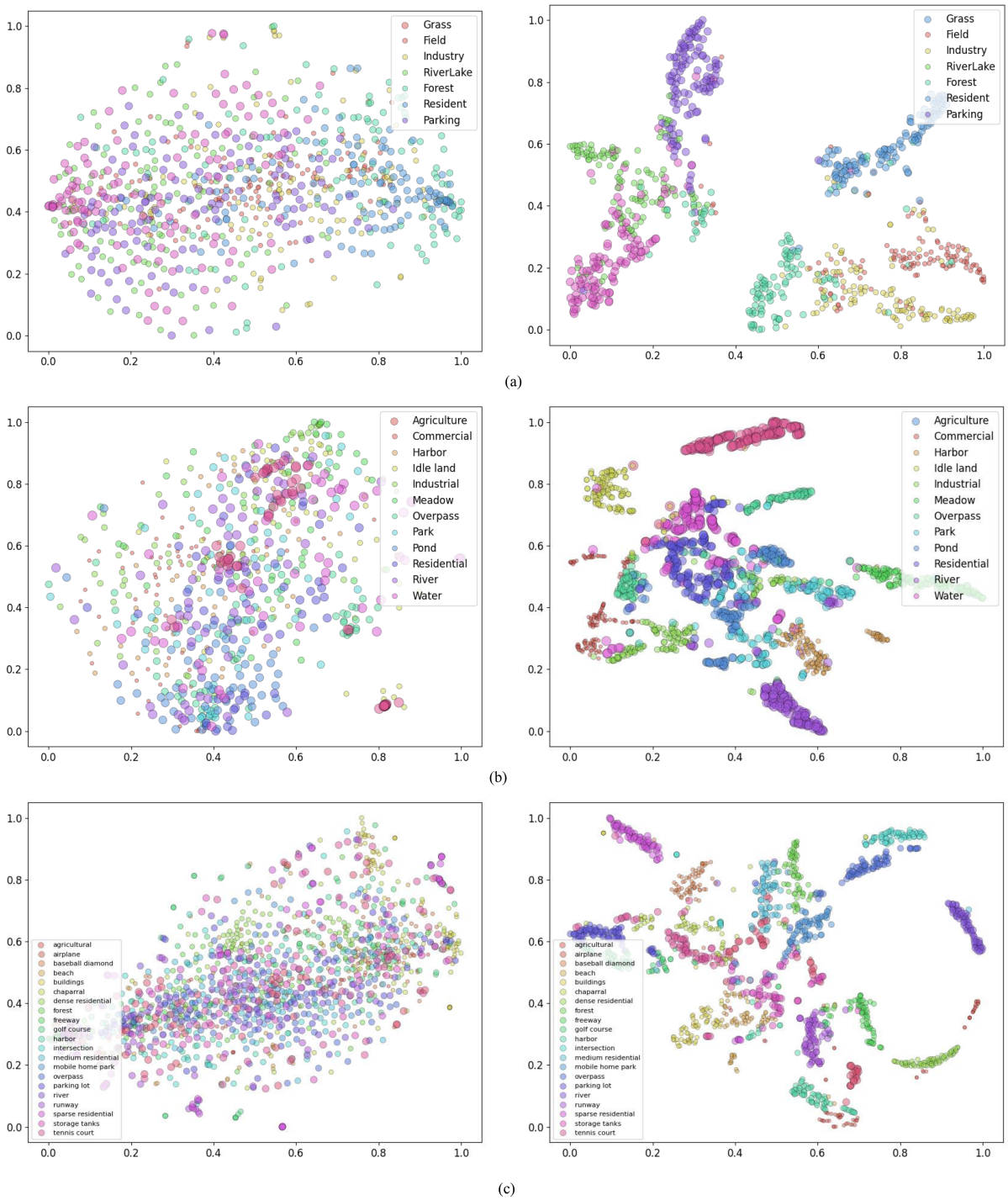

Fig. 7. t-SNE visualization of adversarial sample classification. The left panel displays the original data distribution of adversarial samples ($\epsilon = 0.03$), while the right panel shows the feature distribution across scene categories. Our approach improves intra-class compactness and inter-class separability within the feature space. (a) RSSCN7 dataset. (b) SIRI-WHU dataset. (c) UC Merced Land-Use dataset.

adaptation RAcc is 85.1%. By contrast, at $\epsilon = 0.30$, the image entropy rises to 7.237, while RAcc falls to 26.7%.Table X presents a comprehensive comparison of classification accuracies under varying perturbation magnitudes on the SIRI-WHU dataset. For adversarial samples generated using the FGSM, a significant decline in classification Acc is observed as the perturbation magnitude increases, particularly at perturbations of 0.10 and 0.30. By contrast, PGD adversarial samples exhibit a more gradual decrease in Acc, with domain adaptation models demonstrating enhanced robustness. As the perturbation magnitude escalates, image entropy similarly increases, peaking at $\epsilon = 0.3$. This rise in entropy correlates with the reduction in classification Acc, indicating that higher image entropy reflects a greater intensity of adversarial perturbations, which adversely affects model performance. The data reveal a direct relationship whereby increased image entropy is associated with diminished

TABLE X
PERFORMANCE COMPARISON UNDER VARIOUS PERTURBATIONS ON THE SIRI-WHU DATASET

Sample	Perturbation	Image entropy	StaT RAcc	AdvT RAcc	MIDANet RAcc
FGSM	$\epsilon = 0.01$	6.278	85.8	77.6	86.0
	$\epsilon = 0.03$	6.280	75.0	77.6	85.4
	$\epsilon = 0.06$	6.552	63.1	60.0	73.5
	$\epsilon = 0.10$	6.686	47.1	48.3	51.0
	$\epsilon = 0.30$	7.107	16.2	17.2	18.7
PGD	$\epsilon = 0.03$	6.387	80.9	74.7	85.0
	$\epsilon = 0.06$	6.462	78.4	74.4	83.6
	$\epsilon = 0.10$	6.549	73.7	70.6	80.1
	$\epsilon = 0.30$	6.927	41.4	48.1	43.5

TABLE XI
PERFORMANCE COMPARISON UNDER VARIOUS PERTURBATIONS ON THE UC MERCED LAND-USE DATASET

Sample	Perturbation	Image entropy	StaT RAcc	AdvT RAcc	MIDANet RAcc
FGSM	$\epsilon = 0.01$	6.821	83.2	75.6	80.5
	$\epsilon = 0.03$	6.889	81.8	75.8	80.2
	$\epsilon = 0.06$	6.986	72.7	69.1	80.1
	$\epsilon = 0.10$	7.095	55.9	60.8	51.3
	$\epsilon = 0.30$	7.429	5.87	16.1	21.3
PGD	$\epsilon = 0.03$	6.870	82.5	75.5	81.0
	$\epsilon = 0.06$	6.912	80.4	74.8	80.6
	$\epsilon = 0.10$	6.967	70.2	74.0	77.4
	$\epsilon = 0.30$	7.240	52.0	61.0	54.3

classification Acc.Table XI presents the outcomes for the UC Merced land use dataset. Our methodology achieves an RAcc of approximately 80% for FGSM adversarial samples at lower perturbation levels $\epsilon \leq 0.06$, which is about 5% higher than that obtained through AdvT. In the case of PGD adversarial samples, AdvT maintains relatively stable performance, indicating effective adaptation to PGD perturbations. Furthermore, domain adaptation training significantly surpasses both standard and AdvT at lower perturbation levels. As the magnitude of perturbations increases, image entropy also rises, likely due to the introduction of additional uncertainty and noise by these perturbations. This increase in entropy leads to a more dispersed pixel distribution, which may contribute to a reduction in classification Acc.

V. DISCUSSION

This study introduces MIDANet, a framework integrating MI and UDA to enhance adversarial robustness in RS scene classification. Experimental results demonstrate its effectiveness in improving both standard and robust Acc, outperforming conventional AdvT methods.

However, several aspects require further exploration. First, while MIDANet enhances robustness against FGSM and PGD adversarial examples, its generalization to more sophisticated adversarial perturbations remains uncertain. Second, our findings suggest a correlation between adversarial perturbations and image information entropy, highlighting the need for entropy-aware defense mechanisms. Third, although designed for scene classification, MIDANet's methodology could be extended to other RS tasks like object detection and segmentation.

In addition, the evaluation is conducted on relatively small datasets with limited scene diversity. Future research should validate its scalability on large-scale datasets and refine robustness metrics for a more comprehensive assessment.

VI. CONCLUSION

We propose MIDANet to improve adversarial robustness in RS scene classification by leveraging MI and UDA. The MIR module significantly enhances feature representations, yielding notable Acc improvements across multiple datasets. In addition, the domain adaptation mechanism effectively mitigates distributional shifts, improving model resilience against adversarial perturbations. Comprehensive experiments using FGSM and PGD method, which generate adversarial samples with varying characteristics and magnitudes, validate the effectiveness of MIDANet. Moreover, our findings reveal a correlation between adversarial perturbations and image information entropy, providing valuable insights into the impact of adversarial perturbations on RS classification.

Despite promising results, further validation on large-scale datasets and more refined robustness metrics are needed. our future work will focus on 1) extending research to large-scale datasets for a more thorough evaluation of model performance across diverse scenarios; 2) developing more comprehensive metrics for robustness assessment; and 3) applying our model to a wider range of applications, such as object detection, semantic segmentation, and change detection.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

ACKNOWLEDGMENT

The authors would like to thank LetPub (www.letpub.com.cn) for its linguistic assistance during the preparation of this manuscript.

REFERENCES

[1] C. Sitaula, S. KC, and J. Aryal, "Enhanced multi-level features for very high resolution remote sensing scene classification," *Neural Comput. Appl.*, vol. 36, no. 13, pp. 7071–7083, 2024.
[2] S. Dutta and M. Das, "Remote sensing scene classification under scarcity of labelled samples—a survey of the state-of-the-arts," *Comput. Geosci.*, vol. 171, 2023, Art. no. 105295.
[3] Y. Yang, X. Tang, Y. M. Cheung, X. Zhang, and L. Jiao, "SAGN: Semantic-aware graph network for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1011–1025, 2023.
[4] X. Wang, W. Chen, J. Yin, L. Wang, and H. Guo, "Risk assessment of flood disasters in the Poyang lake area," *Int. J. Disaster Risk Reduct.*, vol. 100, 2024, Art. no. 104208.
[5] M. Rahnemoonfar, T. Chowdhury, and R. Murphy, "Rescuenet: A high resolution UAV semantic segmentation dataset for natural disaster damage assessment," *Sci. Data.*, vol. 10, no. 1, 2023, Art. no. 913.
[6] C. Shi, M. Zhang, Z. Lv, Q. Miao, and C.-M. Pun, "Universal object-level adversarial attack in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens..*, vol. 61, 2023, Art. no. 5532714.

[7] Y. Xu and P. Ghamisi, "Universal adversarial examples in remote sensing: Methodology and benchmark," *IEEE Trans. Geosci. Remote Sens.*., vol. 60, 2022, Art. no. 5619815.

[8] L. Chen, Z. Xu, Q. Li, J. Peng, S. Wang, and H. Li, "An empirical study of adversarial examples on remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*., vol. 59, no. 9, pp. 7419–7433, Sep. 2021.

[9] X. Li, C. Wen, Y. Hu, and N. Zhou, "RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 124, Art. no.103497, 2023.

[10] R. Datla and N. Perveen, "Learning scene-vectors for remote sensing image scene classification," *Neurocomputing.*, vol. 587, Art. no.127679, 2024.

[11] W. Wang, L. Xing, P. Ren, Y. Jiang, G. Wang, and B. Liu, "Subspace prototype learning for few-shot remote sensing scene classification," *Signal Process.*, vol. 208, 2023, Art. no. 108976.

[12] S. Huang, W. Fu, Z. Zhang, and S. Liu, "Global-local fusion based on adversarial sample generation for image-text matching," *Inf. Fusion.*, vol. 103, 2024, Art. no. 102084.

[13] J. Zhang, Y. Huang, W. Wu, and M. R. Lyu, "Transferable adversarial attacks on vision transformers with token gradient regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 16415–1624.

[14] J. Kang et al., "Adversarial attacks and defenses for semantic communication in vehicular metaverses," *IEEE Wireless Commun.*, vol. 30, no. 4, pp. 48–55, Aug. 2023.

[15] R. Venkatesh, E. Wong, and Z. Kolter, "Adversarial robustness in discontinuous spaces via alternating sampling & descent," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2023, pp. 4662–4671.

[16] Y. Zhu and Y. Jiang, "Imperceptible adversarial attacks against traffic scene recognition," *Soft Comput.*, vol. 25, no. 20, pp. 13069–13077, 2021.

[17] C. Yu, Z. Zhang, H. Li, J. Sun, and Z. Xu, "Meta-learning-based adversarial training for deep 3D face recognition on point clouds," *Pattern Recognit.*, vol. 134, 2023, Art. no. 109065.

[18] N. Ruiz et al., "Simulated adversarial testing of face recognition models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4145–4155.

[19] G. Tang, W. Yao, C. Li, T. Jiang, and S. Yang, "Black-box adversarial patch attacks using differential evolution against aerial imagery object detectors," *Eng. Appl. Artif. Intell.*, vol. 137, 2024, Art. no. 109141.

[20] Y. Chen, "The risk and opportunity of adversarial example in military field," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 100–107.

[21] L. Zhao, T. Liu, X. Peng, and D. Metaxas, "Maximum-entropy adversarial data augmentation for improved generalization and robustness," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 14435–14447, 2020.

[22] H. Wang, C. Xiao, J. Kossaifi, Z. Yu, A. Anandkumar, and Z. Wang, "AugMax: Adversarial composition of random augmentations for robust training," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 237–250, 2021.

[23] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data augmentation can improve robustness," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 29935–29948, 2021.

[24] G. Li, S. Ding, J. Luo, and C. Liu, "Enhancing intrinsic adversarial robustness via feature pyramid decoder," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 800–808.

[25] C. Sun, L. Zou, C. Fan, Y. Shi, and Y. Liu, "Enhancing adversarial attack transferability with multi-scale feature attack," *Int. J. Wavelets Multiresol. Inf. Process.*, vol. 19, no. 02, 2021, Art. no. 2050076.

[26] P. Agrawal, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Impact of attention on adversarial robustness of image classification models," in *Proc. IEEE Int. Conf. Big Data.*, Dec. 2021, pp. 3013–3019.

[27] Z. Li, C. Feng, M. Wu, H. Yu, J. Zheng, and F. Zhu, "Adversarial robustness via attention transfer," *Pattern Recogn. Lett.*, vol. 146, pp. 172–178, 2021.

[28] S. Jain and T. Dutta, "Towards understanding and improving adversarial robustness of vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 24736–24745.

[29] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected convnet for aerial scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4911–4926, 2020.

[30] Z. Li, K. Xu, J. Xie, Q. Bi, and K. Qin, "Deep multiple instance convolutional neural networks for learning robust scene representations," *IEEE Trans. Geosci. Remote Sens.*., vol. 58, no. 5, pp. 3685–3702, May 2020.

[31] Q. Bi, B. Zhou, K. Qin, Q. Ye, and G.-S. Xia, "All grains, one scheme (AGOS): Learning multigrain instance representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5629217.

[32] J. Tack, S. Yu, J. Jeong, M. Kim, S. J. Hwang, and J. Shin, "Consistency regularization for adversarial robustness," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2022, vol. 36, no. 8, pp. 8414–8422.

[33] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd. Int. Conf. Learn. Representations*, May 2015, pp. 1–11.

[34] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. 5th. Int. Conf. Learn. Representations*, Apr. 2017, pp. 1–17.

[35] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. 6th. Int. Conf. Learn. Representations*, May 2018, pp. 1–13.

[36] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2574–2582.

[37] W. Nie, B. Guo, Y. Huang, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2022, pp. 16805–16827.

[38] Q. Wu, H. Ye, and Y. Gu, "Guided diffusion model for adversarial purification from random noise," 2022, *arXiv:2206.10875*.

[39] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Déforges, "Adversarial example detection for DNN models: A review and experimental comparison," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4403–4462, 2022.

[40] Z. Zhang, Q. Liu, C. Wu, S. Zhou, and Z. Yan, "A novel adversarial detection method for UAV vision systems via attribution maps," *Drones-Basel*, vol. 7, no. 12, 2023, Art. no. 697.

[41] T. Hickling, N. Aouf, and P. Spencer, "Robust adversarial attacks detection based on explainable deep reinforcement learning for UAV guidance and planning," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 456–467, Oct. 2023.

[42] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi, "Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 14973–14982.

[43] H. Jiang, J. Lin, and H. Kang, "FGMD: A robust detector against adversarial attacks in the IoT network," *Futur. Gener. Comp. Syst.*, vol. 132, pp. 194–210, 2022.

[44] W. Kong, H. Cao, J. Tian, and X. Kuang, "Detecting adversarial samples in neural network with statistical metrics: A practical approach," in *Proc. 2021 IEEE Int. Conf. Data Sci. Cyberspace.*, Jun. 2021, pp. 662–669.

[45] A. Cennamo, I. Freeman, and A. Kummert, "A statistical defense approach for detecting adversarial examples," in *Proc. Int. Conf. Pattern Recognit. Intell. Syst.*, Jan. 2020, pp. 1–7.

[46] R. D. Hjelm et al., "Learning deep representations by MI estimation and maximization," in *Proc. 7th. Int. Conf. Learn. Representations*, May 2019, pp. 1–24.

[47] M. I.. Belghazi, A. Baratin, A. Rajeshwar, S. Ozair, and S. Bengio, "Mutual information neural estimation," in *Proc. 35th Int. Conf. Mach. Learn.*, May 2018, pp. 531–540.

[48] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. 2017 IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[49] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.

[50] I. Csiszár, "The method of types [information theory]," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[51] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time. iv," *Commun. Pure Appl. Math.*, vol. 36, no. 2, pp. 183–212, 1983.

[52] S. Nowozin, B. Cseke, and R. Tomioka, "F-GAN: Training generative neural samplers using variational divergence minimization," *Adv. Neural Inf. Process. Syst.*, vol. 29, pp. 2202–2210, 2016.

[53] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[54] B. Zhao, Y. Zhong, G. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model fusing heterogeneous features for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[55] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2010, pp. 270–279.

Linjuan Li received the B.S. degree in control theory and control engineering from the Taiyuan University of Technology, Taiyuan, China, in 2009, and the Ph.D. degree in control science and engineering from the Taiyuan University of Science and Technology, Taiyuan, in 2024.

She is currently with the School of Electronic Information Engineering, Taiyuan University of Science and Technology. Her research interests include computer vision and deep learning technology.

Gang Xie (Member, IEEE) received the B.S. degree in control theory and the Ph.D. degree in circuits and systems from the Taiyuan University of Technology, Taiyuan, China, in 1994 and 2006, respectively.

He has been the Vice President with the Taiyuan University of Science and Technology, Taiyuan, and a Professor with the Taiyuan University of Technology, since 2008. His research interests include intelligent information processing, complex networks, and Big Data.

Haoxue Zhang received the B.S. degree in measurement and control technology and instrument from the Beijing University of Chemical Technology, Beijing, China, in 2017. She is currently working toward the Ph.D. degree in control science and engineering with the School of Electronic and Information Engineerin, Taiyuan University of Science and Technology, Taiyuan, China.

Her research interests include scene understanding, semantic segmentation, and high-resolution remote sensing image processing.

Xinlin Xie received the B.S. degree in electronics and information engineering and the Ph.D. degree in electronic science and technology from the Taiyuan University of Technology, Taiyuan, China, in 2012 and 2019, respectively.

He is currently with the School of Electronic Information Engineering, Taiyuan University of Science and Technology. His research interests include deep learning technology and computer vision.

Heng Li received the B.S. and M.S. degrees in civil engineering from Tongji University, in 1984 and 1987, respectively, and the Ph.D. degree in architectural science from The University of Sydney, Australia, in 1993.

He is a Chair Professor of construction informatics with The Hong Kong Polytechnic University, Hong Kong. He started his academic career from Tongji University, Shanghai, China, since 1987. He then researched and lectured with the University of Sydney, Sydney, NSW, Australia, James Cook University, Douglas, QLD, Australia, and Monash University, Melbourne, VIC, Australia, before joining The Hong Kong Polytechnic University. During this period, he have also worked with engineering design and construction firms and provided consultancy services to both private and government organizations in Australia, Hong Kong, and China. He has conducted many funded research projects related to the innovative application and transfer of construction information technologies, and has authored or coauthored 2 books and more than 300 journal papers in major journals of his field and numerous conferences papers in proceedings.