Spatial-Aware Conformal Prediction for Trustworthy Hyperspectral Image Classification

Kangdao Liu, Tianhao Sun, Hao Zeng, Yongshan Zhang, *Member, IEEE*, Chi-Man Pun, *Senior Member, IEEE*, and Chi-Man Vong, *Senior Member, IEEE*

Abstract—**Hyperspectral image (HSI) classification involves assigning unique labels to each pixel to identify various land cover categories. While deep classifiers have achieved high predictive accuracy in this field, they lack the ability to rigorously quantify confidence in their predictions. Quantifying the certainty of model predictions is crucial for the safe usage of predictive models, and this limitation restricts their application in critical contexts where the cost of prediction errors is significant. To support the safe deployment of HSI classifiers, we first provide a theoretical proof establishing the validity of the emerging uncertainty quantification technique, conformal prediction, in the context of HSI classification. We then propose a conformal procedure that equips any trained HSI classifier with trustworthy prediction sets, ensuring that these sets include the true labels with a user-specified probability (e.g., 95%). Building on this foundation, we introduce Spatial-Aware Conformal Prediction (SACP), a conformal prediction framework specifically designed for HSI data. This method integrates essential spatial information inherent in HSIs by aggregating the non-conformity scores of pixels with high spatial correlation, which effectively enhances the efficiency of prediction sets. Both theoretical and empirical results validate the effectiveness of our proposed approach. The source code is available at https://github.com/J4ckLiu/SACP.**

Index Terms—**Hyperspectral image classification, uncertainty quantification, conformal prediction**

I. INTRODUCTION

Hyperspectral imaging technology captures spectral information across a wide wavelength range, providing more detailed pixel features compared to traditional three-channel color images [1], [2]. Hyperspectral image (HSI) classification, a fundamental task in this domain, involves assigning unique labels to each pixel to identify various land cover categories. While deep classifiers have demonstrated high accuracy in this field [3]–[7], performance metrics like accuracy alone are

insufficient for their deployment in safety-critical contexts. In these scenarios, it is also crucial to quantify the certainty of predictions. This need is amplified by the significant applications of HSIs, including surveillance, threat detection [8], mineral detection [9], agriculture [10], environmental protection [11], and defense [12], where prediction errors could have serious consequences. For example, in threat detection or mineral exploration, incorrect predictions could result in misidentifying potential threats or valuable resources, leading to severe repercussions. This emphasizes the necessity of advancing methodologies that rigorously quantify the uncertainty of HSI classifiers to ensure their safe deployment.

Various techniques have been developed to estimate and incorporate uncertainty into predictive models, including confidence calibration [13], Monte Carlo Dropout [14], and Bayesian neural networks [15]. Additionally, there are uncertainty quantification strategies specifically tailored to HSI classification. For instance, PL-CNN [16] reduces both predictive and label uncertainty for more reliable predictions, achieving high classification accuracy while lowering the Expected Calibration Error (ECE) [13]. Another approach leverages Bayesian deep learning to estimate prediction uncertainty while enhancing the structural characteristics of HSI classifiers [17]. Furthermore, uncertainty quantification techniques specifically applied to graph-based HSI classifiers have been explored in [18]. This study adapts advanced uncertainty quantification models, initially developed for node classification in graphs, to the context of HSI classification. Collectively, these techniques contribute valuable insights into quantifying or reducing decision uncertainty. However, they lack theoretical guarantees of model performance [19], underscoring the importance of conformal prediction [20]–[22].

Unlike the techniques discussed earlier, conformal prediction—an emerging uncertainty quantification methodology—typically operates as post-hoc processing for the outputs of trained classifiers. This statistical framework can be applied to any predictive model, transforming the output probabilities into prediction sets while ensuring that the true label is included with a user-specified coverage probability. Notably, the validity of the prediction sets is assured in a distribution-free manner, providing explicit, non-asymptotic guarantees without reliance on any distributional or model assumptions [20]. Formally, this guarantee can be expressed as follows:

$$\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) \geq 1 - \alpha, \tag{1}$$

where Y_{test} is the true label, $\mathcal{C}(X_{\text{test}})$ is the generated prediction set, and α is the user-specified error rate. The prediction set

will consistently include the true labels of the inputs with the specified error frequency. The strong theoretical foundation of conformal prediction supports its efficacy across various applications, including classification [19], [23], regression [24], and specialized areas such as image generative models [25], robotic control [26], graph neural networks [27], [28], and large language models (LLMs) [29]–[31].

Despite its success in numerous domains, conformal prediction has remained unexplored in the context of HSI data. There are two primary challenges that need to be addressed for the valid and effective implementation of conformal prediction in HSI classification. Firstly, unlike traditional classification tasks, HSI classifiers typically require access to the entire HSI, including test instances, during training. This practice could potentially undermine the validity of conformal prediction, which relies on the exchangeability of calibration and test data [20]. Therefore, it is imperative to theoretically examine whether the validity of conformal prediction is maintained in the context of HSI classification. Secondly, the performance of conformal prediction algorithms is dictated by the non-conformity score function once the base classifier is determined. Although many effective non-conformity score functions have been proposed—such as Adaptive Prediction Sets (APS) [32], which calculates the non-conformity scores by accumulating descending softmax values, and Regularized Adaptive Prediction Sets (RAPS) [33], which enhances the efficiency of conformal prediction sets through regularization based on label ranks—these non-conformity score functions generally neglect the critical spatial information inherent in HSIs. This oversight can lead to suboptimal prediction sets with unnecessarily large sizes. In summary, addressing the following two questions is essential for the effective application of conformal prediction in HSI classification:

- **Does the validity of conformal prediction persist in the context of HSI classification?**
- **How can spatial information be utilized to enhance the performance of conformal prediction?**

In this paper, we first present a theoretical analysis that confirms the validity of conformal prediction in HSI classification, even when classifiers are exposed to pixels from both calibration and test sets during training. Building on this foundation, we propose a framework for integrating conformal prediction sets into HSI classifiers, ensuring the inclusion of the true label with a user-specified coverage probability. Furthermore, we theoretically illustrate that incorporating neighborhood information can enhance the statistical efficiency of prediction sets. Motivated by this insight, we introduce **Spatial-Aware Conformal Prediction** (SACP), which refines standard conformal prediction by aggregating the non-conformity scores of pixels exhibiting high spatial correlation. This approach effectively leverages the spatial information inherent in HSIs to improve the performance of conformal prediction. Moreover, we present a theoretical demonstration of the validity of SACP, confirming that it guarantees coverage as defined in Eq. (1).

We conduct extensive evaluations on standard HSI classification benchmarks, including Indian Pines, Pavia University, and Salinas. The experimental results confirm the validity

of our proposed conformal framework (standard conformal prediction) and SACP for HSI classification: the generated prediction sets consistently include the true label of data with a user-specified coverage probability. Furthermore, SACP significantly enhances the efficiency of various non-conformity score functions while maintaining satisfactory coverage rates across different scenarios, outperforming standard conformal prediction in this context. This finding underscores the critical importance of incorporating spatial information into conformal prediction. Notably, SACP demonstrates its effectiveness with both simple classifiers, such as CNN-based models, and advanced classifiers, including Transformer-based models, thereby highlighting the adaptability of our proposed approach.

Our main contributions are highlighted as follows:

- To the best of our knowledge, we are the first to explore the application of the emerging uncertainty quantification technique, conformal prediction, to HSI data. We present a thorough theoretical proof that establishes the validity of conformal prediction for HSI classification, thereby opening a new avenue for the safe utilization of HSI.
- We introduce a framework that equips HSI classifiers with reliable conformal prediction sets, ensuring the inclusion of true labels with a user-specified probability.
- We provide a theoretical demonstration that incorporating spatial information can enhance the statistical efficiency of prediction sets. Building on this insight, we propose SACP, a framework that leverages the inherent spatial information in HSIs to improve the performance of conformal prediction. We theoretically demonstrate the validity of SACP by confirming its coverage guarantee.
- Extensive experiments validate the effectiveness of our proposed method. The empirical results confirm the validity of conformal prediction for HSI classification. Moreover, SACP outperforms standard conformal prediction by reducing the average prediction set size while maintaining satisfactory coverage rates across different scenarios.

II. BACKGROUND

A. Conformal Prediction

a) Notation: In this section, we consider the multi-class classification task involving K classes. Let $\mathcal{X} \subset \mathbb{R}^d$ represent the input space, and $\mathcal{Y} = \{1, \ldots, K\}$ denote the label space. Let (X, Y) be a random data pair drawn from the joint distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$. We use $f_\theta : \mathcal{X} \to \mathbb{R}^K$ to denote a pretrained classifier with parameters θ. For the input $\boldsymbol{x}$, $f_{\theta,y}(\boldsymbol{x})$ represents the y-th element of the logits vector $f_\theta(\boldsymbol{x})$. The conditional probability of class y is approximated by:

$$\hat{\pi}_{\theta,y}(\boldsymbol{x}) = \sigma(f_\theta(\boldsymbol{x}))_y = \frac{e^{f_{\theta,y}(\boldsymbol{x})}}{\sum_{i=1}^{K} e^{f_{\theta,i}(\boldsymbol{x})}},$$

where σ denotes the softmax function and $\hat{\pi}_\theta(\boldsymbol{x}) = (\hat{\pi}_{\theta,1}(\boldsymbol{x}), \hat{\pi}_{\theta,2}(\boldsymbol{x}), \ldots, \hat{\pi}_{\theta,K}(\boldsymbol{x}))$ represents the softmax probability distribution. Typically, the classification result is obtained by $\hat{y} = \mathrm{argmax}_{y \in \mathcal{Y}} \hat{\pi}_{\theta,y}(\boldsymbol{x})$.

b) Pipeline: Instead of predicting a single label based on the output probabilities of a classifier, conformal prediction constructs a set-valued mapping $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ that satisfies the following marginal coverage property:

$$\mathbb{P}(Y \in \mathcal{C}(X)) \geq 1 - \alpha,$$

where $\alpha \in (0, 1)$ denotes the user-specified error rate and $\mathcal{C}(X)$ is the prediction set, which is a subset of $\mathcal{Y}$.

Before deployment, conformal prediction starts with a calibration step using a separate calibration set, $\mathcal{D}_{\text{cal}} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, drawn independently from the distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$. Specifically, we compute the non-conformity score $s_i = S(\boldsymbol{x}_i, y_i)$ for each sample in $\mathcal{D}_{\text{cal}}$, which measures how y_i "conforms" to the prediction at $\boldsymbol{x}_i$. The $1 - \alpha$ quantile of the scores $\{s_i\}_{i=1}^n$ is then determined as the threshold τ. Formally, τ is obtained as follows:

$$\tau = \inf \left\{ s \mid \frac{|\{i : s_i \leq s\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\}. \quad (2)$$

Once the threshold τ is determined, the prediction set for the test instance $\boldsymbol{x}_{n+1}$ at the error rate α is generated by:

$$\mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}; \tau) := \{y \in \mathcal{Y} \mid S(\boldsymbol{x}_{n+1}, y) \leq \tau\}, \quad (3)$$

where $\mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}; \tau)$ includes all labels for which their non-conformity score $S(\boldsymbol{x}_{n+1}, y)$ does not exceed the threshold τ. As stated in the following theorem, the prediction sets generated by Eqs. (2) and (3) are guaranteed to achieve finite-sample marginal coverage of the true labels.

Theorem 1. *[20] Let $(\boldsymbol{x}_i, y_i)_{i=1}^n$ and $(\boldsymbol{x}_{n+1}, y_{n+1})$ be exchangeable data samples. Suppose τ is calculated by Eq. (2) using $(\boldsymbol{x}_i, y_i)_{i=1}^n$, and the prediction set for $(\boldsymbol{x}_{n+1}, y_{n+1})$ is given by Eq. (3). Then, the coverage guarantee is as follows:*

$$\mathbb{P}(y_{n+1} \in \mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}; \tau)) \geq 1 - \alpha.$$

Notably, the validity is ensured in a distribution-free manner, offering explicit, non-asymptotic guarantees without relying on distributional or model assumptions. Consequently, as long as the exchangeability assumption holds, the generated prediction sets will have a coverage guarantee. This highlights the flexibility of conformal prediction, as it can be applied to any **pre-trained classifier** when the assumption is satisfied.

c) Non-conformity score function: Once the base classifier is determined, the effectiveness of the resulting prediction set depends critically on the chosen non-conformity score function S. Two popular options are Adaptive Prediction Sets (APS) and Regularized Adaptive Prediction Sets (RAPS). APS calculates the non-conformity scores by accumulating the softmax probabilities, sorted in descending order:

$$S_{\text{APS}}(\boldsymbol{x}, y) = \sum_{y_i \neq y} \hat{\pi}_{\theta, y_i}(\boldsymbol{x}) \cdot \mathbb{1}_{\{\hat{\pi}_{\theta, y_i}(\boldsymbol{x}) > \hat{\pi}_{\theta, y}(\boldsymbol{x})\}} + u \cdot \hat{\pi}_{\theta, y}(\boldsymbol{x}),$$

where u is an independent random variable following a uniform distribution on $[0, 1]$. However, the softmax probabilities typically exhibit a long-tailed distribution, which facilitates the inclusion of tail classes in the prediction sets. To address this issue, RAPS [33] refines the non-conformity scores by incorporating regularization for the label rankings. Through this regularization, RAPS ensures that labels with higher rankings are associated with larger scores, defined as follows:

$$S_{\text{RAPS}}(\boldsymbol{x}, y) = S_{\text{APS}}(\boldsymbol{x}, y) + \lambda \cdot (o(y, \hat{\pi}_\theta(\boldsymbol{x})) - k)^+,$$

where $o(y, \hat{\pi}_\theta(\boldsymbol{x}))$ is the label ranking of y, λ and k are hyperparameters, and $(z)^+$ denotes the positive part of z. Yet, the non-conformity score of RAPS still involves unreliable softmax probabilities, leading to suboptimal performance in conformal prediction. To solve it, Sorted Adaptive Prediction Sets (SAPS) [19] reduces the emphasis on probabilities by retaining only the maximum softmax probability and discarding the others. Formally, SAPS is defined as follows:

$$S_{\text{SAPS}}(\boldsymbol{x}, y) = \begin{cases} u \cdot \hat{\pi}_{\theta, \max}(\boldsymbol{x}), & \text{if } o(y, \hat{\pi}_\theta(\boldsymbol{x})) = 1, \\ \hat{\pi}_{\theta, \max}(\boldsymbol{x}) + (o(y, \hat{\pi}_\theta(\boldsymbol{x})) - 2 + u) \cdot \lambda, \text{else}, \end{cases}$$

where λ balances the weight of ranking information, $\hat{\pi}_{\theta, \max}(\boldsymbol{x})$ denotes the maximum softmax probability of instance $\boldsymbol{x}$, and u is a random variable uniformly distributed on $[0, 1]$. **It can be concluded that these score functions depend exclusively on softmax probabilities and label ranks.**

d) Evaluation Metric: We evaluate the generated prediction sets using three key metrics: marginal coverage rate (Coverage), average size of the prediction sets (Size), and Size-Stratified Coverage Violation (SSCV). Coverage measures the percentage of samples whose sets contain true labels:

$$\text{Coverage} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_{\text{test}}} \mathbb{1}_{\{y \in \mathcal{C}(\boldsymbol{x})\}},$$

which should ideally match the pre-defined coverage rate $1 - \alpha$. Size measures the average length of prediction sets. When coverage is satisfied, a smaller size indicates higher statistical efficiency and is more desirable. Formally, Size is defined as:

$$\text{Size} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_{\text{test}}} |\mathcal{C}(\boldsymbol{x})|.$$

SSCV [33] reflects the conditional coverage rate and is formally defined as:

$$\text{SSCV} = 100 \times \sup_j \left| (1 - \alpha) - \frac{|\{i : y_i \in \mathcal{C}(\boldsymbol{x}_i) \text{ and } i \in \mathcal{J}_j\}|}{|\mathcal{J}_j|} \right|,$$

where $\mathcal{J}_j$ denotes the partitioned sets, with prediction sets categorized according to their sizes. This metric quantifies the maximum deviation between the observed coverage rate and $1 - \alpha$ across the various size categories.

B. HSI Classification

Let $\mathcal{I} \in \mathbb{R}^{H \times W \times U}$ be the HSI to classify, where H and W denote the height and width in pixels, and U denotes the number of spectral bands. Let $M = HW$ be the total number of pixels. Typically, HSI classification aims to predict the label $y \in \mathcal{Y}$ for each pixel in the image, given that a random subset of labels known. The pixels with unknown labels appear randomly at any location in the HSI. When employing a deep classifier, the (i, j)-th pixel is generally considered as the center, with information from the small

cuboid $B \in \mathbb{R}^{P \times P \times U}$ utilized to extract the joint spectral-spatial feature representation for the central pixel [34][1]. Here, P denotes the patch size of B. Correspondingly, the dataset constructed from $\mathcal{I}$ comprises M cuboids, defined as follows:

$$\mathcal{D}_\mathcal{I} := \{(B_i, y_i) \mid 1 \leq i \leq M\},$$

where y_i represents the label of the central pixel of B_i. Here, $\{(B_i, y_i)\}_{i=1}^M$ are exchangeable to $\mathcal{D}_\mathcal{I}$. Typically, the HSI dataset is partitioned into $\mathcal{D}_\text{train}$, $\mathcal{D}_\text{val}$, and $\mathcal{D}_\text{test}$ for training, validation, and evaluation of the deep classifier. Essentially, the HSI classifier is engineered to accurately predict the label of the central pixel within each cuboid in $\mathcal{D}_\text{test}$.

III. VALIDITY OF CONFORMAL PREDICTION ON HSI

A. Theoretical Analysis

As outlined in Theorem 1, conformal prediction requires that calibration and test samples be exchangeable. However, HSI classifiers typically operate on pixel cuboids as the fundamental unit of training, rather than individual pixels, and thus require access to the entire HSI during training. Therefore, when partitioning $\mathcal{D}_\mathcal{I}$ into $\mathcal{D}_\text{train}$, $\mathcal{D}_\text{cal}$, and $\mathcal{D}_\text{test}$, cuboids in $\mathcal{D}_\text{train}$ may overlap with those in $\mathcal{D}_\text{cal}$ or $\mathcal{D}_\text{test}$ due to shared pixels, which potential breaks the validity of conformal prediction. In this section, we demonstrate that, under a general permutation-invariance condition, the exchangeability of the non-conformity scores remains valid even when pixels from both the calibration and test sets are used during training. This theoretical analysis conclusively establishes the validity of applying conformal prediction to HSI classification.

Assumption 1. *For any instance cuboid B and any label $y \in \mathcal{Y}$, the non-conformity score $S(B, y)$ is invariant under any permutation of the union $\mathcal{D}_{cal} \cup \mathcal{D}_{test}$.*

Assumption 1 implies that the non-conformity scores for any instance-label pairs should remain unaffected by the choice of $\mathcal{D}_\text{cal}$, thereby imposing a permutation-invariance condition on the training of the HSI classifier. This ensures that the output probabilities and, consequently, the non-conformity scores are invariant to permutations of the calibration and test data of the HSI. **HSI classifiers typically satisfy this assumption, as they do not take into account the ordering of pixels (or cuboids) during training.** Based on this fact, we then demonstrate that any HSI classifier can produce valid prediction sets and offer a test-time coverage guarantee.

Theorem 2. *Let $\{(B_i, y_i)\}_{i=1}^n$ and $\{(B_{n+j}, y_{n+j})\}_{j=1}^m$ denote the calibration and test sets, respectively. Assume that the non-conformity score function S, derived from the pre-trained HSI classifier, satisfies Assumption 1. Define the non-conformity scores for the calibration and test sets as $\{s_i\}_{i=1}^n$ and $\{s_{n+j}\}_{j=1}^m$, where $s_i := S(B_i, y_i)$. Under these conditions, the scores $\{s_i\}_{i=1}^n$ and $\{s_{n+j}\}_{j=1}^m$ are exchangeable.*

The proof of Theorem 2 is provided in Appendix A. This theorem establishes that exchangeability is preserved, thereby

[1]In prior research, the terms "cuboids", "cubes", and "patches" are often used interchangeably to describe local regions of HSIs. For clarity, this work consistently uses the term "cuboids".

validating the application of conformal prediction to HSI classification. Specifically, it guarantees that the generated prediction sets will include the true label with a user-specified coverage probability. Additionally, we empirically demonstrate this coverage guarantee in our experiments, as detailed in Table II. Building on this theoretical foundation, we propose a conformal procedure that equips any trained HSI classifier with reliable prediction sets, ensuring that these sets contain the true labels with a user-specified probability.

B. Conformal HSI Classification

For an HSI $\mathcal{I}$ with a randomly selected subset of known labels, the following steps outline the process for generating conformal prediction sets with a coverage rate of $1 - \alpha$:

1. For each labeled pixel, construct a cuboid and randomly partition these cuboids into two subsets: $\mathcal{D}_\text{train}$, containing N_1 samples, and $\mathcal{D}_\text{cal}$, containing N_2 samples. Similarly, construct a cuboid for each unlabeled pixel to form the set $\mathcal{D}_\text{test}$, which includes N_3 instances.
2. Train a HSI classifier $\hat{\pi}_\theta$ using the training subset $\mathcal{D}_\text{train}$.
3. Define the score function S based on the HSI classifier $\hat{\pi}_\theta$ and compute the non-conformity scores $s_i = S(B_i, y_i)$ for each sample in the calibration set $\mathcal{D}_\text{cal}$.
4. Determine the threshold τ using $\{s_i\}_{i=1}^{N_2}$ by solving:

$$\tau = \inf \left\{ s \mid \frac{|\{i \mid s_i \leq s\}|}{N_2} \geq \frac{\lceil (N_2 + 1)(1 - \alpha) \rceil}{N_2} \right\}.$$

5. For each $B_i \in \mathcal{D}_\text{test}$, derive the prediction set as follows:

$$\mathcal{C}_{1-\alpha}(B_i; \tau) := \{y \in \mathcal{Y} \mid S(B_i, y) \leq \tau\}.$$

Theorem 2 indicates that the non-conformity scores of $\mathcal{D}_\text{cal}$ and $\mathcal{D}_\text{test}$ are exchangeable. Thus, according to Theorem 1, it follows that the prediction set $\mathcal{C}_{1-\alpha}(B_i; \tau)$ will include the true label of the central pixel of B_i with a probability of $1 - \alpha$. For the non-conformity score function S, we can utilize any well-defined non-conformity score functions, such as APS, RAPS, or SAPS. However, these score functions rely solely on softmax probabilities and label ranks, overlooking the critical spatial information inherent in HSIs. This observation motivates the exploration of integrating spatial information into the score calculation for enhanced conformal prediction.

IV. SPATIAL-AWARE CONFORMAL PREDICTION

A. Motivation

Numerous studies have demonstrated that leveraging spatial information can significantly enhance HSI classification accuracy [35], [36]. However, in the context of conformal prediction, existing non-conformity scores are computed solely based on the output conditional probability of the central pixel and label rankings, resulting in a considerable neglect of spatial information. In the following, we present an theoretical analysis illustrating the impact of incorporating spatial information into the calculation of non-conformity scores.

Given a non-conformity score function S based on an HSI classifier $\hat{\pi}_\theta$, the score for instance-label pair (B_i, y_i)
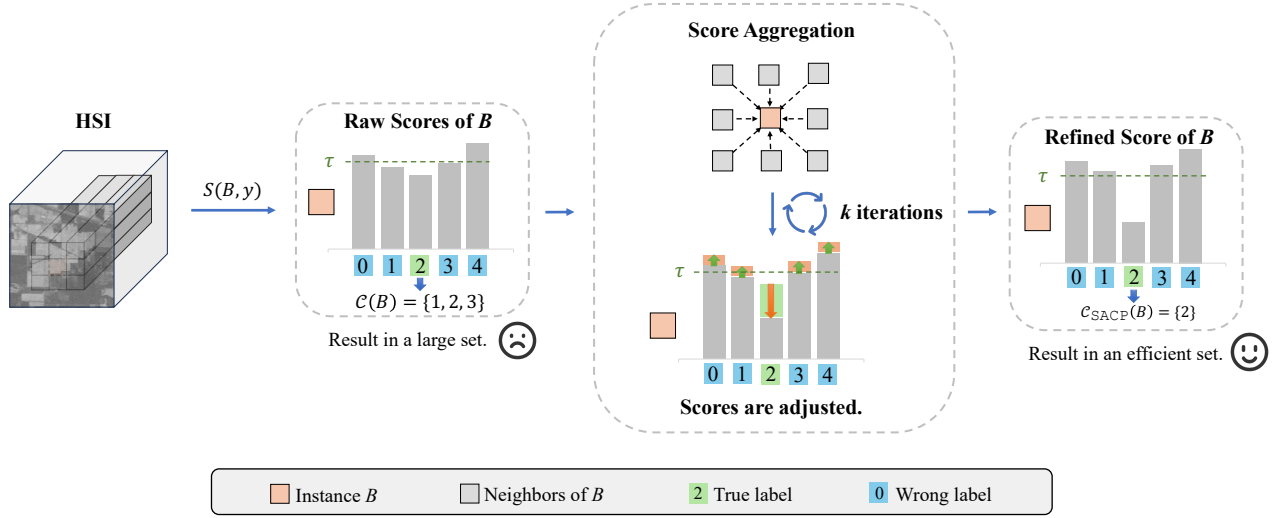
Fig. 1. **Overview of SACP**: Raw non-conformity scores lead to excessively large prediction sets. SACP addresses this issue by refining the scores through aggregation of instances with high spatial correlation. This approach is a simple yet effective solution for enhancing the efficiency of prediction sets.

is $S(B_i, y_i)$. To incorporate information from neighboring cuboids, we define the refined score function $\tilde{S}(B_i, y_i)$ by:

$$\tilde{S}(B_i, y_i) = S(B_i, y_i) + \frac{1}{|\mathcal{N}_i|} \sum_{B_j \in \mathcal{N}_i} S(B_j, y_i), \quad (4)$$

where $\mathcal{N}_i$ denotes the set of neighboring cuboids around the central pixel of B_i. Intuitively speaking, since neighboring pixels in a given HSI often belong to the same class, if y_i is the true label for B_i, it is likely that y_i is also the true label for $B_j \in \mathcal{N}_i$. Consequently, the term $\frac{1}{|\mathcal{N}_i|} \sum_{B_j \in \mathcal{N}_i} S(B_j, y_i)$ will generally be small. Conversely, if y_i is not the correct label for B_i, this summation is expected to be larger. Therefore, compared to $S(B_i, y_i)$, the refined score $\tilde{S}(B_i, y_i)$ increases modestly for correct instance-label pairs and substantially for incorrect ones. To clarify, the advantages of this aggregation are formally demonstrated in the following proposition.

Proposition 1. *Consider two different score functions S and $\hat{S}$ based on an HSI classifier. Let $\mathcal{D}_{cal}$ represent the calibration set, containing n samples, and let $\mathcal{D}_{test}$ denote the test set, containing n' instances. We define $\hat{\mathcal{D}}_{test}$ as follows:*

$$\hat{\mathcal{D}}_{test} := \{(B_i, y) \mid (B_i, y_i) \in \mathcal{D}_{test}, y \in \mathcal{Y}\}.$$

Let $\{s_i\}_{i=1}^{n}$ and $\{\hat{s}_i\}_{i=1}^{n}$ be the respective scores for $\mathcal{D}_{cal}$ based on S and $\hat{S}$. The scores for $\hat{\mathcal{D}}_{test}$ are denoted by $\{t_k\}_{k=1}^{n'|\mathcal{Y}|}$ and $\{\hat{t}_k\}_{k=1}^{n'|\mathcal{Y}|}$, respectively. We define

$$R(\{s_i\}_{i=1}^{n}, \{t_k\}_{k=1}^{n'|\mathcal{Y}|}) = \frac{1}{n \times n'|\mathcal{Y}|} \sum_{i=1}^{n} \sum_{k=1}^{n'|\mathcal{Y}|} \mathbb{1}_{\{s_i > t_k\}},$$

which represents the empirical probability that a score from $\mathcal{D}_{cal}$ is greater than the one from $\hat{\mathcal{D}}_{test}$. Then, we have

$$R(\{s_i\}_{i=1}^{n}, \{t_k\}_{k=1}^{n'|\mathcal{Y}|}) > R(\{\hat{s}_i\}_{i=1}^{n}, \{\hat{t}_k\}_{k=1}^{n'|\mathcal{Y}|}) \Leftrightarrow$$

$$\sum_{B:(B,y) \in \mathcal{D}_{test}} \int_0^1 |\mathcal{C}_{1-\alpha}(B)| \, d\alpha > \sum_{B:(B,y) \in \mathcal{D}_{test}} \int_0^1 |\hat{\mathcal{C}}_{1-\alpha}(B)| \, d\alpha,$$

where $\mathcal{C}(\cdot)$ and $\hat{\mathcal{C}}(\cdot)$ represent the prediction sets yielded by S and $\hat{S}$, and α denotes the error rate.

The proof of Proposition 1 is provided in Appendix B. This proposition suggests that the probability of a randomly selected score from $\mathcal{D}_{cal}$ exceeding a randomly selected one from $\hat{\mathcal{D}}_{test}$ serves as an indicator of the efficiency of conformal predictors: a lower probability indicates more efficient prediction sets. For the refined score $\tilde{S}$ in Eq. (4), which incorporates aggregation within neighborhoods, scores in $\hat{\mathcal{D}}_{test}$ typically show a larger increase than those in $\mathcal{D}_{cal}$. This discrepancy occurs because the instance-label pairs in $\mathcal{D}_{cal}$ are always correctly matched, while the defined set $\hat{\mathcal{D}}_{test}$ mostly contains mismatched pairs. Therefore, compared to standard conformal prediction, the integration of spatial information (via Eq. (4)) is expected to lower the probability defined in Proposition 1, resulting in more efficient prediction sets.

B. Methodology

In our preceding analysis, we demonstrate that incorporating scores from neighboring pixels potentially enhances the efficiency of prediction sets. To fully leverage the spatial information inherent in HSIs, we introduce Spatial-Aware Conformal Prediction (SACP). Central to this framework is the Score Aggregation Operator $\mathcal{V}_k$, which is defined as follows:

$$\mathcal{V}_k(B_i, y) = (1 - \lambda)\mathcal{V}_{k-1}(B_i, y) + \frac{\lambda}{|\mathcal{N}_i|} \sum_{B_j \in \mathcal{N}_i} \mathcal{V}_{k-1}(B_j, y),$$

$$(5)$$

where k represents the iteration number, and $\mathcal{V}_0 = S$ is the utilized base score function. The parameter λ serves as a trade-off parameter to balance the influence of the neighborhood information. The term $\mathcal{N}_i$ denotes the set of neighboring cuboids around the central pixel of B_i that are not in $\mathcal{D}_{train}$. For $k = 1$, each score is aggregated with its immediate neighbors. As k increases, the scores are computed based on previously aggregated scores, thus incorporating spatial information from

progressively larger neighborhoods. An overview of SACP is illustrated in Figure 1. While raw scores often result in excessively large prediction sets, SACP refines these scores by aggregating them according to spatial correlations, thereby yielding more efficient prediction sets. Additionally, the complete pipeline of SACP is detailed in Algorithm 1.

Algorithm 1 Spatial-Aware Conformal Prediction (SACP)

Input: Training set $\mathcal{D}_{\text{train}}$, calibration set $\mathcal{D}_{\text{cal}}$, test cuboid B_{n+1}, error rate α, and parameters k and λ.

Output: Conformal prediction set $\mathcal{C}_{1-\alpha}(B_{n+1}; \hat{\tau})$.

Step 1: Train a deep HSI classifier $\hat{\pi}_{\theta}$ using $\mathcal{D}_{\text{train}}$.

Step 2: Define the score function S based on $\hat{\pi}_{\theta}$.

Step 3: Compute the scores for $\mathcal{D}_{\text{cal}}$ and aggregate the results using the Score Aggregation Operator defined in Eq. (5). Denote the aggregated scores as $\{\hat{s}_i\}_{i=1}^{|\mathcal{D}_{\text{cal}}|}$.

Step 4: Determine the threshold $\hat{\tau}$ by:

$$\hat{\tau} = \inf \left\{ s \mid \frac{|\{i \mid \hat{s}_i \leq s\}|}{N_2} \geq \frac{\lceil (N_2 + 1)(1 - \alpha) \rceil}{N_2} \right\}.$$

Step 5: For B_{n+1}, the prediction set is derived as:

$$\hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau}) := \{ y \in \mathcal{Y} \mid \mathcal{V}_k(B_{n+1}, y) \leq \hat{\tau} \},$$

where $\mathcal{V}_k$ is the Score Aggregation Operator in Eq. (5).

To demonstrate that SACP can produce trustworthy prediction sets, we also theoretically establish that SACP ensures coverage, which is formalized in the following proposition.

Proposition 2. *Consider the Score Aggregation Operator $\mathcal{V}_k$, with $\mathcal{V}_0 = S$ as the initial score function. Let $\{(B_i, y_i)\}_{i=1}^n$ be the calibration set and (B_{n+1}, y_{n+1}) be a sample from the test set. Denote the aggregated scores for the calibration set by $\{\hat{s}_i\}_{i=1}^n$, where $\hat{s}_i = \hat{s}_i^{(k)} := \mathcal{V}_k(B_i, y_i)$. Suppose the threshold $\hat{\tau}$ is computed by:*

$$\hat{\tau} = \inf \left\{ s \mid \frac{|\{i \mid \hat{s}_i \leq s\}|}{n} \geq \frac{\lceil (n + 1)(1 - \alpha) \rceil}{n} \right\}.$$

Then, the following coverage guarantee holds:

$$\mathbb{P}\left(y_{n+1} \in \hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau}) \right) \geq 1 - \alpha,$$

where $\hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau})$ is formulated as

$$\hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau}) := \{ y \in \mathcal{Y} \mid \mathcal{V}_k(B_{n+1}, y) \leq \hat{\tau} \}.$$

Proposition 2 can be proved using mathematical induction, and the proof is provided in Appendix C. This proposition formally establishes that SACP maintains the coverage guarantee and can generate prediction sets with the specified coverage probability. Moreover, it offers several compelling advantages:

- **Model-agnostic:** SACP is a post-hoc method that can be incorporated into **any** pre-trained HSI classifier without necessitating modifications to the model itself.
- **Easy to use:** SACP requires minimal hyperparameter tuning, as it exhibits insensitivity to the iteration count k and the term weight λ (see Figure 3).
- **Computationally efficient:** SACP maintains computational efficiency comparable to that of standard conformal prediction (see the discussion in Section V-C0b).

TABLE I
ALLOCATION OF THE THREE DATASETS ACROSS DIFFERENT CLASSIFIERS.

Model	$\mathcal{D}_{\mathcal{I}}$	$|\mathcal{D}_{\text{train}}|$	$|\mathcal{D}_{\text{cal}}| + |\mathcal{D}_{\text{test}}|$
1D-CNN	IP	250	9999
	PU	103	42673
	SA	244	53885
3D-CNN	IP	128	10113
	PU	52	42724
	SA	122	54007
HybridSN	IP	128	10113
	PU	52	42724
	SA	122	54007
SSTN	IP	128	10113
	PU	52	42724
	SA	122	54007

TABLE II
PERFORMANCE OF THE PRE-TRAINED HSI CLASSIFIERS USED IN OUR EXPERIMENTS.

$\mathcal{D}_{\mathcal{I}}$	Metrics	1D-CNN	3D-CNN	HybridSN	SSTN
IP	OA (%)	68.44	68.93	75.25	88.35
	AA (%)	64.38	69.10	74.53	87.43
PU	OA (%)	70.45	68.97	68.31	89.39
	AA (%)	67.88	65.81	51.82	88.26
SA	OA (%)	86.44	86.11	87.91	88.52
	AA (%)	85.20	87.61	87.35	88.61

V. EXPERIMENTS

A. Experimental Setup

a) Datasets: We utilize the commonly used benchmarks Indian Pines (IP), Pavia University (PU), and Salinas (SA) datasets[2] for our evaluations. Specifically, IP consists of 145×145 pixels and 200 valid spectral bands, representing 16 different land-cover classes. PU includes 610×340 pixels with 103 usable spectral bands, covering 9 land-cover categories. SA comprises 512×217 pixels with 204 effective spectral bands, corresponding to 16 ground-truth classes. For all HSI datasets, we start by randomly sampling HSI cuboids to create the training set. The remaining data is then divided into calibration and test sets according to a pre-defined ratio.

b) Pre-trained HSI classifiers: To validate the generalizability of standard conformal prediction and SACP in HSI classification, we employ four widely used HSI classifiers, including both CNN-based and Transformer-based models, each featuring distinct scales and architectures. For CNN-based models, we select 1D-CNN [37], 3D-CNN [38], and HybridSN [39]. The 1D-CNN and 3D-CNN models differ in their convolutional kernel dimensions, while HybridSN leverages multi-dimensional kernels to improve feature extraction. As for the Transformer-based model, we employ SSTN [40]. Each

---

[2] These datasets are available at https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

TABLE III

Performance comparison of prediction sets generated by SACP and standard conformal prediction (SCP) across various HSI classifiers with different error rates α. **Bold** numbers indicate that SACP is superior. $\downarrow$ indicates that smaller values are better.

$\mathcal{D}_\mathcal{I}$	Model	Score	w/ SCP \ w/ SACP					
			$\alpha = 0.05$			$\alpha = 0.1$		
			Coverage	Size ($\downarrow$)	SSCV ($\downarrow$)	Coverage	Size ($\downarrow$)	SSCV ($\downarrow$)
Indian Pines	1D-CNN	APS	0.95 \ 0.95	3.68 \ **2.28**	0.41 \ **0.28**	0.90 \ 0.90	2.52 \ **1.75**	0.51 \ **0.45**
		RAPS	0.95 \ 0.94	4.09 \ **2.29**	0.54 \ 0.57	0.90 \ 0.90	2.54 \ **1.74**	0.88 \ **0.30**
		SAPS	0.95 \ 0.95	6.68 \ **4.31**	2.48 \ **1.83**	0.90 \ 0.90	5.56 \ **3.02**	4.84 \ **1.95**
	3D-CNN	APS	0.94 \ 0.95	5.73 \ **3.27**	0.47 \ **0.38**	0.90 \ 0.90	2.85 \ **2.06**	0.35 \ **0.34**
		RAPS	0.95 \ 0.95	4.12 \ **3.92**	0.10 \ 0.21	0.90 \ 0.90	3.21 \ **2.38**	0.11 \ 0.25
		SAPS	0.95 \ 0.95	6.40 \ **5.44**	0.98 \ **0.95**	0.91 \ 0.90	4.97 \ **3.72**	1.61 \ **1.51**
	HybridSN	APS	0.95 \ 0.95	5.56 \ **4.83**	0.20 \ **0.16**	0.90 \ 0.90	3.28 \ **2.74**	0.92 \ **0.15**
		RAPS	0.95 \ 0.95	7.34 \ **7.12**	0.40 \ 0.96	0.90 \ 0.90	4.29 \ **3.95**	0.98 \ **0.75**
		SAPS	0.95 \ 0.95	6.79 \ **6.07**	0.19 \ 0.41	0.90 \ 0.90	4.03 \ **3.40**	0.64 \ **0.61**
	SSTN	APS	0.95 \ 0.95	2.81 \ **1.73**	0.42 \ **0.18**	0.90 \ 0.90	2.00 \ **1.38**	0.41 \ 0.47
		RAPS	0.95 \ 0.95	2.52 \ **1.62**	0.29 \ 0.30	0.90 \ 0.90	1.87 \ **1.36**	0.29 \ 0.34
		SAPS	0.95 \ 0.95	6.98 \ **4.16**	4.38 \ **1.74**	0.90 \ 0.90	5.33 \ **3.25**	6.35 \ **3.09**
Pavia University	1D-CNN	APS	0.95 \ 0.95	2.26 \ **1.92**	0.39 \ **0.37**	0.90 \ 0.90	1.65 \ **1.57**	0.40 \ **0.27**
		RAPS	0.95 \ 0.95	2.00 \ **1.83**	0.23 \ 0.31	0.90 \ 0.90	1.59 \ **1.54**	0.40 \ **0.29**
		SAPS	0.95 \ 0.95	3.92 \ 3.99	1.77 \ 2.21	0.90 \ 0.90	3.44 \ **3.04**	3.62 \ **2.68**
	3D-CNN	APS	0.94 \ 0.95	2.77 \ **2.34**	1.04 \ **0.69**	0.89 \ 0.89	2.14 \ **1.79**	0.94 \ **0.66**
		RAPS	0.94 \ 0.94	2.57 \ **2.28**	0.41 \ 0.46	0.89 \ 0.89	2.04 \ **1.76**	0.64 \ **0.56**
		SAPS	0.94 \ 0.94	4.80 \ **4.31**	4.56 \ **3.85**	0.89 \ 0.89	4.04 \ **3.32**	6.56 \ **4.23**
	HybridSN	APS	0.95 \ 0.95	4.79 \ **4.59**	4.59 \ **3.58**	0.90 \ 0.90	3.38 \ **3.01**	2.39 \ **1.33**
		RAPS	0.94 \ 0.95	5.50 \ **5.36**	0.47 \ **0.36**	0.89 \ 0.90	3.70 \ 3.74	0.81 \ **0.07**
		SAPS	0.95 \ 0.95	5.57 \ **5.44**	1.98 \ 2.35	0.90 \ 0.90	3.99 \ **3.71**	3.33 \ 3.47
	SSTN	APS	0.95 \ 0.95	1.75 \ **1.24**	0.22 \ 0.26	0.90 \ 0.90	1.39 \ **1.11**	0.23 \ 0.29
		RAPS	0.95 \ 0.95	1.60 \ **1.22**	0.20 \ 0.20	0.90 \ 0.90	1.33 \ **1.10**	0.13 \ 0.23
		SAPS	0.95 \ 0.95	3.26 \ **2.24**	2.07 \ **0.92**	0.90 \ 0.90	2.75 \ **1.91**	2.96 \ **1.64**
Salinas	1D-CNN	APS	0.95 \ 0.95	1.40 \ **1.20**	0.15 \ 0.15	0.90 \ 0.90	1.20 \ **1.09**	0.18 \ 0.25
		RAPS	0.95 \ 0.95	1.37 \ **1.20**	0.06 \ 0.20	0.90 \ 0.90	1.19 \ **1.07**	0.23 \ 0.25
		SAPS	0.95 \ 0.95	3.63 \ **1.71**	1.20 \ **0.16**	0.90 \ 0.90	2.97 \ **1.28**	2.04 \ **0.16**
	3D-CNN	APS	0.95 \ 0.95	1.48 \ **1.25**	0.20 \ **0.19**	0.90 \ 0.90	1.17 \ **1.08**	0.12 \ 0.16
		RAPS	0.95 \ 0.95	1.47 \ **1.24**	0.13 \ 0.18	0.90 \ 0.90	1.15 \ **1.07**	0.15 \ 0.17
		SAPS	0.95 \ 0.95	2.94 \ **2.02**	0.58 \ **0.28**	0.90 \ 0.90	2.39 \ **1.29**	1.07 \ **0.13**
	HybridSN	APS	0.95 \ 0.95	1.37 \ **1.09**	0.18 \ **0.07**	0.90 \ 0.90	1.10 \ **1.03**	0.18 \ 0.23
		RAPS	0.95 \ 0.95	1.20 \ **1.07**	0.12 \ **0.10**	0.90 \ 0.90	1.06 \ **1.00**	0.12 \ 0.31
		SAPS	0.95 \ 0.95	1.90 \ **1.37**	0.42 \ **0.33**	0.90 \ 0.90	1.66 \ **1.18**	0.69 \ **0.31**
	SSTN	APS	0.95 \ 0.95	1.70 \ **1.19**	0.21 \ **0.16**	0.90 \ 0.90	1.37 \ **1.08**	0.23 \ **0.10**
		RAPS	0.95 \ 0.95	1.60 \ **1.18**	0.12 \ 0.14	0.90 \ 0.90	1.29 \ **1.06**	0.10 \ 0.11
		SAPS	0.95 \ 0.95	5.42 \ **2.65**	3.58 \ **0.86**	0.90 \ 0.90	3.91 \ **2.07**	3.64 \ **1.26**

HSI classifier is initially pre-trained on the three datasets and subsequently evaluated on the corresponding calibration and test data. We follow the SSTN training framework available on GitHub[3] for pre-training these classifiers. Table I details the sizes of the training, calibration, and test sets for each dataset across the models. By default, we set the calibration and test sets to be of equal size. Notably, our method remains stable across varying calibration set sizes (see Table IV).

Due to its relatively poor performance, the 1D-CNN is allocated a larger amount of training data compared to the other three models. For clarity, Table II demonstrates the Overall Accuracy (OA) and Average Accuracy (AA) results for the pre-trained HSI classifiers across the three datasets. The definitions of OA and AA are given as follows:

$$\text{OA} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \mathbb{1}_{\{\hat{y}_i = y_i\}}, \quad (6)$$

$$\text{AA} = \frac{1}{|\mathcal{Y}|} \sum_{y_j \in \mathcal{Y}} \frac{\sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \mathbb{1}_{\{\hat{y}_i = y_i, y_i = y_j\}}}{\sum_{i=1}^{|\mathcal{D}_{\text{test}}|} \mathbb{1}_{\{y_i = y_j\}}}, \quad (7)$$

where $\mathcal{D}_{\text{test}} := \{(B_i, y_i)\}_{i=1}^{|\mathcal{D}_{\text{test}}|}$, $\hat{y}_i = \operatorname{argmax}(\hat{\pi}_\theta(B_i))$, $\mathbb{1}$ is the indicator function, and $j \in \{1, \ldots, |\mathcal{Y}|\}$. **Note**: In HSI classifier pre-training, the primary focus is not on maximizing their performance metrics such as OA and AA, but rather on
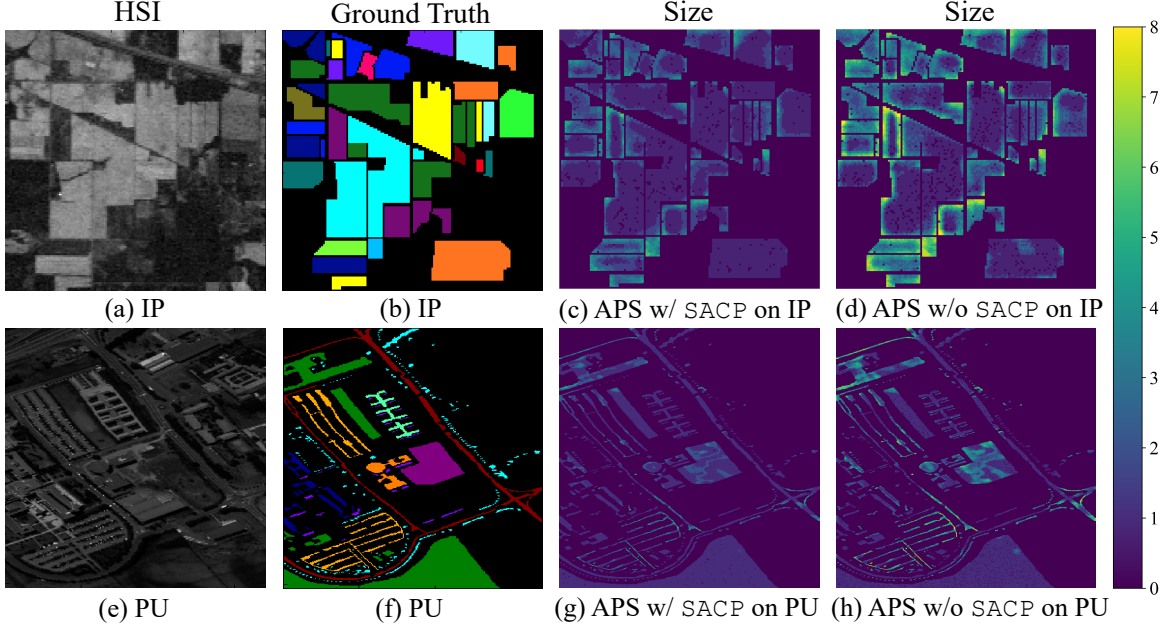
[3]https://github.com/zilongzhong/SSTN

Fig. 2. Visualizations of the average size of prediction sets for each pixel in $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{test}}$ are presented for the (a) IP and (e) PU datasets. The true labels for each class are shown in (b) and (f). The brightness of each pixel indicates the size of its prediction set, with brighter pixels representing larger sets. Panels (c) and (g) display the set sizes with SACP, while panels (d) and (h) show the set sizes with standard conformal prediction.

achieving model diversity. Our goal is to validate the effectiveness—i.e., the generalizability—of our approach across different model architectures and training processes, irrespective of their performance levels. Given that our primary goal is not to optimize classification performance, we utilize only a minimal amount of data for pre-training these classifiers.

c) Implementation Details: The training configurations for all four HSI classifiers adhere to their recommended settings. Each model is trained for 200 epochs utilizing the Adam optimizer [41], with a batch size of 128 and a learning rate of 0.002. For SACP, the trade-off parameter λ is **fixed** at 0.5, and the iteration count k is set to 1 for all experiments. The primary evaluation metrics are Coverage, Size, and SSCV, as detailed in Section II-A0d. To enhance statistical stability, each experiment is repeated 30 times, and the results are averaged for reporting. All experiments are performed on an NVIDIA GeForce RTX 3090 using PyTorch [42].

B. Main Results

a) SACP generates more efficient prediction sets: We compare the performance of prediction sets generated by standard conformal prediction and SACP across various HSI classifiers and datasets with different error rates α. The results are presented in Table III, showing that SACP consistently maintains the desired coverage of $1 - \alpha$. The results empirically validates the coverage guarantee of SACP, as outlined in Proposition 2. In other words, for each pixel, its true label will be reliably included in the prediction set with the user-specified probabilities. Additionally, SACP consistently produces more efficient prediction sets while maintaining a satisfactory conditional coverage rate. For instance, when applying 3D-CNN to the IP dataset, SACP reduces the size
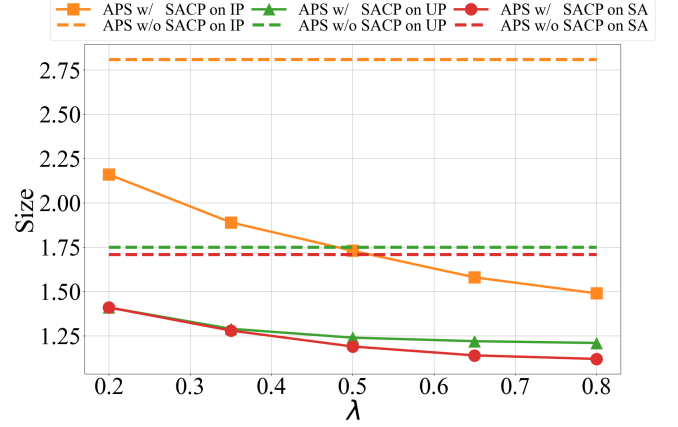
(a) Effect of λ on Size of SACP.
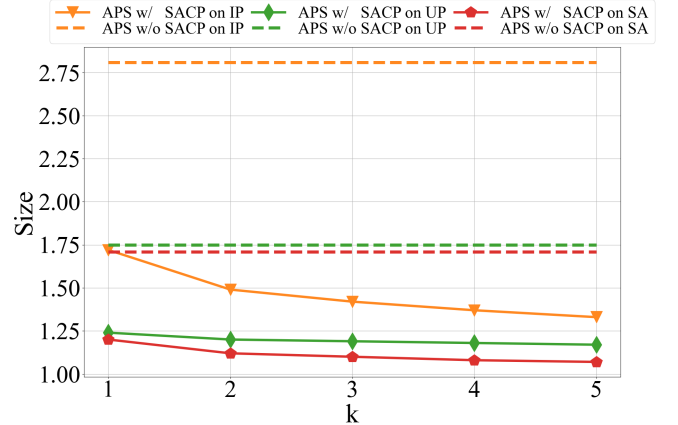
(b) Effect of k on Size of SACP.

Fig. 3. **Effect of λ and k on Prediction Set Size:** Each subplot presents results from three comparative experiments conducted on different datasets. The dashed line indicates the performance of APS without SACP.

TABLE IV

EFFECT OF CALIBRATION SET SIZE ON SACP: THE RATIO OF CALIBRATION SAMPLES IS GIVEN BY $\gamma = \frac{|\mathcal{D}_{\text{CAL}}|}{|\mathcal{D}_{\text{CAL}} \cup \mathcal{D}_{\text{TEST}}|}$.

$\mathcal{D}_{\mathcal{I}}$		$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.35$	$\gamma = 0.4$	$\gamma = 0.45$	$\gamma = 0.5$	$\gamma = 0.55$	$\gamma = 0.6$
IP	Coverage	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	Size	1.83	1.70	1.73	1.75	1.74	1.73	1.75	1.73	1.73	1.76	1.72
PU	Coverage	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	Size	1.24	1.25	1.24	1.24	1.26	1.25	1.25	1.24	1.24	1.25	1.24
SA	Coverage	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	Size	1.23	1.20	1.21	1.19	1.19	1.20	1.20	1.19	1.19	1.20	1.19

from 5.73 for APS to 3.27, representing a 42.9% decrease. Similarly, for RAPS on the same dataset with 1D-CNN, SACP lowers the size from 3.64 to 2.26, reflecting a 37.9% reduction. These improvements are observed across different error rates α, non-conformity score functions, model architectures, and datasets, highlighting the robustness and generalizability of SACP. Notably, unlike traditional classification tasks, SAPS and RAPS do not consistently outperform APS in the context of HSI classification, underscoring the challenges associated with conformal prediction in this field. Nevertheless, SACP consistently improves the performance of various non-conformity score functions, demonstrating its effectiveness.

b) SACP consistently reduces prediction set size across various pixels: To further demonstrate the effectiveness of SACP, we visualize the size of prediction sets for each pixel of $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{test}}$ in Figure 2. This experiment is conducted on the IP and PU datasets using SSTN, with APS as the base non-conformity score function and α set to 0.05. In the visualization, brighter pixels represent larger prediction sets. The results clearly indicate that SACP significantly reduces the size of prediction sets compared to standard conformal prediction. Moreover, standard conformal prediction often produces unusually large prediction sets, even for interior pixels, as evidenced in the lower-left corner of Figure 2(d). Typically, edge pixels, which are more challenging to predict, tend to have larger prediction sets, while interior pixels have smaller sets. SACP effectively addresses this issue by consistently reducing prediction set sizes for both edge and interior pixels. The experimental results further illustrate the superior performance of our approach in conformal HSI classification.

c) SACP is insensitive to hyperparameters: In SACP, two hyperparameters are crucial: λ and k. The parameter λ serves as a trade-off, balancing the influence of the neighborhood, while k specifies the number of iterations. Increasing k allows SACP to incorporate spatial information from progressively larger neighborhoods. This section examines the impact of variations in λ and k on the performance of SACP. The experiment is conducted on all selected datasets using SSTN, with APS as the score function and α set to 0.05. The results are illustrated in Figure 3. Figure 3a demonstrates that SACP consistently outperforms the baseline across different values of λ, with the size of prediction sets decreasing as λ increases. Figure 3b illustrates that SACP consistently achieves
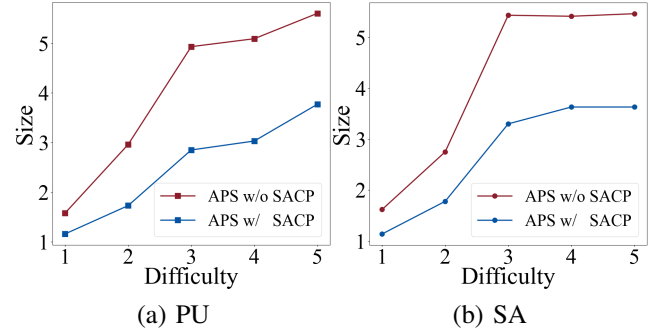

(a) PU (b) SA

Fig. 4. **Size vs. Difficulty**: The term "Difficulty" refers to the ranking of the true class label in the output probability, reflecting the difficulty of accurately classifying the pixel. We report the average size of prediction sets for the PU and SA datasets using SSTN, categorized by their Difficulty.

superior performance as the number of iterations k changes. Overall, the results suggest that SACP is relatively insensitive to variations in λ and k, consistently surpassing the baseline. For simplicity, we employ a moderate range of hyperparameters, consistently setting $\lambda = 0.5$ and $k = 1$ throughout the experiments. Further tuning these parameters on a hold-out set may further enhance the performance of conformal prediction.

d) SACP is robust to the size of calibration set: As shown in Eq. (2), the calibration set is used to determine the threshold τ corresponding to the coverage rate of $1 - \alpha$. A precise estimate of τ is essential. In this work, all experiments are conducted with a fixed ratio for the calibration set size, specifically $|\mathcal{D}_{\text{cal}}| \approx 0.5 \times |\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}|$. In this section, we investigate whether SACP is sensitive to the size of the calibration set $|\mathcal{D}_{\text{cal}}|$. Specifically, we exam whether Coverage and Size remain satisfactory when varying the ratio of the calibration set size. The experiment is conducted on all three datasets using SSTN, with APS as the base score function and the error rate α set to 0.05. The ratio of calibration samples is defined as $\gamma = (|\mathcal{D}_{\text{cal}}| / |\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}|)$. γ ranges from 0.1 to 0.6 in increments of 0.05. The experimental results are presented in Table IV. Coverage consistently maintains $1 - \alpha$ as the calibration set size decreases relative to $|\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}|$, even for very small values of γ. Additionally, the size of the prediction sets remains stable, only marginally influenced by changes in the calibration set size, and fluctuates within a narrow range.

These findings collectively demonstrate that SACP exhibits significant robustness to variations in the calibration set size.

C. Discussion

a) SACP exhibits adaptiveness: Adaptiveness is a crucial attribute of conformal prediction, ensuring that prediction sets accurately reflect the uncertainty associated with each instance. Specifically, prediction sets should be larger for more challenging instances and smaller for easier ones. This analysis evaluates whether SACP maintains this characteristic. We use the true label rank $o(y, \hat{\pi}_\theta(\boldsymbol{x}))$ to represent the difficulty of the instance $\boldsymbol{x}$, where a higher true label rank indicates greater classification difficulty. Instances are categorized into five groups based on their true label rank: 1, 2, 3, 4, and $\{5, \ldots, |\mathcal{Y}|\}$. We then compute the average size of prediction sets for each group. This experiment is conducted on the SA and PU datasets using SSTN, with APS as the base score function and the error rate α set to 0.05.

Figure 4 demonstrates that SACP exhibits adaptiveness across various datasets. In both cases, with or without SACP, APS consistently shows that prediction sets are smaller for easier instances compared to more challenging ones. We conclude that SACP effectively optimizes the average prediction set size for each difficulty group while preserving its adaptive characteristics. These results collectively underscore the significance of SACP in HSI classification.

b) SACP is computationally efficient: In this section, we analyze the computational efficiency of SACP. The base non-conformity score functions APS, RAPS, and SAPS involve only sorting and summation operations, resulting in a time complexity of $\mathcal{O}(|\mathcal{Y}| \log |\mathcal{Y}| N)$, where N is the number of instances and $|\mathcal{Y}|$ represents the total number of labels. The dominant computational overhead arises from the sorting operation. In comparison, SACP introduces an additional aggregation step, which adds a time complexity of $\mathcal{O}(kN)$, where k is the number of iterations. Consequently, the overall time complexity of SACP is $\mathcal{O}(|\mathcal{Y}| \log |\mathcal{Y}| N + kN)$. For simplicity, we approximate $(k + |\mathcal{Y}| \log |\mathcal{Y}|)$ as a constant C, as these terms are negligible compared to N. Therefore, the complexity can be expressed as $\mathcal{O}(CN)$, which is comparable in magnitude to that of standard conformal prediction. This analysis demonstrates that SACP exhibits computational efficiency.

VI. CONCLUSION

In this paper, we first present a comprehensive theoretical proof establishing the validity of conformal prediction for HSI classification. We then introduce a conformal procedure that integrates conformal prediction sets into HSI classifiers, ensuring the inclusion of true labels with a user-specified probability. Building on this foundation, we propose SACP, a conformal prediction framework specifically designed for HSI data. SACP refines non-conformity scores by leveraging the intrinsic spatial information present in HSIs, effectively enhancing the efficiency of conformal prediction sets. To the best of our knowledge, this is the first study to explore conformal prediction in the context of HSI data. Extensive experiments confirm the validity of conformal prediction for

HSI classification. Moreover, SACP outperforms standard conformal prediction in HSI classification by reducing the average prediction set size while maintaining satisfactory marginal and conditional coverage rates. Furthermore, SACP is practical and user-friendly, exhibiting robustness to hyperparameter variations and incurring minimal computational costs.

Limitations. The number of pixels within different classes in HSIs can sometimes be imbalanced, leading to unusually large prediction sets and potential undercoverage for classes with fewer pixels. This issue has not been addressed in our work and represents a promising avenue for future research.

APPENDIX A
PROOF OF THEOREM 2

Proof. Here, we emphasize that the non-conformity score $S(B, y)$ is dependent on the sample partitioning, i.e.,

$$S(B, y) \equiv S(B, y; \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}). \tag{8}$$

By Assumption 1, the non-conformity score $S(B, y)$ remains invariant under any permutation π of the union $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$. Therefore, we have

$$\begin{aligned} & S(B, y; \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}) \\ = & S(B, y; \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, (\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}})^\pi), \end{aligned} \tag{9}$$

where $(\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}})^\pi$ denotes permuted samples according to π, and π is permutation of $|\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}|$. Hence, the score $S(B, y)$ remains invariant with respect to the selection of any subset of $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$ as the calibration set $\mathcal{D}_{\text{cal}}$. Let $[t_1, t_2, \ldots, t_{n+m}]$ be the unordered scores from $\{S(B_i, y_i)\}_{i \in \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}}$, where $[\cdot]$ denotes unordered sets. We have $\{s_i\}_{i=1}^n$ is a subset of size n of $[t_1, t_2, \ldots, t_{n+m}]$. Under random splitting of $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$, any permutation π happens with the same probability, i.e.,

$$\mathbb{P}\left(\{s_i\}_{i=1}^n = \{t_1, t_2, \ldots, t_n\} \mid \{s_i\}_{i=1}^{n+m}\right) = \frac{1}{\binom{n+m}{n}}. \tag{10}$$

As a result, any score of test sample in $\{s_{n+j}\}_{j=1}^m$ is exchangeable with $\{s_i\}_{i=1}^n$. $\qquad \square$

APPENDIX B
PROOF OF PROPOSITION 1

Proof. We introduce the superscript α to τ as follows:

$$\tau^\alpha = \inf\left\{s \mid \frac{|\{i \mid s_i \leq s\}|}{N_2} \geq \frac{\lceil (N_2 + 1)(1 - \alpha) \rceil}{N_2}\right\}. \tag{11}$$

Denote the cuboids in $\mathcal{D}_{\text{test}}$ as $\mathcal{B}_{\text{test}}$. Then, for each cuboid B in $\mathcal{B}_{\text{test}}$, we have

$$\begin{aligned} \int_0^1 |\mathcal{C}_{1-\alpha}(B)| \mathrm{d}\alpha &= \int_0^1 |\{y \in \mathcal{Y} \mid S(B, y) \leq \tau^\alpha\}| \; \mathrm{d}\alpha \\ &= \int_0^1 \sum_{y \in \mathcal{Y}} \mathbb{1}_{\{S(B, y)) \leq \tau^\alpha\}} \; \mathrm{d}\alpha. \end{aligned} \tag{12}$$

Moreover, the event $\{S(B,y) \leq \tau^\alpha\}$ in Eq. (12) is equivalent to the event that at least α percent of the sample in $\mathcal{D}_{\text{cal}}$ have scores larger than that of (B,y). Thus, we have

$$\int_0^1 \sum_{y \in \mathcal{Y}} \mathbb{1}_{\{S(B,y)) \leq \tau^\alpha\}} \ d\alpha$$

$$= \int_0^1 \sum_{y \in \mathcal{Y}} \mathbb{1}_{\{\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}} \geq \lfloor (1+|\mathcal{D}_{\text{cal}}|)\alpha \rfloor\}} \ d\alpha$$

$$= \int_0^1 \sum_{y \in \mathcal{Y}} \mathbb{1}_{\left\{\frac{1+\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}}}{1+|\mathcal{D}_{\text{cal}}|} > \alpha\right\}} \ d\alpha. \tag{13}$$

It is obvious that

$$0 < \frac{1+\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}}}{1+|\mathcal{D}_{\text{cal}}|} \leq 1. \tag{14}$$

Thus, we have

$$\int_0^1 \sum_{y \in \mathcal{Y}} \mathbb{1}_{\left\{\frac{1+\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}}}{1+|\mathcal{D}_{\text{cal}}|} > \alpha\right\}} \ d\alpha$$

$$= \sum_{y \in \mathcal{Y}} \int_0^1 \mathbb{1}_{\left\{\frac{1+\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}}}{1+|\mathcal{D}_{\text{cal}}|} > \alpha\right\}} \ d\alpha \tag{15}$$

$$= \sum_{y \in \mathcal{Y}} \frac{1+\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}}}{1+|\mathcal{D}_{\text{cal}}|}$$

For $\mathcal{B}_{\text{test}}$, we have

$$\sum_{B \in \mathcal{B}_{\text{test}}} \int_0^1 |\mathcal{C}_{1-\alpha}(B)| \, d\alpha$$

$$= \sum_{B \in \mathcal{B}_{\text{test}}} \sum_{y \in \mathcal{Y}} \frac{1+\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}}}{1+|\mathcal{D}_{\text{cal}}|}$$

$$= \sum_{(B,y) \in \hat{\mathcal{D}}_{\text{test}}} \frac{1+\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}}}{1+|\mathcal{D}_{\text{cal}}|} \tag{16}$$

$$= \sum_{(B,y) \in \hat{\mathcal{D}}_{\text{test}}} \frac{1+\sum_{(\tilde{B},\tilde{y}) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(\tilde{B},\tilde{y}) \geq S(B,y)\}}}{1+n}$$

$$= \frac{n'|\mathcal{Y}|}{1+n} + \frac{1}{1+n} \sum_{i=1}^{n} \sum_{k=1}^{n'|\mathcal{Y}|} \mathbb{1}_{\{s_i > t_k\}}.$$

In addition, we have

$$R(\{s_i\}_{i=1}^n, \{t_k\}_{k=1}^{n'|\mathcal{Y}|}) = \frac{1}{n \times n'|\mathcal{Y}|} \sum_{i=1}^{n} \sum_{k=1}^{n'|\mathcal{Y}|} \mathbb{1}_{\{s_i > t_k\}}, \tag{17}$$

Finally, we have

$$\sum_{(B,y) \in \mathcal{D}_{\text{test}}} \int_0^1 |\mathcal{C}_{1-\alpha}(B)| \, d\alpha$$

$$= \frac{n \times n'|\mathcal{Y}|}{1+n} R(\{s_i\}_{i=1}^n, \{t_k\}_{k=1}^{n'|\mathcal{Y}|}) + \frac{n'|\mathcal{Y}|}{1+n}, \tag{18}$$

where $\frac{n \times n'|\mathcal{Y}|}{1+n}$ and $\frac{n'|\mathcal{Y}|}{1+n}$ are positive constants. Based on Eq. (18), we obviously have

$$R(\{s_i\}_{i=1}^n, \{t_k\}_{k=1}^{n'|\mathcal{Y}|}) > R(\{\hat{s}_i\}_{i=1}^n, \{\hat{t}_k\}_{k=1}^{n'|\mathcal{Y}|}) \Leftrightarrow$$

$$\sum_{(B,y) \in \mathcal{D}_{\text{test}}} \int_0^1 |\mathcal{C}_{1-\alpha}(B)| \, d\alpha > \sum_{(B,y) \in \mathcal{D}_{\text{test}}} \int_0^1 |\hat{\mathcal{C}}_{1-\alpha}(B)| \, d\alpha. \tag{19}$$

Then, we complete the proof. □

APPENDIX C
PROOF OF PROPOSITION 2

Proof. We will employ mathematical induction to demonstrate that the statement is valid for all k.

a) Base Case ($k = 0$): For $k = 0$, the operator $\mathcal{V}_k$ reduces to S. By our Theorem 2, we have $\hat{s}_i, (i = 1, \cdots, n)$ and $S(B_{n+1}, y_{n+1})$ are exchangeable. Thus,

$$\mathbb{P}\left(y_{n+1} \notin \hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau})\right)$$

$$= \mathbb{P}\left(S(B_{n+1}, y_{n+1}) > \hat{\tau}\right)$$

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(S(B_i, y_i) > \hat{\tau}\right)$$

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}\mathbb{1}_{\{i:S(B_i,y_i)>\hat{\tau}\}} \tag{20}$$

$$= \frac{1}{n+1} \mathbb{E}|\{i : S(B_i, y_i) > \hat{\tau}\}|$$

$$\leq \frac{1}{n+1}(n+1)\alpha$$

$$= \alpha.$$

Therefore, we have

$$\mathbb{P}\left(y_{n+1} \in \hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau})\right) \geq 1 - \alpha \tag{21}$$

b) Inductive Hypothesis: Assume the proposition holds for $k = m$. Specifically, assume:

$$\mathbb{P}\left(y_{n+1} \in \hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau})\right) \geq 1 - \alpha, \tag{22}$$

where

$$\hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau}) := \{y \in \mathcal{Y} \mid \mathcal{V}_m(B_{n+1}, y) \leq \hat{\tau}\}, \tag{23}$$

and $\hat{s}_i^{(m)}, i = 1, \ldots, n+1$ are exchangeable.

c) Inductive Step: To prove the proposition for $k = m+1$, we need to show:

$$\mathbb{P}\left(y_{n+1} \in \hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau})\right) \geq 1 - \alpha, \tag{24}$$

where

$$\hat{\mathcal{C}}_{1-\alpha}(B_{n+1}; \hat{\tau}) := \{y \in \mathcal{Y} \mid \mathcal{V}_{m+1}(B_{n+1}, y) \leq \hat{\tau}\}, \tag{25}$$

and $\hat{s}_i^{(m+1)}, i = 1, \ldots, n+1$ are exchangeable.

The aggregation operator for $k = m+1$ is:

$$\mathcal{V}_{m+1}(B_i, y) = (1-\lambda)\mathcal{V}_m(B_i, y) + \frac{\lambda}{|\mathcal{N}_i|} \sum_{B_j \in \mathcal{N}_i} \mathcal{V}_m(B_j, y), \tag{26}$$

where $\mathcal{N}_i$ denotes the set of neighboring cuboids around the central cuboid B_i included in $\mathcal{B}$. Here, $\mathcal{B}$ represents the set of all exchangeable cuboids of an HSI to be aggregated. Let $g(\hat{s}_1^{(m)}, \ldots, \hat{s}_{n+1}^{(m)}) = (\hat{s}_1^{(m+1)}, \ldots, \hat{s}_{n+1}^{(m+1)})^\top$. Since the

operator $g(\cdot)$ preserves exchangeability by Theorem 3 in [43], we have the aggregated scores $\hat{s}_1^{(m+1)}, \ldots, \hat{s}_{n+1}^{(m+1)}$ are still exchangeable. Correspondingly, the coverage guarantee is maintained for $k = m + 1$. By Mathematical Induction, the proposition holds for all non-negative integers k. $\qquad\square$

REFERENCES

[1] Y. Zhang, S. Yan, L. Zhang, and B. Du, "Fast projected fuzzy clustering with anchor guidance for multimodal remote sensing imagery," *IEEE Transactions on Image Processing*, Aug 2024.

[2] Y. Zhang, G. Jiang, Z. Cai, and Y. Zhou, "Bipartite graph-based projected clustering with local region guidance for hyperspectral imagery," *IEEE Transactions on Multimedia*, Apr 2024.

[3] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 1559–1572, Jan 2022.

[4] C. Zhao, B. Qin, S. Feng, W. Zhu, W. Sun, W. Li, and X. Jia, "Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning," *IEEE Transactions on Image Processing*, vol. 32, pp. 3606–3621, Jun 2023.

[5] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 2810–2825, Feb 2021.

[6] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 4251–4265, May 2022.

[7] B. Xi, J. Li, Y. Li, R. Song, D. Hong, and J. Chanussot, "Few-shot learning with class-covariance metric for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 5079–5092, Jul 2022.

[8] M. Kütük, I. Geneci, O. B. Özdemir, A. Koz, O. Esentürk, Y. Y. Çetin, and A. A. Alatan, "Ground-based hyperspectral image surveillance system for explosive detection: Methods, experiments, and comparisons," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 8747–8763, 2023.

[9] R. J. Murphy, S. Schneider, and S. T. Monteiro, "Consistency of measurements of wavelength position from hyperspectral imagery: Use of the ferric iron crystal field absorption at ∼900 nm as an indicator of mineralogy," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2843–2857, Jul 2013.

[10] X. Zhang, Y. Sun, K. Shang, L. Zhang, and S. Wang, "Crop classification based on feature band set construction and object-oriented approach using hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4117–4128, Jun 2016.

[11] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, Jul 2013.

[12] V. C. Coffey, "Hyperspectral imaging for safety and security," *Optics and Photonics News*, vol. 26, no. 10, pp. 26–33, 2015.

[13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[14] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1050–1059.

[15] R. C. Smith, *Uncertainty quantification: theory, implementation, and applications*. SIAM, 2013.

[16] X. He, Y. Chen, and L. Huang, "Toward a trustworthy classifier with deep cnn: uncertainty estimation meets hyperspectral image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[17] X. He, L. Huang, and Y. Chen, "Bayesian deep learning for hyperspectral image classification with low uncertainty," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[18] L. Yu, Y. Lou, and F. Chen, "Uncertainty-aware graph-based hyperspectral image classification," in *International Conference on Learning Representations*, 2023.

[19] J. Huang, H. Xi, L. Zhang, H. Yao, Y. Qiu, and H. Wei, "Conformal prediction for deep classifier via label ranking," 2024.

[20] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.

[21] V. Balasubramanian, S.-S. Ho, and V. Vovk, *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[22] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," 2021.

[23] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 223–234, Jun 2019.

[24] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] E. Horwitz and Y. Hoshen, "Conffusion: Confidence intervals for diffusion models," 2022.

[26] J. Wang, J. Tong, K. Tan, Y. Vorobeychik, and Y. Kantaros, "Conformal temporal logic planning using large language models: Knowing when to do what and when to ask for help," 2023.

[27] S. H. Zargarbashi, S. Antonelli, and A. Bojchevski, "Conformal prediction sets for graph neural networks," in *International Conference on Machine Learning*. PMLR, 2023, pp. 12 292–12 318.

[28] K. Huang, Y. Jin, E. Candes, and J. Leskovec, "Uncertainty quantification over graph with conformalized graph neural networks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] J. J. Cherian, I. Gibbs, and E. J. Candès, "Large language model validity via enhanced conformal prediction methods," 2024.

[30] J. Su, J. Luo, H. Wang, and L. Cheng, "Api is enough: Conformal prediction for large language models without logit-access," 2024.

[31] V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. S. Jaakkola, and R. Barzilay, "Conformal language modeling," 2023.

[32] Y. Romano, M. Sesia, and E. Candes, "Classification with valid and adaptive coverage," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3581–3591, 2020.

[33] A. N. Angelopoulos, S. Bates, M. I. Jordan, and J. Malik, "Uncertainty sets for image classifiers using conformal prediction," in *International Conference on Learning Representations*. OpenReview.net, 2021.

[34] M. Imani and H. Ghassemian, "An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges," *Information Fusion*, vol. 59, pp. 59–83, Jan 2020.

[35] N. He, M. E. Paoletti, J. M. Haut, L. Fang, S. Li, A. Plaza, and J. Plaza, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 755–769, Aug 2018.

[36] J. Zhang, Y. Zhang, and Y. Zhou, "Quantum-inspired spectral-spatial pyramid network for hyperspectral image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9925–9934.

[37] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, no. 1, p. 258619, 2015.

[38] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-d deep learning approach for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4420–4434, Apr 2018.

[39] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, Jun 2019.

[40] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, Oct 2021.

[41] D. Kingma, "Adam: a method for stochastic optimization," 2014.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[43] A. K. Kuchibhotla, "Exchangeability, conformal prediction, and rank tests," Jun. 2021.