Contextual Adversarial Consistency Distillation for Remote Sensing

Yanze Gao[1], Changxin Rong[1], Lyuzhou Chen[1], Xiangyu Wang[1*], Qiuju Chen[1] and Huanhuan Chen[1*]
[1]School of Computer Science and Technology, University of Science and Technology of China
*{sa312, hchen}@ustc.edu.cn

***Abstract*—Knowledge distillation has become a pivotal strategy for deploying efficient remote sensing classifiers on resource-constrained platforms. However, student models in existing methods primarily focus on mimicking the observational distribution of high-performance teacher models, inheriting the contextual biases embedded in them. These biases induce spurious correlations where models rely on statistical shortcuts from non-causal backgrounds instead of intrinsic object features, leading to poor discrimination and generalization. To overcome this limitation, a Contextual Adversarial Consistency Distillation (CACD) framework is proposed to actively intervene in the data generation process. Specifically, the framework first computes a soft inter-class affinity measure to uncover potential contextual biases of teacher models. Guided by this prior knowledge, we then employ a latent diffusion model to synthesize adversarial backgrounds, generating interventional samples where the foreground is preserved but the background is replaced with confusing textures. Finally, we transfer this causal invariance to student models through a dual-level consistency objective, which reinforces the stability of both probabilistic outputs and feature representations under different background interventions. Extensive experiments on multiple benchmark datasets demonstrate that our framework significantly outperforms state-of-the-art baselines, effectively severing spurious correlations and enhancing both classification accuracy and robustness against contextual variations.**

***Index Terms*—Causal Intervention, Remote Sensing, Knowledge Distillation**

I. Introduction

Remote sensing image scene classification is a fundamental task in geospatial analysis, providing critical support for a wide range of applications, including precision agriculture, urban planning, and real-time disaster response [1–3]. Benefiting from recent advances in deep learning, neural networks have enabled the interpretation of complex geospatial data with high classification accuracy by extracting deep features, even for visually similar classes [4]. However, deploying these complex models in real-world applications remains challenging due to the strict computation and energy constraints of edge platforms such as satellites and drones [5]. To achieve efficient extraction of geospatial information on resource-constrained hardware, Knowledge Distillation (KD) has become a widely adopted strategy [6]. By transferring knowledge from a cumbersome teacher to a compact student network, KD allows lightweight models to achieve competitive performance, making geospatial analysis feasible in resource-constrained environments [7].

This research was supported in part by the National Nature Science Foundation of China (No. 62406302, 62137002, 62576327), in part by the Natural Science Foundation of Anhui province (No. 2408085QF195). This work was funded by the USTC Research Funds of the Double First-Class Initiative (Grant No. YD9110002085), Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM (Grant No. 2025IHM01030).
* Corresponding authors

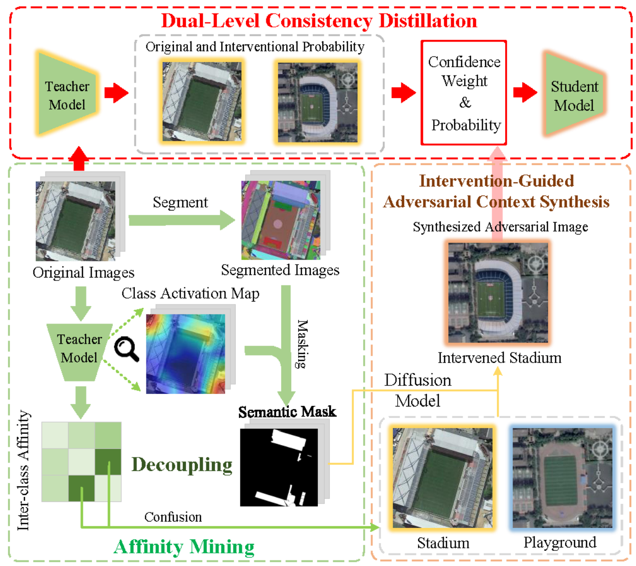

Fig. 1: The overall architecture of the proposed method.

In practical remote sensing scenarios, objects are often embedded in complex and variable geographical contexts, which poses a higher demand on the ability to classify target objects [8]. Existing KD methods mainly force lightweight student networks to mimic the discriminative behaviors of high-performance teacher models under the observational distribution [9]. However, they ignore the predictive biases of teacher models in such correlated contexts. As a result, teachers may exploit spurious correlations as statistical shortcuts, linking non-causal contextual information to class labels instead of learning intrinsic causal features[1]. When a student model is trained to fit the probability distribution of the teacher model, it inevitably inherits contextual biases and becomes overly dependent on background textures for recognition. This reliance renders student models extremely vulnerable when facing out-of-distribution scenarios, severely limiting generalization.

[1]Consider a model that classifies a *Stadium* not by recognizing its architectural structure, but merely by associating green fields with sports. This leads to confusion when another class is *Playground*, which also has green fields.

To overcome this limitation of contextual bias inherent in traditional KD methods, this paper introduces the perspective of causal intervention, aiming to fundamentally break the interference of environmental contexts on class predictions [10]. This mechanism involves performing the do-operator to actively control the state of background variables, severing the shortcut paths between contexts and class labels [11]. By leveraging this capability, the model is forced to extract robust object features under contextual variations, enabling it to break the reliance on contexts and capture intrinsic causal features.

However, applying the causal intervention faces challenges. First, discriminative objects are often deeply intertwined with diverse backgrounds, and the biases of teacher models are exposed when confronted with highly similar confusing backgrounds. Therefore, simply extracting backgrounds and randomly replacing them between class pairs will damage intrinsic features, struggling to reach the true decision boundaries due to a lack of targeted guidance. Second, interventional samples in remote sensing need to alter the background texture while maintaining geospatial consistency. Therefore, random noise padding or simple cross-image stitching often disrupts geographical consistency, introducing irrelevant artifacts and failing to generate confusing backgrounds. Finally, the differences between the original and interventional samples need to be reflected in the distillation process. This process requires that changes in the background do not interfere with either the internal feature extraction or the final prediction.

To address these challenges, this paper proposes a Contextual Adversarial Consistency Distillation (CACD) framework that preserves the intrinsic semantics of classes and prevents interference from contextual variations, as shown in Fig. 1. Instead of passively accepting the data distribution, the method intervenes data generation by introducing Structural Causal Models (SCMs), where the objective is to simulate the interventional operation $\mathrm{do}(\cdot)$ on background variables, effectively severing the backdoor path from backgrounds to labels. Specifically, a soft inter-class affinity measure is first designed to identify the most confusing contextual biases, and extract the activation map from the teacher model to divide images into segments. Second, guided by this prior, a latent diffusion model is utilized to synthesize adversarial contexts by replacing the original background with confusing textures while preserving the foreground object. Finally, a dual-level consistency constraint transfers this causal invariance to the student model by enforcing stability in both output and feature spaces. The contributions of this paper are listed as follows:

- A soft inter-class affinity measure is designed that quantifies feature overlaps in the teacher model, identifying confusing contextual biases for targeted interventions.
- An adversarial synthesis strategy based on SCM is introduced to replace confusing backgrounds and generate intervention samples with the diffusion model, decoupling the correlation between the foreground and background.
- A dual-level consistency constraint is designed that combines entropy-aware gating and feature regularization to force the student model to learn invariant representations.

II. Related Work

In this section, we review related work on remote sensing knowledge distillation and causal inference, highlighting the need for causal intervention to address spurious correlations.

A. Knowledge Distillation in Remote Sensing

Knowledge Distillation (KD), originally developed to transfer knowledge from a cumbersome teacher model to a more compact student model, has become a key technique for deploying deep learning models on resource-constrained remote sensing platforms [12]. Early KD works in this field primarily focus on logit-based distillation, where the student network is trained to mimic the class probability distributions of a teacher model to capture its high-level decision logic [13]. Subsequent research has extended this to feature-based distillation, aiming to align intermediate feature maps to preserve rich spatial and textural information [14, 15]. Recent advancements have also explored cross-modal distillation to leverage multi-source data to improve student performance [16–18].

However, a critical limitation of these methods is their exclusive reliance on observational distributions [19]. In complex remote sensing scenes, teacher models often inadvertently develop contextual biases, establishing spurious correlations between non-causal background elements and class labels. Since existing KD methods lack a mechanism to distinguish between causal object features and environmental noise, the student model is forced to blindly inherit these statistical shortcuts, which severely limits the robustness and generalization in out-of-distribution scenarios. Unlike these methods, our work introduces the causal intervention to decouple these spurious correlations, ensuring that the distilled knowledge is rooted in intrinsic semantic attributes rather than contextual variations.

B. Causal Inference and Intervention in Vision

To mitigate reliance on non-robust features, causal inference has been increasingly integrated into computer vision tasks [20]. The core paradigm involves modeling the data generation process through Structural Causal Models (SCMs) and eliminating confounding factors via the do-operator [11]. In the task of visual classification, several studies have proposed decoupling spurious correlations between objects and their backgrounds by generating counterfactual samples or reweighting training data based on estimated causal effects [21]. These methods aim to force the model to prioritize causal foreground attributes over environmental contexts [22].

However, research applying causal inference within the specific domain of remote sensing knowledge distillation remains insufficient. Furthermore, existing intervention strategies primarily rely on random background replacement or simple masking, which are inefficient because they fail to target the specific decision boundaries where the model is most confused. In contrast, our method introduces an adversarial consistency mechanism, which actively synthesizes the most challenging contextual examples based on inter-class affinity, forcing student models to learn more robust and invariant representations that existing intervention methods fail to achieve.

III. METHOD

This section provides a detailed description of the proposed framework, consisting of three stages: confusion identification through affinity mining, intervention-guided with adversarial context synthesis, and dual-level consistency distillation.

A. Affinity Mining and Decoupling

The prerequisite for effective contextual consistency distillation is to identify which environmental contexts are most likely to induce spurious correlations in the teacher model, and subsequently isolate these contextual regions from the semantic foreground. Since teacher biases are not randomly distributed but are class-specific, interventions must target the most confusing class pairs to be efficient. Therefore, a soft, probability-based inter-class affinity mining mechanism is proposed to uncover potential contextual biases and capture fine-grained dependencies beyond rigid decision boundaries.

Specifically, we first quantify dependencies between classes by constructing an inter-class affinity matrix, denoted as $\mathcal{A}$. Let $P_T(x)$ represent the probability output of the teacher model for an input image x belonging to the ground truth class c_i. To quantify the intrinsic feature overlap between classes, we aggregate the teacher's soft confidence scores across the entire training dataset $\mathcal{D}$. The affinity score $\mathcal{A}_{ij}$ between the source class c_i and the potentially interfering class c_j is calculated as the expected probability mass assigned to c_j:

$$\mathcal{A}_{ij} = \mathbb{E}_{x \sim \mathcal{D}_{c_i}} \left[P_T(x)_j \right], \tag{1}$$

where $P_T(x)_j$ denotes the predicted probability for class c_j. High values of $\mathcal{A}_{ij}$ where $i \neq j$ indicate that the teacher model perceives a strong feature overlap between c_i and c_j, which is driven by shared or confusing background textures, e.g., the visual similarity in the soil color tones between a *BareLand* and a *Desert*. Based on this matrix, for each target class c_i, we can retrieve the class $c_{\text{adv}} = \arg\max_{i \neq j} \mathcal{A}_{ij}$ as the target for adversarial context synthesis. This ensures that subsequent interventions target the most vulnerable decision boundaries in the feature space of the teacher model.

To physically implement these interventions, we need to structurally decouple semantic objects from their backgrounds. We employ a dual-stream localization strategy that harmonizes bottom-up segmentation with top-down attention. Specifically, we utilize the Segment Anything Model (SAM) to generate a class-agnostic pool of candidate masks $\mathcal{S} = \{m_1, m_2, ..., m_k\}$. Simultaneously, we extract Class Activation Map (CAM) from the teacher model corresponding to ground truth labels, which highlight the discriminative regions used for prediction. The optimal semantic mask M_{obj} is determined by selecting the candidate segment m_k that maximizes the Intersection-over-Union (IoU) with the binarized CAM threshold at a significant activation level. Therefore, the contextual region is explicitly defined as the complement of the object region, formulated as $M_{\text{bg}} = 1 - M_{\text{obj}}$. This structural decoupling allows us to freeze the semantic foreground while exposing the background pixels M_{bg} for the subsequent generative intervention.

B. Intervention-Guided Adversarial Context Synthesis

To eliminate the reliance on spurious background correlations, we formulate the image generation process from the perspective of an SCM. Assume that a remote sensing image x consists of semantic objects O and backgrounds B. In the observational distribution P_{obs}, the teacher model tends to learn a predictor $(O, B) \to Y$ that exploits the confounding correlation between O and B with the class label Y. For instance, the *Airplane* frequently co-occurs with "Runway". This creates a statistical shortcut $B \to Y$ that compromises the robustness of the model. Our objective is to transition from the observational probability $P(Y|O, B)$ to the interventional probability $P(Y|O, do(B))$, effectively severing the backdoor path from the background to the label.

This intervention can be achieved by forcing the background B to transition from its natural state to an adversarial state determined by the confounding prior c_{adv} identified in Section III-A. This constitutes a counter-intuitive intervention, denoted as $\text{do}(B := G(c_{\text{adv}}))$, where $G(\cdot)$ represents the texture synthesis function. Specifically, we construct a semantic prompt π_{adv} that explicitly describes the textural characteristics of the adversarial class c_{adv}. To ensure high semantic fidelity of the synthesized context, the prompt can be described as "Aerial view of the $[\text{Desc}(c_{\text{adv}})]$ texture". This prompt acts as the conditional guidance for the generative intervention, steering the synthesis process towards decision boundary regions where the teacher model exhibits maximum uncertainty.

To technically implement this intervention while strictly preserving the causal integrity of objects, we perform a masked latent fusion process. Specifically, we utilize a pre-trained Latent Diffusion Model (LDM) that operates in a compressed feature space. Let z_0 denote the latent representation of the original image encoded by the LDM encoder, and z_{adv} denote the latent features generated based on the adversarial prompt π_{adv}. Since the binary background mask M_{bg} is derived in the pixel space, we first spatially downsample it to match the resolution of the latent feature map, denoted as M'_{bg}. Then, we enforce the preservation of the causal foreground by fusing the latent features, which can be formulated as:

$$z_{\text{final}} = z_0 \odot (1 - M'_{\text{bg}}) + z_{\text{adv}} \odot M'_{\text{bg}}, \tag{2}$$

where $\odot$ denotes the Hadamard product of elements. This operation ensures that the semantic foreground is frozen by directly copying its original latent values, while the background region is entirely replaced by the synthesized adversarial context. Finally, the adversarial sample is reconstructed through the decoder, that is, $x_{\text{adv}} = \text{Dec}(z_{\text{final}})$. The entire process described above guarantees that the student model is subjected to controlled intervention, in which the background is modified according to the previous confusion, without altering the inherent object features. However, the key to robust knowledge transfer lies in how the student model is enforced to remain invariant across these original and interventional scenarios. This requires injecting true causal invariance into the compact student network to bridge the gap between these two distributions, which is discussed in the following section.

C. Dual-Level Consistency Distillation

The final stage of the proposed framework aims to transfer robust causal knowledge to the student model S, where a reliable model should maintain consistent predictions and internal feature representations regardless of contextual changes. By enforcing stability across both probabilistic outputs and high-level features, we can force the student model to ignore the adversarial background x_{adv} and focus exclusively on the invariant foreground semantics. To achieve this, we introduce a dual-level consistency distillation objective: output consistency distillation and feature consistency regularization.

At the probabilistic level, we require the student model to produce stable predictions even when facing adversarial backgrounds. We use the prediction of the teacher model on the original image x as a reliable supervisory target, training the student model to minimize the divergence between this target distribution $P_T(x)$ and its prediction $P_S(x_{\text{adv}})$ on the synthesized adversarial sample. To mitigate the risk of transferring incorrect knowledge from uncertain teacher predictions, we calculate the confidence weight ω using an exponential decay function based on the prediction entropy $H(P_T(x))$:

$$\omega = \exp\left(-\frac{H(P_T(x))}{\tau}\right), \tag{3}$$

where τ is the temperature parameter. This non-linear weighting scheme significantly reduces the importance of high-entropy samples, ensuring that the distillation focuses on high-confidence regions[2]. The output consistency loss is defined as:

$$\mathcal{L}_{\text{out}} = \omega \cdot D_{\text{KL}}(P_T(x) \| P_S(x_{\text{adv}})), \tag{4}$$

where D_{KL} represents the Kullback-Leibler (KL) divergence.

To ensure that the model implicitly decouples the foreground from the background, a feature consistency regularization is imposed, which means that the high-level semantic features of the original image x and the adversarial image x_{adv} should remain consistent because the causal foreground object is preserved. Let $f_S(\cdot)$ denote the feature vector of the student model's penultimate layer. We minimize the cosine distance between the features of paired images:

$$\mathcal{L}_{\text{feat}} = 1 - \frac{f_S(x) \cdot f_S(x_{\text{adv}})}{\|f_S(x)\| \|f_S(x_{\text{adv}})\|}. \tag{5}$$

This regularization explicitly penalizes feature shifts caused by background variations, focusing on invariant foreground semantics. The total loss function combines the standard cross-entropy loss $\mathcal{L}_{\text{ce}}$ with these two consistency terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}}(S(x), y) + \lambda_1 \mathcal{L}_{\text{out}} + \lambda_2 \mathcal{L}_{\text{feat}}. \tag{6}$$

Through this dual-layer supervision, the student model effectively learns to decouple spurious contextual correlations from the true label, achieving better generalization performance.

[2]High-entropy regions occur when the teacher is uncertain about complex backgrounds, while low-entropy regions correspond to samples where the teacher is highly confident and produces a sharp probability distribution.

TABLE I: Comparative accuracy of the proposed framework with existing baselines on the AID and NWPU datasets.

Type	Method	AID	NWPU
CNN-based	GoogLeNet [23]	83.44 ± 0.40	86.02 ± 0.18
	VGG-VD-16 [24]	86.59 ± 0.29	90.36 ± 0.18
	ResNet50 [25]	92.57 ± 0.17	91.86 ± 0.19
	SCCov [26]	93.12 ± 0.25	92.10 ± 0.25
	BiMobileNet [27]	94.83 ± 0.24	94.08 ± 0.11
	MF²CNet [28]	95.54 ± 0.17	93.85 ± 0.27
ViT-based	Fine-tune VIT [29]	94.90 ± 0.29	93.90 ± 0.20
	ET-GSNet [30]	95.58 ± 0.18	94.50 ± 0.18
KD-based	DKD [31]	94.78 ± 0.24	93.23 ± 0.26
	SSKDNet [6]	95.96 ± 0.12	94.92 ± 0.12
Ours	CACD	**97.02 ± 0.16**	**95.17 ± 0.10**

IV. Experiments

This section first sets up the experiments and then presents performance comparisons with state-of-the-art baselines. In addition, ablation studies and further analyses are conducted, including the sensitivity and qualitative analysis, demonstrating the effectiveness of the proposed framework.

A. Experimental Setup

Datasets. The proposed framework is evaluated on two remote sensing scene classification datasets, including AID [24] and NWPU. For both datasets, we follow standard protocols [1], randomly selecting a subset for training and reserving the remainder for testing to ensure a fair comparison. Specifically, the AID dataset contains 10,000 images of 30 different aerial scene types, while the NWPU dataset spans 31,500 images covering 45 classes with significant background variations.

Baselines. We benchmark the framework against comprehensive state-of-the-art methods, divided into three groups: 1) Standard CNN architectures, including GoogLeNet [23], VGG-VD-16 [24], ResNet50 [25], SCCov [26], BiMobileNet [27], and MF²CNet [28]; 2) Visual transformation methods, including fine-tuning ViT [29] and ET-GSNet [30]; and 3) Advanced KD methods, such as DKD [31] and SSKDNet [6].

Implementation Details. For the distillation, we adopt a pre-trained ResNet-50 as the teacher network due to its robust and mature feature extraction capabilities in remote sensing tasks, which has 25.6 million parameters and requires 4.1 billion FLOPs. The lightweight EfficientNet-B0 [32] is employed as the student network, with only 5.3 million parameters and 0.39 billion FLOPs. For adversarial context synthesis, we utilize pre-trained stable diffusion as the generative backbone, where the sampling process is configured with 50 DDIM inference steps and a classifier-free guidance scale of 7.5 to balance diversity and fidelity. The positive prompts are dynamically constructed based on the mined inter-class affinity matrix, and the negative prompts are fixed to low quality and blurry to ensure visual clarity. All experiments are conducted on a system equipped with a 4.5 GHz AMD Ryzen 9 7950X CPU, an NVIDIA GeForce RTX 4090 GPU, and 32GB of memory.

TABLE II: Ablation results on the core components of the proposed framework on the AID and NWPU datasets.

Method Ablation	AID	NWPU
CACD (Ours)	**97.02 ± 0.16**	**95.17 ± 0.10**
w/o Causal intervention	94.86 ± 0.23	91.09 ± 0.18
w/o Affinity mining	95.32 ± 0.18	91.88 ± 0.20
w/o Adversarial context synthesis	95.54 ± 0.12	92.53 ± 0.14
w/o Confidence weight ω	95.12 ± 0.21	93.06 ± 0.17
w/o Feature consistency $\mathcal{L}_{\text{feat}}$	95.83 ± 0.10	94.81 ± 0.13

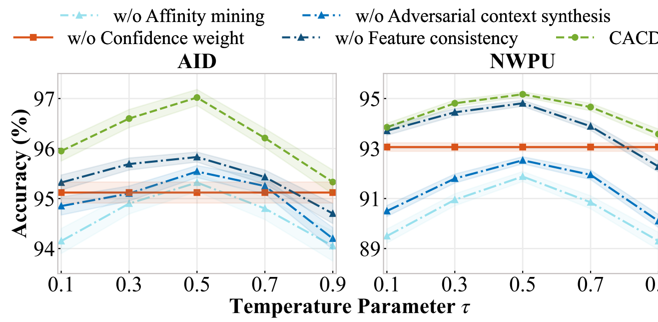

Fig. 2: Sensitivity analysis on the effect of temperature parameter τ in terms of accuracy on the AID and NWPU datasets.

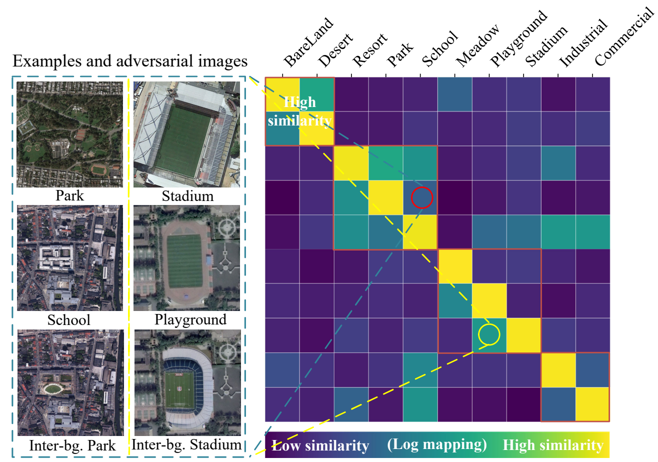

Fig. 3: Visualization of inter-class affinity matrix and synthesized adversarial images guided by confusion pairs on the AID.

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

To evaluate the effectiveness of the proposed framework, we conduct a comprehensive comparative analysis against various state-of-the-art baselines on the AID and NWPU datasets. As shown in Table I, our method consistently outperforms all baseline methods on both datasets. Notably, on the challenging NWPU dataset, which is characterized by high intra-class variance, our method achieves a significant accuracy gain of approximately 1.9% over traditional KD methods. This performance advantage suggests that simply mimicking the teacher's output is insufficient for robust knowledge transfer in complex remote sensing scenes. Traditional KD methods often encounter difficulties in this scenario, because student models tend to memorize specific background textures associated with the teacher's predictions. In contrast, the CACD framework effectively breaks this spurious correlation. By reinforcing consistency across adversarial background variations, our student model learns to focus on the intrinsic structural features of the foreground, achieving superior generalization capabilities even when the background context is misleading.

C. Ablation Study

To independently validate the contribution of each core component, we conduct a step-by-step ablation study, where we start from our full model and replace or remove only one component at a time. Specifically: 1) Remove the introduction of causal intervention. 2) Replace affinity mining with random class pairs selection. 3) Replace adversarial context synthesis with simple background masking. 4) Remove the confidence weight. 5) Remove the feature-level consistency regularization.

As demonstrated in Table II, the replacement or removal of any component degrades the performance of the model, validating their necessity. The most significant performance drop occurs when the causal intervention is entirely removed, suggesting that standard distillation forces the student to inherit the contextual biases of the teacher model, whereas intervention is essential to decouple these spurious correlations. In addition, our proposed affinity-based synthesis targets confusing contexts, providing more effective gradients for robust learning. Furthermore, removing both the confidence weight and feature consistency regularization leads to a performance drop, highlighting that the the former prevents the student model from mimicking the teacher's incorrect predictions on uncertain adversarial examples, while the latter focuses on intrinsic causal attributes to decouple semantic representations.

D. Parameter Sensitivity Analysis

To investigate how the temperature parameter τ in the entropy-aware reliability gate regulates the trade-off between knowledge quantity and quality, a sensitivity analysis is conducted on the AID and NWPU datasets by varying τ from 0.1 to 0.9 with an interval of 0.2. The experimental results, reported in Fig. 2, reveal a bell-shaped performance curve. Specifically, when τ is too small, the gate becomes restrictive, filtering out a significant portion of valid supervisory signals and leading to suboptimal convergence. Conversely, when τ exceeds 0.7, the gate fails to suppress high-entropy predictions, allowing the teacher's uncertainty on complex adversarial samples to mislead the student's learning process. Optimal performance is consistently observed around $\tau = 0.5$, indicating a balanced trade-off where the student learns from the confident predictions of the teacher while ignoring ambiguous ones. This stability across datasets suggests that our method is robust to hyperparameter settings within a reasonable range.

E. Visualization and Analysis of Contextual Affinity

To intuitively understand the contextual bias captured by our method, we visualize the learned inter-class affinity matrix in Fig. 3. Unlike standard confusion matrices that only compute hard errors, our affinity matrix reveals subtle semantic dependencies. For example, we observe a high affinity score between *Stadium* and *Playground*. Visually, these two classes share similar background textures, which often become a confounding factor for the teacher model. The model tends to use the statistical shortcut of "green grass" for classification, rather than extracting the essential structural features of the target. Guided by these discovered affinities, our generative module synthesizes "Playground-background" for stadium images. By successfully intervening in these specific contextual shortcuts, the CACD framework forces the model to abandon the background confusion and learn causal features.

V. Conclusion

This paper proposes the Contextual Adversarial Consistency Distillation (CACD) framework to address the spurious background correlation problem in remote sensing knowledge distillation. By combining inter-class affinity mining with adversarial context synthesis, our framework actively severs the shortcut between environmental contexts and class labels. The proposed dual-level consistency constraint further ensures that the student model learns invariant representations even under drastic changes in the background. Experiments confirm that the method significantly improves robustness and generalization compared to state-of-the-art baselines. Future work will extend this causal framework to cross-modal tasks.

References

[1] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[2] R. Ahmad, "Smart remote sensing network for disaster management: An overview," *Telecommunication Systems*, vol. 87, no. 1, pp. 213–237, 2024.

[3] J. Chen, X. Dai, Y. Guo, J. Zhu, X. Mei, M. Deng, and G. Sun, "Urban built environment assessment based on scene understanding of high-resolution remote sensing imagery," *Remote Sensing*, vol. 15, no. 5, p. 1436, 2023.

[4] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[5] R. Yang, Z. Chen, B. Wang, Y. Guo, and L. Hu, "A lightweight detection method for remote sensing images and its energy-efficient accelerator on edge devices," *Sensors*, vol. 23, no. 14, p. 6497, 2023.

[6] Y. Zhao, J. Liu, J. Yang, and Z. Wu, "Remote sensing image scene classification via self-supervised learning and knowledge distillation," *Remote Sensing*, vol. 14, no. 19, p. 4813, 2022.

[7] H. Song, C. Wei, and Z. Yong, "Efficient knowledge distillation for remote sensing image classification: A CNN-based approach," *International Journal of Web Information Systems*, vol. 20, no. 2, pp. 129–158, 2024.

[8] P. Ciampi, M. Mangifesta, L. M. Giannini, C. Esposito, G. Scalella, B. Burchini, and N. Sciarra, "Geophysical and remote sensing techniques for large-volume and complex landslide assessment," *Remote Sensing*, vol. 17, no. 12, p. 2029, 2025.

[9] H. Song, J. Xie, L. Liang, Y. Su, Y. Xiao, X. Zhang, Y. Ouyang, X. Li, S. Chen, and Y. Li, "Symmetrical learning and transferring: Efficient knowledge distillation for remote sensing image classification," *Symmetry*, vol. 17, no. 7, p. 1002, 2025.

[10] L. Chen, T. Ban, X. Wang, D. Lyu, and H. Chen, "Mitigating prior errors in causal structure learning: A resilient approach via bayesian networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[11] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: Foundations and learning algorithms*. The MIT press, 2017.

[12] H.-I. Liu, M. Galindo, H. Xie, L.-K. Wong, H.-H. Shuai, Y.-H. Li, and W.-H. Cheng, "Lightweight deep learning for resource-constrained environments: A survey," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–42, 2024.

[13] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit standardization in knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15731–15740, 2024.

[14] Z. Yang, Z. Li, A. Zeng, Z. Li, C. Yuan, and Y. Li, "ViTKD: Feature-based knowledge distillation for vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1379–1388, 2024.

[15] T. Ban, C. Rong, X. Wang, L. Chen, X. Wang, D. Lyu, Q. Zhu, and H. Chen, "Differentiable structure learning with ancestral constraints," in *Forty-second International Conference on Machine Learning*, 2025.

[16] D. Lyu, X. Wang, T. Ban, L. Chen, X. Zhou, and H. Chen, "Expanding the category of classifiers with llm supervision," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5905–5913, 2025.

[17] T. Ban, L. Chen, D. Lyu, X. Wang, Q. Zhu, and H. Chen, "LLM-driven causal discovery via harmonized prior," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[18] Z. Tao, H. Chen, Y. Sun, Q. Tu, F. Gao, L. Chen, W. Wang, and Y. Gao, "Active differentiable structure learning for clinical causal discovery," *Knowledge-Based Systems*, p. 114145, 2025.

[19] X. Wang, T. Ban, L. Chen, D. Lyu, Q. Zhu, and H. Chen, "Large-scale hierarchical causal discovery via weak prior knowledge," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[20] X. Wang, Y. Gao, C. Rong, L. Chen, D. Lyu, X. Zhou, T. Ban, and H. Chen, "Counterfactual-driven zero-shot classifier expansion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, pp. 26508–26516, 2026.

[21] T. Ban, C. Rong, X. Wang, L. Chen, Y. Gao, X. Wang, and H. Chen, "Structure learning from time-series data with lag-agnostic structural prior," in *The Fourteenth International Conference on Learning Representations*, 2025.

[22] L. Chen, Y. Sun, Y. Gao, X. Wang, D. Lyu, T. Ban, X. Wang, X. Zhou, and H. Chen, "Pattern-guided adaptive prior for structure learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[24] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[25] Z. Zhao, J. Li, Z. Luo, J. Li, and C. Chen, "Remote sensing image scene classification based on an enhanced attention module," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 1926–1930, 2020.

[26] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 1461–1474, 2019.

[27] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, "An efficient and lightweight convolutional neural network for remote sensing image scene classification," *Sensors*, vol. 20, no. 7, p. 1999, 2020.

[28] L. Bai, Q. Liu, C. Li, Z. Ye, M. Hui, and X. Jia, "Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[29] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[30] K. Xu, P. Deng, and H. Huang, "Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[31] D. Li, Y. Nan, and Y. Liu, "Remote sensing image scene classification model based on dual knowledge distillation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[32] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019.