Rapid Hyperspectral Classification of Camouflaged Objects in Complex Environments

Anonymous CVPR submission
Paper ID #*****

Abstract

Hyperspectral object classification has predominantly focused on satellite-based top-down imagery for large-scale land-cover classification. However, hyperspectral object classification in ground-based natural scenes remains largely underexplored due to the lack of publicly available datasets and the complexity of real-world environments. In this work, we construct a hyperspectral dataset simulating defense scenarios where hazardous targets may be concealed within complex environments such as vegetation, trees, and camouflage structures. The constructed dataset consists of hyperspectral imagery acquired in the Visible–VNIR spectral range (479–900 nm), enabling scene-level analysis of concealed objects in natural environments. Conventional hyperspectral classification models typically process hyperspectral cubes in a patch-wise manner. While effective for local feature extraction, this approach requires repeated patch inference to cover the entire scene, resulting in long inference times for scene-level analysis. To address this limitation, we propose a time-efficient inference framework for fast scene-level hyperspectral object classification. The proposed method enables the model to predict multiple pixels simultaneously from a single hyperspectral patch, significantly reducing the number of required inference operations. Experimental results demonstrate that the proposed method reduces inference time by 86.56% (7.44× speedup) while reliably classifying concealed hazardous objects under complex and adverse environmental conditions.

1. Introduction

Hyperspectral imaging (HSI) provides rich spectral information beyond conventional RGB imagery and has demonstrated significant potential in object classification tasks [1], [2]. Most existing hyperspectral classification studies have focused on satellite or airborne top-down imagery, where large and relatively homogeneous land regions are analyzed for land-cover classification and remote sensing applications [3], [4]. While such environments are well suited for large-scale terrain analysis, they do not fully reflect the complexity of real-world ground-level scenes.

In contrast, hyperspectral object classification at the scene level in ground-based natural environments remains relatively underexplored compared to satellite-based remote sensing studies. Representative public hyperspectral datasets, such as the Indian Pines and Pavia University datasets, were acquired from airborne or satellite platforms and are primarily designed for large-scale land-cover classification. These datasets typically exhibit relatively regular spatial structures and limited object-level complexity, and therefore do not adequately represent the challenges encountered in real-world ground-level scene understanding. Ground-based hyperspectral scenes differ significantly from satellite imagery as they include objects observed from various viewpoints and distances, with irregular object boundaries, occlusions, and complex background mixtures. Moreover, in real-world environments, the spectral signatures of materials can vary significantly due to illumination changes, atmospheric conditions, material mixing, and bidirectional reflectance distribution function (BRDF) effects [5]. These factors increase spectral variability within the same class while simultaneously reducing inter-class separability, making object classification considerably more challenging.

To better reflect these real-world conditions, we construct a ground-based scene-level hyperspectral dataset acquired in natural environments. The dataset consists of hyperspectral imagery captured in the visible to near-infrared (VNIR) spectral range (479–900 nm), and includes hazardous objects intentionally concealed within complex backgrounds such as vegetation, trees, and camouflage structures. This dataset enables the investigation of hyperspectral object classification under realistic concealment conditions encountered in defense environments. In defense environments, hazardous objects may be intentionally concealed by surrounding environmental elements such as vegetation, trees, or camouflage structures. As illustrated in Fig. 1, the target objects can become visually indistinguishable in RGB imagery. Moreover, the spectral signatures of the target objects may also be mixed with those of surround-

(a) (b) (c) (d)

Figure 1. Examples of concealed hazardous objects in natural environments. (a) Friendly and enemy personnel lying prone and concealed within vegetation. (b) A friendly soldier concealed behind a tree. (c) A mock tank camouflaged with grass to mimic the surrounding environment. (d) A mock tank concealed using a camouflage net. These examples highlight the difficulty of detecting hazardous objects when they are intentionally hidden within complex natural backgrounds.

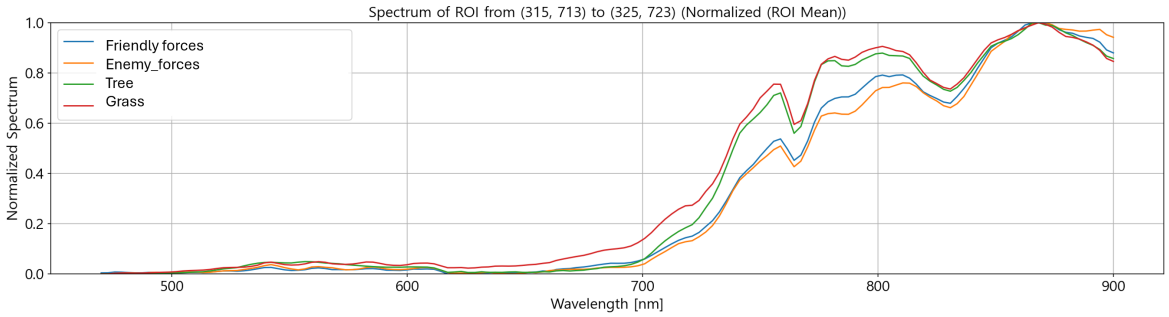

Figure 2. Example spectral signatures extracted from a hyperspectral scene. The spectra correspond to vegetation (grass and tree) and human subjects (friendly and enemy personnel). Although these objects belong to different semantic classes, their spectral curves exhibit very similar shapes across the VNIR range, making it difficult to distinguish them using simple spectral features alone.

ing background materials. Fig. 2 shows an example of spectral signatures extracted from a hyperspectral scene containing vegetation (grass and tree) and human subjects (friendly and enemy personnel). Although these objects belong to different semantic categories, their spectral curves exhibit very similar shapes across the VNIR spectral range. This spectral similarity makes it difficult to reliably distinguish objects based solely on simple spectral characteristics.

Consequently, effective object classification in such environments requires not only high spectral resolution to capture subtle spectral differences but also the ability to incorporate spatial context across the scene. Recently, CTMixer, a hybrid architecture combining convolutional neural networks (CNNs) and Transformers, has demonstrated strong performance in hyperspectral classification by jointly modeling local spatial features and global contextual information [6]. However, most hyperspectral models, including CTMixer, process hyperspectral cubes in a patch-wise manner. While patch-based processing alleviates memory constraints during training, it requires repeated inference over numerous patches when applied to scene-level data, resulting in substantial computational latency. In operational defense environments where real-time decision-making is critical, such inference delays become a significant limitation.

In this work, we propose a time-efficient hyperspectral object classification framework that enables efficient scene-level inference. Our method extends the CTMixer architecture to mitigate the computational bottleneck caused by patch-based processing and enables efficient scene-level prediction. Experimental results demonstrate that the proposed approach reduces inference time by 86.56% (7.44× speedup) while maintaining robust classification performance under complex and adverse environmental conditions.

The main contributions of this work are summarized as follows:

- We construct a scene-level hyperspectral dataset that includes hazardous objects concealed in complex natural environments.
- We propose a time-efficient hyperspectral object classification framework that significantly reduces inference latency.

2. Related Work

Deep learning has significantly advanced hyperspectral image (HSI) analysis in recent years [2]. Existing approaches can be broadly categorized into CNN-based methods, Transformer-based methods, and hybrid architectures that combine both paradigms.

Early studies mainly relied on convolutional neural networks (CNNs) to exploit spatial and spectral characteristics of hyperspectral data. Roy et al. [7] proposed a hybrid 2D–3D convolutional architecture (HybridSN) that jointly models spatial and spectral information for hyperspectral image classification. Similarly, Zhang et al. [8] introduced an encoder–decoder convolutional architecture that progressively reduces and restores spatial resolution to capture contextual information for hyperspectral object classification. Although CNN-based approaches are effective at extracting local spatial features, they have limited capability to model global spectral relationships across hyperspectral bands, which is an important characteristic of hyperspectral data.

More recently, Transformer-based models have been explored to capture long-range dependencies and spectral relationships in hyperspectral data. Hong et al. [9] proposed SpectralFormer, which represents spectral bands as token embeddings and models spectral correlations using a Transformer architecture. He et al. [10] further introduced HSI-BERT by adapting the BERT architecture from natural language processing to hyperspectral data for contextual spectral representation learning. Despite their strong capability in modeling long-range spectral dependencies, Transformer-based approaches are relatively less effective at capturing fine-grained local spatial structures, which remain important for hyperspectral object classification [6].

Hybrid architectures combining CNN and Transformer components have also been proposed to leverage both local spatial feature extraction and global contextual modeling. Chen et al. [6] proposed a hybrid CNN–Transformer architecture (CTMixer) to jointly capture spatial and spectral features in hyperspectral imagery. More recently, Wang et al. [11] introduced a hyperspectral foundation model (HyperSIGMA) pre-trained on large-scale datasets to improve generalization across various hyperspectral recognition tasks.

Despite these advances, most existing hyperspectral deep learning models rely on patch-based processing of hyperspectral cubes. While this strategy reduces memory consumption during training, it requires repeated patch inference when applied to high-resolution scenes, resulting in significant inference latency. This limitation becomes critical in scenarios where fast scene-level analysis is required. To address this issue, we propose a time-efficient hyperspectral object classification framework that enables efficient scene-level hyperspectral analysis.

3. Method

3.1. Overview

Hyperspectral image (HSI) classification methods commonly adopt patch-based processing, where a small hyperspectral cube centered at a target pixel is used as the input of a deep neural network. The network predicts the class label of the center pixel, and the process is repeated for all pixels in the image. Although this strategy improves classification accuracy by utilizing local spatial context, it leads to inefficient inference for high-resolution scenes since each pixel requires a separate patch inference. To address this limitation, we propose a multi-pixel inference strategy that enables the model to predict multiple pixels simultaneously from a single hyperspectral patch. Our framework is built upon CTMixer, a hybrid CNN–Transformer architecture designed for hyperspectral image classification. The proposed method modifies the prediction mechanism of CTMixer to generate predictions for multiple pixels within a patch, significantly reducing the number of required inference operations for scene-level hyperspectral object classification.

3.2. Baseline: CTMixer

CTMixer is a hybrid architecture that integrates convolutional neural networks and transformer encoders to jointly exploit spatial and spectral information in hyperspectral images.

As illustrated in Fig. 3, the network consists of three main components:

- a spectral–spatial feature extraction module,
- a dual-branch feature extraction structure composed of a transformer encoder branch and a CNN branch,
- a classification head.

First, the spectral–spatial features of hyperspectral patches are extracted by a convolutional preprocessing module. In this process, group-parallel residual blocks (GPRBs) are used to preserve the spectral-spatial structure information of the hyperspectral cube by simultaneously capturing spectral and spatial features through grouped convolution operations. Next, the extracted features are processed by a dual-branch architecture. The transformer encoder branch captures global contextual information by modeling long-range spectral and spatial dependencies using a self-attention mechanism, while the CNN branch focuses on modeling local spatial structures. The outputs from the two branches are fused through element-wise addition. Finally, global feature aggregation is performed using average pooling, and the final class label is predicted by a linear classifier.

Given an input hyperspectral patch $X \in \mathbb{R}^{C\times P\times P}$, where C denotes the number of spectral bands and P is the spatial patch size, the CTMixer model predicts the class label of the center pixel of the patch. For scene-level infer-

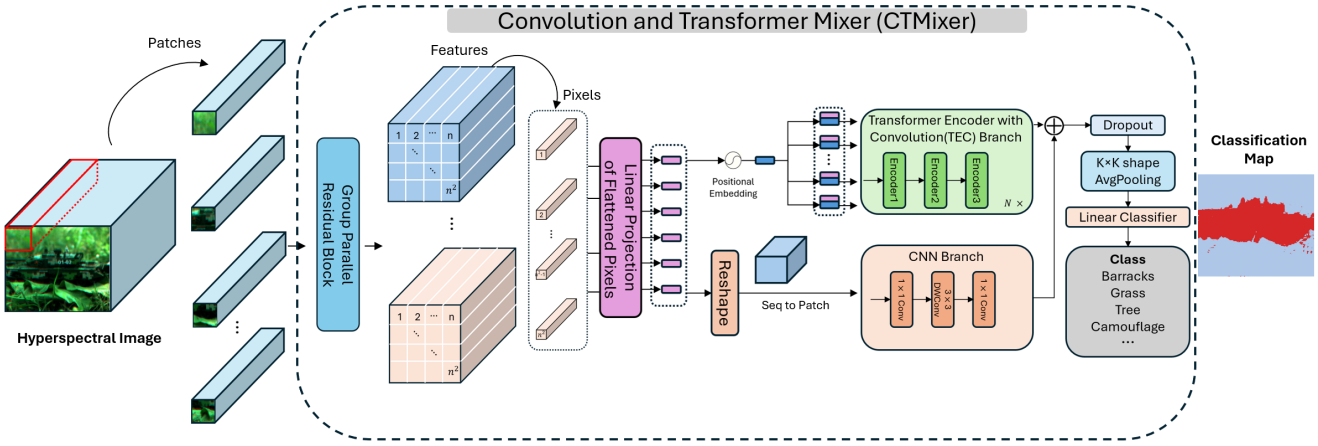

Figure 3. Framework of the proposed multi-pixel inference built upon CTMixer.

ence, a sliding-window strategy is typically adopted, where patches centered at each pixel location are sequentially extracted from the hyperspectral image and processed by the model.

3.3. Proposed Multi-Pixel Inference

In conventional patch-based hyperspectral classification frameworks, a single forward pass of the network produces the prediction of only the center pixel of the input patch. Therefore, to generate a full prediction map for an image with N pixels, the model must perform N forward passes.

To improve inference efficiency, we propose a multi-pixel prediction mechanism that enables the model to generate predictions for multiple pixels simultaneously from a single patch. Specifically, instead of predicting only the center pixel, the network is modified to predict a square region of pixels around the center. Let k denote the side length of the prediction region. The model therefore produces predictions for a $k \times k$ pixel region from a single input patch. Formally, for an input hyperspectral patch X, the prediction function can be expressed as

$$f(X) \rightarrow Y \in \mathbb{R}^{k \times k}.$$

Here, X denotes the input hyperspectral patch and Y represents the predicted labels for the $k \times k$ pixels in the center region of the patch. In this study, the spatial patch size is set to $P = 11$, and an 11×11 hyperspectral patch is used as the input to the model. Experiments are conducted using the 9×9 multi-pixel prediction configuration to evaluate the effectiveness of the proposed inference strategy. Let N denote the number of pixels in a hyperspectral image. In the conventional center-pixel prediction framework, the model must perform N forward passes to generate the full prediction map. In contrast, when predicting a $k \times k$ pixel region at once, the number of required forward passes becomes approximately

$$\frac{N}{k^2}.$$

Therefore, the proposed method reduces the number of inference operations by a factor of approximately k^2 compared with the conventional patch-based approach. This modification enables efficient scene-level hyperspectral object classification while maintaining the representation capability of the original CTMixer architecture.

4. Dataset

We collect ground-based hyperspectral scenes using an IMEC SNAPSCAN VNIR hyperspectral camera (IMEC). The camera covers the visible to near-infrared (VNIR) spectral range of 470–900 nm, and each scene consists of 150 spectral bands. The original frame resolution is 1088×2048. The dataset is designed to reflect realistic defense scenarios where hazardous targets may be intentionally concealed within complex natural backgrounds (e.g., vegetation, trees, and camouflage structures), which aligns with our goal of scene-level hyperspectral object classification.

Since the raw frames contain large background areas and targets often occupy only a small portion of the scene, we crop a region-of-interest (ROI) around the target area and use the cropped hyperspectral cubes for training and evaluation. In total, 106 ROI scenes are constructed. The average ROI resolution is 477.70×618.18 (H×W), and the ROI

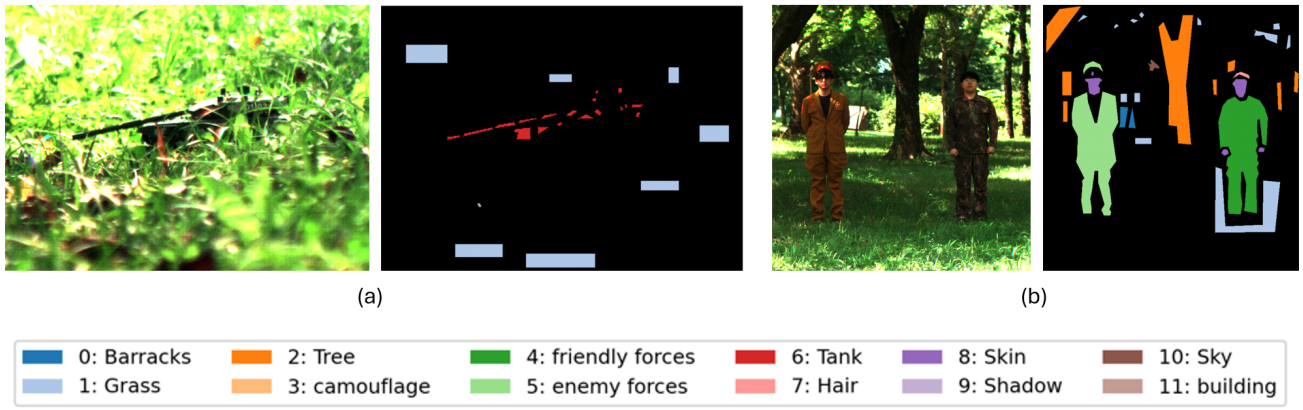

Figure 4. RGB reconstruction and pixel-wise annotation examples of our hyperspectral dataset. (a) Labeling example for a camouflaged tank scene. (b) Labeling example for friendly and enemy personnel in a ground scene.

width ranges from 194 to 1087 pixels. This ROI-based setting preserves realistic background mixtures and concealment effects around targets while enabling systematic analysis of inference efficiency in scene-level processing.

For annotation, we reconstruct RGB images from the hyperspectral cubes (Fig. 4) and perform pixel-wise semantic labeling for all pixels. We define 12 classes: {Barracks, Grass, Tree, Camouflage, Friendly forces, Enemy forces, Tank, Hair, Skin, Shadow, Sky, Building}. The primary targets of interest in this work are *Barracks*, *Camouflage*, *Friendly forces*, *Enemy forces*, and *Tank*, which represent objects commonly appearing in defense environments and can be visually indistinguishable or spectrally mixed with surrounding backgrounds. The remaining classes (e.g., Grass, Tree, Shadow, Sky) are included as auxiliary background categories to explicitly model complex scene composition and mixed pixels, thereby reducing confusion when identifying concealed targets.

5. Experiments and Results

5.1. Experimental Setup

All experiments were conducted on a workstation equipped with an AMD Ryzen Threadripper 3970X 32-Core CPU, an NVIDIA TITAN RTX GPU, and DDR4 memory (2666 MT/s). The proposed model was trained for 300 epochs using the Adam optimizer with a learning rate of 5×10^{-4}. Following a commonly used protocol in hyperspectral image classification, the labeled pixels within each scene were divided into training, validation, and test sets with ratios of 10%, 10%, and 80%, respectively.

5.2. Quantitative Results

Table 1 compares the classification performance and inference time between the conventional single-pixel prediction (1×1) and the proposed multi-pixel prediction (9×9).

Table 1. Comparison of inference performance between conventional and proposed methods.

Inference Size	OA	AA	Kappa	Time (sec)
1×1	0.9946	0.9512	0.9764	30.370
9×9	0.9968	0.9503	0.9787	4.084

As shown in Table 1, the proposed 9×9 multi-pixel inference achieves nearly identical classification performance compared with the conventional 1×1 inference while significantly reducing the processing time. Specifically, the proposed method reduces the average inference time per image from 30.370 seconds to 4.084 seconds, corresponding to approximately a $7.44\times$ speed improvement. Although additional prediction region sizes such as 3×3 and 5×5 were initially considered, they were omitted in this study due to time constraints. Since the 9×9 prediction already demonstrates comparable classification accuracy to the baseline while significantly improving inference speed, the intermediate configurations were not further evaluated.

5.3. Hard Case Visualization

Figure 5 visualizes several challenging scenarios in which hazardous objects are intentionally concealed within natural environments. The proposed hyperspectral approach successfully identifies targets such as prone personnel and camouflaged tanks even when the objects are visually difficult to distinguish in RGB imagery. For comparison, segmentation results obtained using an RGB-based model (SegNet) are also shown. [12] The RGB-based predictions exhibit significantly higher noise and fail to consistently detect the concealed objects. These results demonstrate that hyperspectral imagery provides valuable spectral information that enables more reliable object identification compared with conventional RGB imagery.

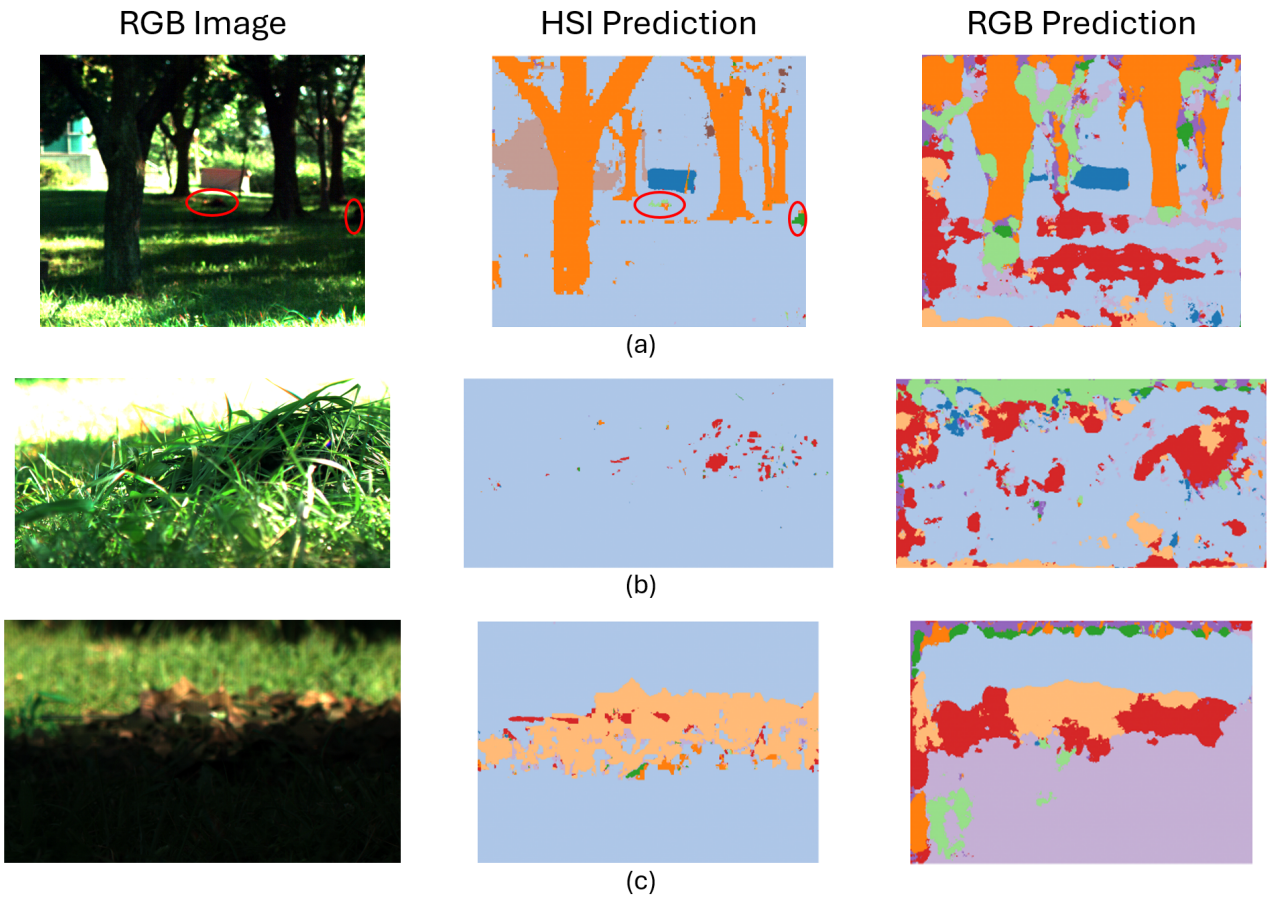

Figure 5. Visualization results under challenging concealment scenarios. From left to right: RGB image, hyperspectral prediction, and RGB-based prediction. (a) Friendly and enemy personnel lying prone on the ground. (b) A tank camouflaged with surrounding grass. (c) A tank concealed under a camouflage net. The hyperspectral prediction uses the proposed 9×9 multi-pixel inference based on CTMixer, which successfully identifies concealed objects even under difficult environmental conditions. In contrast, the RGB-based segmentation model (SegNet) produces significantly noisier predictions and fails to reliably detect concealed targets, highlighting the advantage of hyperspectral information for object identification in complex environments.

5.4. Spectral Importance Analysis

To further analyze how the model utilizes spectral information, we perform a spectral importance analysis using SHAP (SHapley Additive exPlanations). [13], [14]. Since the hyperspectral data contain 150 spectral bands, the bands were grouped into 15 spectral groups, each consisting of 10 adjacent bands. Figure 6 illustrates the SHAP importance values for several representative object classes. As observed earlier in Fig. 2, the spectral curves of different objects appear visually similar, making it difficult to distinguish them through simple visual inspection. However, the SHAP analysis reveals that CTMixer implicitly learns to emphasize different spectral regions when identifying different object classes. For example, certain band groups contribute more significantly to the classification of tanks, while others are more important for distinguishing vegetation or human subjects. This analysis indicates that the proposed model effectively exploits subtle spectral variations across the VNIR range to improve object discrimination in complex environments.

6. Conclusion

In this paper, we proposed an efficient hyperspectral object classification framework for identifying concealed objects in complex natural environments. In addition to the proposed classification framework, we constructed a ground-based scene-level hyperspectral dataset designed to simulate defense scenarios where hazardous objects are intentionally concealed within complex natural backgrounds such as vegetation, trees, and camouflage structures.

Conventional hyperspectral classification models typically adopt a patch-based center-pixel prediction strategy,

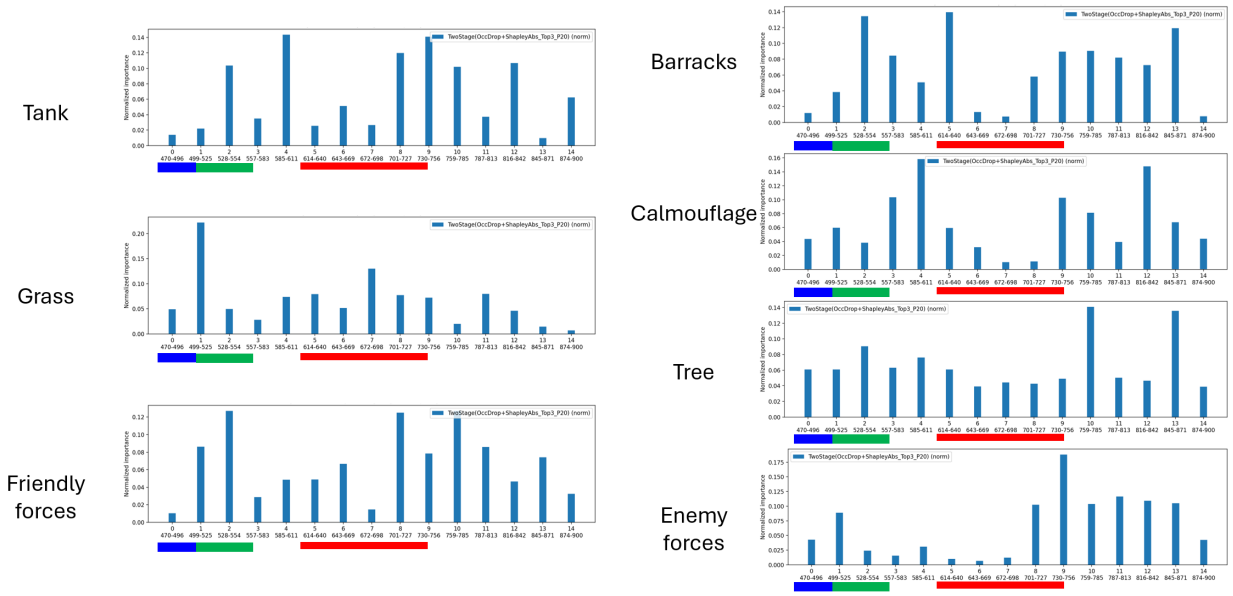

Figure 6. Spectral importance analysis using SHAP. The 150 spectral bands are grouped into 15 band groups (10 bands per group). The results show that CTMixer assigns different spectral importance depending on the target object class.

which requires a large number of repeated inference operations to generate a scene-level prediction map. This limitation results in significant computational overhead when processing high-resolution hyperspectral scenes. To address this issue, we introduced a multi-pixel inference strategy built upon the CTMixer architecture. The proposed method enables the model to predict multiple pixels simultaneously from a single hyperspectral patch. Experimental results demonstrate that the proposed approach achieves approximately 7.44× faster inference compared with the conventional 1×1 center-pixel prediction strategy while maintaining comparable classification performance in terms of OA, AA, and Kappa metrics.

Qualitative visualization results further show that hyperspectral imagery provides significant advantages in challenging scenarios where objects are intentionally concealed within complex natural backgrounds. In particular, the RGB-based segmentation model (SegNet) produces noisy predictions and struggles to reliably identify concealed objects, whereas the hyperspectral model successfully detects targets by leveraging spectral information beyond the RGB spectrum. Furthermore, the spectral importance analysis using SHAP reveals that the CTMixer model implicitly assigns different importance to specific spectral regions when distinguishing different object classes. This indicates that the proposed model effectively exploits subtle spectral variations across the VNIR range to improve object discrimination.

Overall, this work contributes both a newly constructed hyperspectral dataset for ground-based concealed-object scenarios and an efficient inference framework for scene-level hyperspectral classification. These contributions provide a practical foundation for future research on hyperspectral object recognition in complex real-world environments. Future work will explore larger-scale hyperspectral datasets and investigate the integration of hyperspectral foundation models to further enhance classification performance and generalization capability.

References

[1] Alexander F. H. Goetz, Gregg Vane, Jerry E. Solomon, and Barrett N. Rock. Imaging spectrometry for earth remote sensing. *Science*, 228(4704):1147–1153, 1985. 1

[2] Wei Li, Hantao Fu, Le Yu, and Arthur Cracknell. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019. 1, 3

[3] José M. Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser M. Nasrabadi, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, 2012. 1

[4] Antonio Plaza, Jon Atli Benediktsson, James W. Boardman, et al. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009. 1

[5] Ben Somers, Gregory P. Asner, Laurent Tits, and Pol Cop-

pin. Endmember variability in spectral mixture analysis: A review. *Remote Sensing of Environment*, 115(7):1603–1616, 2011. 1

[6] Junjie Zhang, Yao Lu, Xiaoqian Zhang, et al. Convolution transformer mixer for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 2, 3

[7] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri. Hybridsn: Exploring 3d–2d cnn feature hierarchy for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 17(2):277–281, 2019. 3

[8] Liangpei Zhang, Lefei Zhang, and Bo Du. Spectral–spatial deep guided learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):1–14, 2020. 3

[9] Danfeng Hong, Zhu Han, Jing Yao, and Xiao Xiang Zhu. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60(1):1–15, 2022. 3

[10] Xiangyu He, Yushi Chen, and Liangpei Zhang. Hsi-bert: Hyperspectral image classification using bidirectional encoder representation from transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 61(1):1–14, 2023. 3

[11] Jia Wang, Rui Zhao, and Xin Huang. Hypersigma: A foundation model for hyperspectral image understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 62(1):1–15, 2024. 3

[12] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 5

[13] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 6

[14] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *IEEE Signal Processing Magazine*, 38(3):56–67, 2021. 6