

INVESTIGATION OF TARGET DETECTION IN HYPERSPECTRAL IMAGES FOR
AIDED TARGET RECOGNITION

by

Mazharul Hossain

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Computer Science

The University of Memphis

May 2026

Copyright© Mazharul Hossain
All rights reserved

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Lan Wang, for her invaluable guidance, continuous support, and mentorship throughout my doctoral studies. Her insight, encouragement, and patience have been instrumental in shaping this work.

I am sincerely grateful to my co-advisor, Dr. Aaron L. Robinson, and my dissertation committee members, Dr. Chrysanthé Preza, Dr. Deepak Venugopal, and Dr. Xiaolei Huang, for their thoughtful feedback and constructive suggestions, which significantly improved this research.

I also acknowledge Dr. Eddie Jacobs for providing financial support and research opportunities through CAESER and the "Multi-UAS Multi-Sensor Intelligence, Surveillance, and Reconnaissance" project, which made this work possible.

I appreciate the iTiger GPU cluster [1], supported by the NSF MRI program (CNS-2318210), for providing the computing resources. Additionally, I am grateful to all the authors of open-source projects for their dedication and contributions.

I thank my colleagues, especially Mohammed Younis, and my labmates in the Networking Research Lab (NetLab), for their collaboration, insightful discussions, and support throughout this journey.

I am especially grateful to my wife and daughter for their patience, understanding, and unwavering support throughout the many long nights and demanding periods of this journey. Their encouragement has been my constant source of motivation and strength.

Finally, I am deeply grateful to my family and friends for their unwavering support and encouragement.

Funding statement: This material is based upon work supported by the U.S. Army Contracting Command under Contract No. W911QX-24-C-0011. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Army Contracting Command.

Abstract

Aided/Automatic Target Recognition (AiTR/ATR) enhances human decision-making by improving the detection, localization, and interpretation of targets in sensor imagery. Hyperspectral (HS) imaging provides rich spectral information for AiTR that can significantly improve target detection; however, HS detectors often perform unreliably in real-world conditions due to high data dimensionality, limited labeled training data, class imbalance, sensitivity to environmental and sensor variations, and a shortage of pre-trained models. These challenges frequently lead to degraded performance when models trained in one setting are deployed in different scenes, sensors, or operating conditions.

The overarching goal of this dissertation is to make hyperspectral target detection for AiTR systems reliable under real-world conditions. This work focuses on improving robustness to sensor variability, environmental changes, and limited labeled data while maintaining computational efficiency for deployment on resource-constrained airborne platforms such as unmanned aerial vehicles (UAVs). To address these challenges, the dissertation improves the quality of HS data, combines complementary detectors via ensemble learning, incorporates physically interpretable spectral features, and develops domain adaptation techniques to maintain performance across different sensors and scenes. Finally, the work extends beyond anomaly detection for target localization to downstream scene understanding through hyperspectral semantic segmentation to support AiTR interrogation systems.

This dissertation first examines the importance of HS image processing and enhancement techniques for enhancing data quality, improving feature extraction, and interpretability for both humans and machine learning models. Furthermore, this dissertation investigates robust anomaly and target detection in airborne HS imagery, with an emphasis on unsupervised and weakly supervised learning paradigms suitable for label-scarce operational settings. Individual detectors often have limitations, as the backgrounds in which they operate may not meet their underlying assumptions and can be cluttered. To overcome these challenges, a Greedy Ensemble Anomaly Detection (GE-AD) framework is proposed that systematically selects and weights heterogeneous

statistical and machine-learning-based detectors via supervised and unsupervised stacking.

To enhance robustness and interpretability, physically grounded spectral features derived from spectral unmixing are integrated into GE-AD. The resulting ensemble methods are designed to be lightweight and CPU-efficient, enabling deployment on resource-constrained unmanned aerial systems while maintaining competitive detection performance.

To improve generalization under domain shift, the dissertation further explores transfer learning and unsupervised domain adaptation (DA) within a unified MoE stacking ensemble architecture. These approaches mitigate performance degradation caused by sensor variability, environmental changes, and acquisition inconsistencies without requiring retraining on target-domain data. Finally, the work extends beyond anomaly detection or target localization to broader scene understanding through HS semantic segmentation. A channel-attention-based adaptation of existing deep segmentation models enables effective processing of high-dimensional HS data with minimal architectural modifications, as demonstrated by agricultural crop-mapping case studies.

Taken together, this dissertation demonstrates that robust hyperspectral target detection under real operational variability can be achieved by combining improved hyperspectral data processing, ensemble anomaly detection, physically interpretable spectral features, and adaptive generalization strategies while preserving computational efficiency for practical AiTR deployment.

Keywords: hyperspectral imaging, remote sensing, anomaly detection, ensemble learning, domain adaptation, spectral unmixing, near-infrared (NIR), unmanned aerial vehicles (UAVs), aided target recognition (AiTR), semantic segmentation, transfer learning

Table of Contents

Chapter	Page
List of Tables	x
List of Figures	xii
1 Introduction	1
1.1 Hyperspectral Imaging and Anomaly Detection in ATR Systems	3
1.2 Hyperspectral Image Semantic Segmentation in ATR Systems	6
1.3 Challenges in Robust Hyperspectral Target Detection	7
1.3.1 High Dimensionality and Background Variability	7
1.3.2 Class Imbalance and Evaluation Instability	7
1.3.3 Domain Shift Across Scenes and Sensors	7
1.4 Research Objectives and Questions	8
1.5 Scope and Organization of the Dissertation	9
1.6 Contributions	11
2 Literature Review and Conceptual Framework	14
2.1 Hyperspectral Image Preprocessing	14
2.1.1 Initial HS Image Preprocessing	15
2.1.2 Data Normalization for Machine Learning	16
2.1.3 Image Processing Algorithms for Visual Enhancement	17
2.1.4 Image Quality Evaluation Metrics	17
2.2 Classical Hyperspectral Target and Anomaly Detection Methods	18
2.3 Hyperspectral Unmixing for Background and Target Modeling	21
2.3.1 Role of Abundance Information in Anomaly Detection	22
2.4 Ensemble Learning for Hyperspectral Anomaly Detection	24
2.4.1 Stacking and Meta-Learning Strategies	27
2.5 Hyperspectral Unmixing-Based Voting Ensemble Anomaly Detector (HUE-AD)	28
2.6 Hyperspectral Anomaly Detection Datasets	29
2.7 Performance Evaluation under Class Imbalance in HS-AD	31
2.7.1 Limitations of ROC-AUC in Hyperspectral Detection	33
2.7.2 F1-Based Metrics and Thresholding Strategies	34
2.8 Domain Shift, Generalization, and Transfer Learning in Hyperspectral Imaging	35
2.8.1 Cross-Scene and Cross-Sensor Variability	35
2.8.2 Domain Adaptation and Representation Alignment	36
2.8.3 Test-time Domain Adaptation	37
2.9 Mixture-of-Experts Architectures	39
2.10 Hyperspectral Semantic Segmentation	39
2.11 Hyperspectral Semantic Segmentation Dataset	42
2.12 Performance Evaluation in HS-SS	43

3	Hyperspectral Image Preprocessing	45
3.1	Process HS images to Improve Machine Detection Accuracy	45
3.1.1	Preprocessing Methods to Improve Hyperspectral AD Algorithms for Arizona Dataset	45
3.1.2	Removal of Noisy Bands	46
3.1.3	Normalization of Anomaly Detection Results	48
3.1.4	Unmixing-based Anomaly Detection	48
3.1.5	Discussion and Summary	51
3.2	Image Enhancement Methods for Human Understandability	52
3.2.1	Motivation	53
3.2.2	Dataset Generation	53
3.2.3	Methodological Design Choice	55
3.2.4	Proposed Method	55
3.2.5	Evaluation and Results	57
3.2.6	Discussion and Summary	58
4	Ensemble-Based Hyperspectral Anomaly Detection	59
4.1	Background	59
4.2	Motivation	60
4.3	Greedy Ensemble Hyperspectral Anomaly Detection (GE-AD)	61
4.3.1	Ensemble Learning Formulation	61
4.3.2	Weighted Stacking with Supervised Meta-Models	62
4.3.3	Greedy Base-Model Selection Strategy	62
4.3.4	Computational Complexity Analysis	65
4.4	Unsupervised Extension of GE-AD (UGE-AD)	66
4.5	Incorporating Spectral Information into GE-AD (mGE-AD)	67
4.6	Evaluation of the Proposed GE-AD Framework	67
4.6.1	Comparison with Individual AD Methods	68
4.6.2	Comparison with Baseline SOTA Ensemble Methods	72
4.6.3	Public Dataset Benchmark Evaluation: Hydice Urban	73
4.6.4	Public Dataset Benchmark Evaluation: Salinas	73
4.6.5	Public Dataset Benchmark Evaluation: San-Diego Airport	75
4.6.6	Private Dataset Benchmark Evaluation: Arizona	75
4.6.7	Discussion	76
4.7	Evaluation of the Unsupervised GE-AD Extension	79
4.7.1	Evaluation Protocol and Metrics	80
4.7.2	Single-Dataset Greedy Search Evaluation	80
4.7.3	Multi-Dataset Greedy Search Evaluation	80
4.7.4	Per-Dataset Quantitative Results	81
4.7.5	Discussion	84
4.8	Summary	86

5	Generalization of Hyperspectral Anomaly detection	89
5.1	Limitations of Classical Ensemble Methods	89
5.1.1	Cross-Scene Transfer: ABU-Airport to San-Diego	90
5.1.2	Cross-Sensor Transfer: ABU-Airport to Arizona	90
5.1.3	Problem Formulation	92
5.2	Unified MoE Ensemble framework for HS-AD	92
5.2.1	System Overview	92
5.2.2	Foundation Model as Backbone	95
5.2.3	Experts	95
5.2.4	Mixture-of-Experts (MoE) Based Framework	96
5.2.5	Weighted Prediction Fusion	98
5.2.6	Training-time Domain Adaptation and Generalization	99
5.2.7	Training	100
5.3	Test-time Domain Adaptation	103
5.3.1	Drift-Aware Adaptation Policy	105
5.4	Results	106
5.4.1	Leave-One-Out Evaluation	110
5.4.2	Results from Base Methods	111
5.4.3	Results from Individual GE-AD Methods	112
5.4.4	Ablation Test	114
5.5	Discussion	115
5.5.1	Impact of MoE on Anomaly Detection Metrics	116
5.5.2	Experiment Leave-One-Out	116
5.5.3	Train-time adaptation	116
5.6	Summary	117
6	Hyperspectral Image Semantic Segmentation for Aided Target Recognition	118
6.1	Motivation for Target-Focused Hyperspectral Segmentation	118
6.2	Task-Adapted Deep Learning for Hyperspectral Crop Mapping	119
6.3	Methodology	120
6.3.1	Using Residual Blocks from ResNetv2 in Contraction Path	121
6.3.2	Using Attention in Residual Blocks of ResNetv2	123
6.3.3	Transformer as Encoder Bottleneck	124
6.3.4	Expansion Path	124
6.4	Experimental Setup	125
6.4.1	Loss Function	125
6.4.2	Training Details	126
6.5	Experimental Results on UAV Hyperspectral Imagery	127
6.5.1	Qualitative Evaluation	129
6.6	Computational Considerations and Practical Implications	130
6.6.1	Model Complexity and Memory Footprint	130
6.6.2	Architectural Efficiency under Hyperspectral Constraints	131
6.6.3	Transfer Learning and Weight Reuse	131
6.6.4	Implications for Real-World Deployment	132
6.7	Summary and Integration with Detection Pipelines	132

7 Discussion, Conclusions, and Future Research	134
References	139

List of Tables

Table		Page
1	Examples from various types of HS-AD algorithms.	18
2	Feature comparison of ensemble learning algorithms.	25
3	Quantitative performance of the proposed model on overall degradation classification.	57
4	Quantitative performance of the proposed model on overall enhancement method classification.	58
5	Comparison of F1-macro scores showing our proposed ensemble method (GE-AD) performs better compared to HUE-AD, and individual methods using the ABU dataset.	71
6	Performance results of all methods regarding AUC metric (%) along with Friedman rank using the ABU-airport dataset.	71
7	ABU-IV quantitative performance (as reported in the authors' papers [84, 86, 87]) comparison between our proposed ensemble method (GE-AD) and other ensemble methods. Here, the best scores are shown in bold.	73
8	Comparison between our proposed ensemble method (GE-AD) and individual methods using the HYDICE-Urban dataset.	74
9	Comparison between our proposed ensemble method (GE-AD) and individual methods using the Salinas dataset.	74
10	Comparison between our proposed ensemble method (GE-AD) and individual methods using the San-Diego Airport dataset.	75
11	Comparison between our proposed ensemble method (GE-AD) and individual methods using the Arizona dataset.	77
12	Average ROC-AUC scores of our newly proposed Unsupervised ensemble UGE-AD over various datasets. The first column shows the dataset used in the greedy search to find suitable base methods, and the second column shows those base methods. Where the dataset names in rows and columns match, those cells show the training score after the greedy search.	81

13	Average and standard deviation of ROC-AUC scores of various methods computed over the ABU, Arizona, HYDICE-Urban, Salinas, and San-Diego datasets.	82
14	Average and standard deviation of F1-macro scores of our ensemble methods computed over the ABU, Arizona, HYDICE-Urban, Salinas, and San-Diego datasets.	83
15	Comparison of ROC-AUC scores between various methods using the HYDICE-Urban, Salinas, and San-Diego datasets	84
16	Comparison of ROC-AUC scores between various methods using the ABU-Airport dataset	85
17	Comparison of ROC-AUC scores between various methods using the ABU-Beach dataset	86
18	Comparison of ROC-AUC scores between various methods using the ABU-Urban dataset	87
19	Comparison of ROC-AUC scores between various methods using the Arizona dataset	88
20	Quantitative performance of the proposed model across all datasets (mean $\pm$ standard deviation over 5 runs with different random seeds).	107
21	Performance of our new model, where we excluded one dataset completely during training. The Dataset column shows the excluded dataset.	111
22	ROC-AUC performance comparison between the proposed method and base HS-AD and unmixing-AD methods	112
23	Performance of GE-AD models trained and evaluated on the same dataset (for completeness, we report standard classification metrics for the single-domain setting, where a fixed threshold is well-defined).	113
24	Performance of our new model without the domain classifier head and with the domain classifier head but without the Gradient Reversal Layer, averaged among all datasets.	115
25	Performance comparison of HRViTUNet using segmentation metrics. Bold marks the better result than other models for that metric.	128
26	Performance comparison of HRViTUNet using classification metrics. Bold marks the better result than other models for that metric.	128

List of Figures

Figure		Page
1	Generic algorithm processing pipeline for an AiTR system [2].	1
2	Example of plant diseases and pests detection problem [7].	2
3	AiTR is helping to identify enemy vehicles in military applications.	3
4	The Electromagnetic spectrum [12].	4
5	A generic scheme of HS imager mapping a scene identifying soil, vegetation, and water through reflectance signature [5].	4
6	Abundance map for Indian Pines.	5
7	(a) Endmembers extracted from the ABU-Airport-I dataset. (b) The complete process for obtaining target abundance using a target endmember [30]. A target endmember is a reference spectrum extracted from a reference HS image. The process of extracting target abundance is similar to object detection.	22
8	The architecture of a stacked model, where the new training data consist of the outputs of the first-level models.	28
9	Design for Hyperspectral Unmixing-Based Voting Ensemble Anomaly Detector (HUE-AD).	29
10	Overview of the TransUNet framework [43]: (a) schematic of the Transformer layer; (b) architecture of the TransUNet.	41
11	A block diagram of the Attention U-Net [140] segmentation model. Input image is progressively filtered and downsampled by a factor of 2 at each scale in the encoding part of the network (e.g., $H_4 = H_1 / 8$). N_c denotes the number of classes. Attention gates (AGs) filter the features propagated through the skip connections.	42
12	Channel Attention Module [143].	42
13	SNR in [dB] showing the impact of removing noisy bands and other denoising techniques from Image V in the Arizona dataset.	46
14	Visual comparison showing quantitative improvement of removing noisy bands from Florida Image 1. The green line shows SNR in [dB] after deleting noisy bands, whereas the blue line shows the dropped channels	47

15	Irregular results from kernel isolation forest anomaly detection are transformed into a normal distribution. (a) Distribution of Original data. (b) Irregularities. (c) Transformed normal distribution.	48
16	Normalization can give a boost in performance for tricky data like ABU-IV. (a) RGB, (b) ground-truth, (c) GE-AD with no normalization (F1 = 0.727), (d) GE-AD with z score normalization (F1 = 0.744), (e) GE-AD with quantile normalization (F1 = 0.757).	49
17	Example of digital image processing. From left to right: original image, auto-brightness adjustment (SSIM 0.685), and HSV CLAHE (SSIM 0.677).	52
18	Previously proposed enhancement method selection framework.	53
19	Image Enhancement Method Selection using ML Model	56
20	Overall flowchart of GE-AD to find the best AD methods for ensemble fusion.	63
21	Flowchart of the greedy search to find the best AD methods for ensemble fusion.	64
22	Visual comparison showing better performance of our proposed ensemble method (GE-AD) compared to other methods using ABU-IV data. (a) RGB, (b) ground-truth, (c) GE-AD (F1 = 0.856), (d) HUE-AD (F1 = 0.649), (e) Abundance (F1 = 0.735), (f) AED (F1 = 0.610), (g) KIFD (F1 = 0.587), (h) KRX (F1 = 0.604), (i) LSUNRSORAD (F1 = 0.569), (j) FCBAD (F1 = 0.631), (k) ERRX MF (F1 = 0.666), (l) ERCRD (F1 = 0.545), and (m) SED (F1 = 0.592). Please note that, as the implementations of algorithms for Figures (k–m) were not publicly available, we binarized the original outputs from the authors’ papers [84, 86, 87], respectively.	69
23	Detection accuracy evaluation through ROC curve and AUC scores (shown in the legend) of (a) ABU-I, (b) ABU-II, (c) ABU-III, and (d) ABU-IV.	72
24	Visual comparison showing better performance of our proposed ensemble method (GE-AD) compared to other methods using HYDICE-Urban data. (a) RGB, (b) ground-truth, (c) GE-AD (F1 = 0.944), (d) FCBAD (F1 = 0.575), and (e) LSUNRSORAD (F1 = 0.567).	74
25	Visual comparison showing better performance of our proposed ensemble method (GE-AD) compared to other methods using Arizona image-III data. (a) RGB, (b) ground-truth, (c) GE-AD (F1 = 0.858), (d) FCBAD (F1 = 0.773), (e) KIFD (F1 = 0.583), (f) KRX (F1 = 0.494), and (g) LSUNRSORAD (F1 = 0.639).	76

26	Average ROC-AUC scores of our UGE-AD over various datasets showing consistent performance against various training datasets. The X-axis shows the dataset used in the greedy search to find suitable base methods. The Y-axis shows the testing ROC AUC scores. Except where the dataset names in rows and columns match, those bars show the training score after the greedy search.	83
27	The visualization shows the generalization of our proposed ensemble method (GE-AD) trained on the ABU dataset compared to other methods using San-Diego-02 data. (a) RGB, (b) ground-truth, (c) GE-AD (F1 = 0.826), (d) Abundance (F1 = 0.655), (e) AED (F1 = 0.636), (f) FCBAD (F1 = 0.536), and (g) KRX (F1 = 0.627).	90
28	The visualization shows the performance of our proposed ensemble method (GE-AD) trained on the ABU dataset compared to other methods using Arizona-V data. (a) RGB, (b) ground-truth, (c) GE-AD (F1 = 0.490), (d) Abundance (F1 = 0.701), (e) AED (F1 = 0.584), (f) FCBAD (F1 = 0.810), (g) KRX (F1 = 0.503).	91
29	Visualization of difference of feature importance between models.	91
30	The overall system design of the proposed Generalized Anomaly Detector consists of the following components: (a) base methods that produce initial anomaly detection maps; (b) HyperSL embedding including Universal Spectral Tokenizer and Spectral Wavelength Positional Encoding; (c) HyperSL encoder, which, together with (b), forms the unified feature-extraction backbone; (d) a group of FFNN-AD experts; (e) soft-gate network that predicts expert weights from latent features; (f) weighted fusion module to produce the final anomaly prediction; (g) Gradient Reversal Layer (GRL); and (h) domain classifier, together with (g), enforces domain-invariant feature learning during training (Note that the dashed connections indicate training-only paths. The GRL (g) and domain classifier (h) are removed at inference.).	93
31	Outline of our proposed FFNN-AD expert	96
32	Outline of our proposed Soft-Gate Network	97
33	Qualitative results across datasets. Rows (top to bottom): HYDICE-Urban (F1=0.958), Salinas (F1=0.933), San-Diego image 02 (F1=0.913), ABU-Airport image 02 (F1=0.819), ABU-Beach image 01 (F1=0.976), ABU-Urban image 02 (F1=0.942), Arizona image 04 (F1=0.923). Columns (left to right): RGB image, ground-truth mask, anomaly score map, and binary detection result.	108
33	Qualitative results across datasets (continued).	109

34	Comparison of cross-domain F1-macro performance. Rows correspond to models trained on a single dataset (rows 1–7) and the proposed mixture-of-experts model (last row), while columns denote the evaluation dataset. Single-domain GE-AD models exhibit strong performance primarily along the diagonal (in-domain) but degrade significantly on unseen datasets (off-diagonal). In contrast, the proposed model maintains consistently higher and more stable performance across all domains, highlighting its ability to generalize under domain shift.	114
35	Outline of our proposed segmentation network.	121
36	ResNetV2 block with Pre-Activation Bottleneck, Attention mechanism, and Residual block.	122
37	Adaptive Channel Attention Mechanism	123
38	Design of our residual ViT	124
39	The visualization shows an example of semantic segmentation from our model using an image from UAV-HSI-CropDataset. (a) RGB, (b) ground-truth, (c) Our HRViTUNet fine-tuned on this agricultural HSI dataset, (d) TransUNet fine-tuned on RGB dataset, (e) TransUNet fine-tuned on HSI dataset.	127
40	Confusion-Matrix from one of the runs shows low missing classes.	129
41	The visualization of our semantic segmentation model’s predictions using images from UAV-HSI-CropDataset. In each section: (a) RGB, (b) ground-truth, (c) Our HRViTUNet’s prediction.	130

Chapter 1

Introduction

Aided/Automatic Target Recognition (AiTR/ATR) refers to the use of advanced sensing and computational systems to assist in identifying and classifying objects or targets, typically in military, security, environmental, and agricultural contexts. AiTR systems enhance target recognition and tracking by integrating multi-modal sensor data with automated processing, computer vision, and machine learning (ML) algorithms (see Fig. 1) [2, 3]. By reducing operator workload and human error, AiTR enables faster and more reliable decision-making in time-critical scenarios.

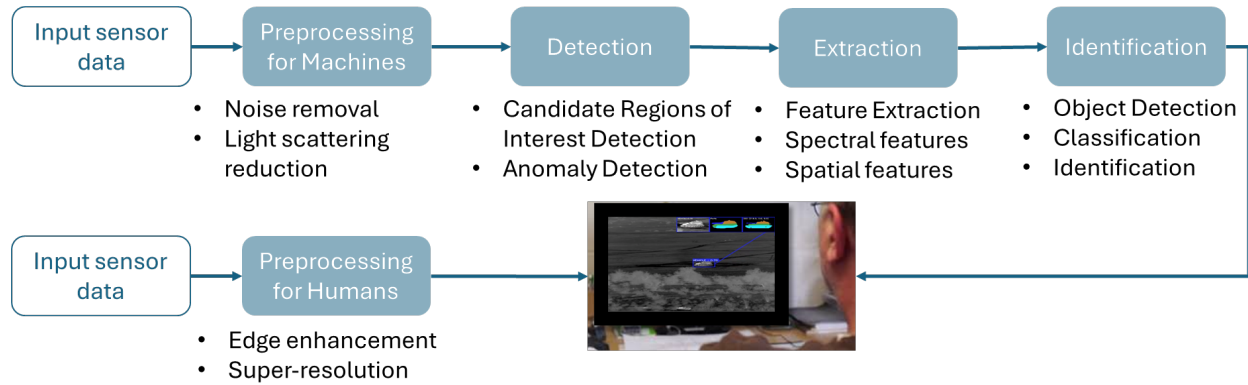


Figure 1: Generic algorithm processing pipeline for an AiTR system [2].

AiTR systems increasingly rely on data collected from diverse sensing platforms, including satellites, ground-based sensors, and unmanned aerial vehicles (UAVs). Emerging UAV and broader unmanned aircraft system (UAS) platforms equipped with various sensors provide agile, close-range sensing capabilities to enhance Intelligence, Surveillance, and Reconnaissance (ISR), agriculture, and environmental assessment applications, but introduce strict constraints on onboard computation, energy consumption, and real-time processing [4]. These constraints motivate the development of efficient, robust algorithms that can operate reliably under limited resources.

Hyperspectral (HS) imaging plays a central role in modern AiTR by enabling

material-level scene understanding that is not achievable with conventional RGB or multispectral sensors. By capturing hundreds of contiguous spectral bands, HS imaging allows fine-grained discrimination of objects, materials, and anomalies based on their spectral signatures [5, 6]. This capability supports a wide range of applications, including camouflaged target detection in military environments, crop health monitoring and pest detection in agriculture (Fig. 2), environmental monitoring, and mineral exploration.

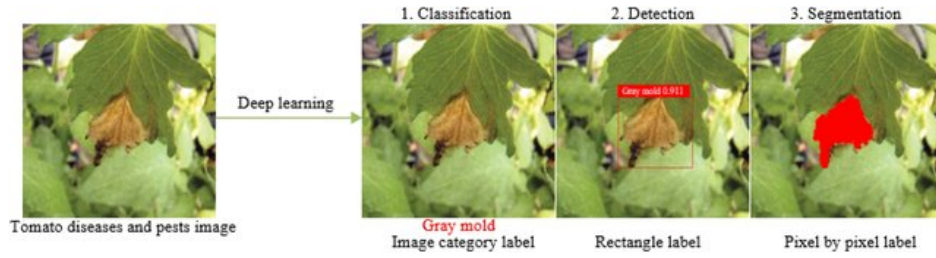


Figure 2: Example of plant diseases and pests detection problem [7].

Despite the advantages of HS imaging, it poses unique and nontrivial challenges for AiTR systems. Unlike conventional imagery, hyperspectral data amplify *domain shift* effects due to their high dimensionality and sensitivity to physical acquisition conditions, making even minor environmental changes statistically significant for detection algorithms. This high dimensionality also increases computational cost and exacerbates the curse of dimensionality, particularly for covariance-based and statistical detection methods. Moreover, sensor variability, atmospheric effects, illumination changes, and differences in acquisition protocols across platforms and missions further intensify domain shift, degrading detection and classification performance when models are deployed outside their training conditions.

In operational AiTR pipelines, HS-based target detection must satisfy competing requirements: a low false alarm rate (FAR) to prevent operator overload [8, 9], and a high detection rate (i.e., recall) to ensure that potential threats are not missed [10]. Although these objectives are inversely related, AiTR systems often prioritize detection performance, relying on subsequent analysis or human-in-the-loop verification to resolve false alarms. Figure 3 illustrates a military scenario in which HS-enabled AiTR assists operators in identifying vehicles under

challenging sensing conditions.

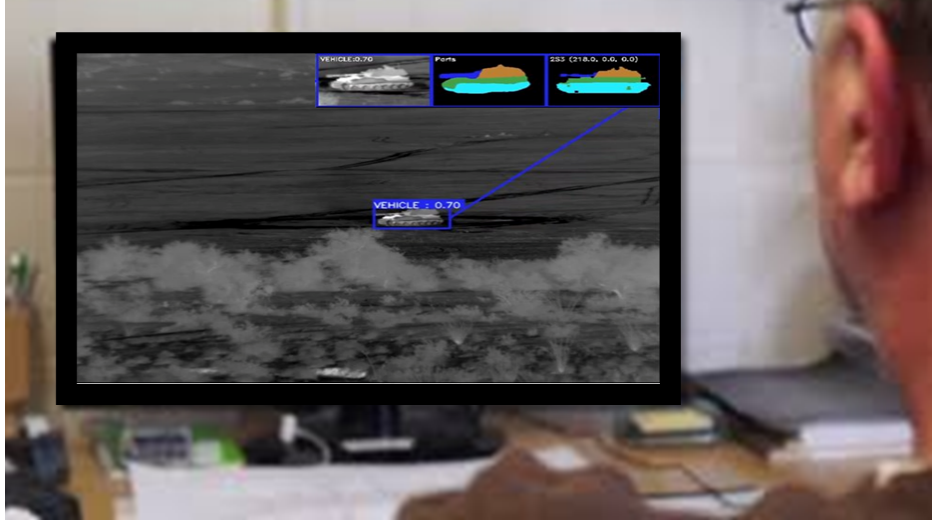


Figure 3: AiTR is helping to identify enemy vehicles in military applications.

In this dissertation, we focus on HS data collected from airborne platforms and investigate anomaly detection at the detection stage and classification and segmentation at subsequent stages of the AiTR pipeline. Particular emphasis is placed on understanding how task-oriented image preprocessing, statistical modeling, and machine learning affect both detection performance and human interpretability under real-world deployment constraints.

1.1 Hyperspectral Imaging and Anomaly Detection in ATR Systems

Hyperspectral imaging (HSI) combines spectroscopy with imaging to capture spatial information alongside dense, contiguous spectral measurements across the electromagnetic (EM) spectrum (see Fig. 4). Whereas conventional RGB imagery encodes scenes using only three broad bands in the visible range (380 nm to 700 nm), hyperspectral sensors acquire hundreds of narrow bands spanning ultraviolet (from 350 nm) through shortwave infrared wavelengths (up to 2500 nm) [5, 11]. This dense spectral sampling enables fine-grained characterization of material reflectance properties, allowing each pixel to exhibit a distinctive spectral signature.

Because different materials absorb, reflect, and transmit incident radiation in characteristic ways, HSI supports material discrimination, object identification, and the detection

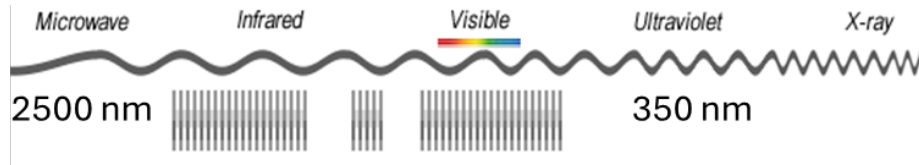


Figure 4: The Electromagnetic spectrum [12].

of spectrally anomalous targets that deviate from their surrounding background [5, 13]. Figure 5 illustrates a typical hyperspectral image cube and representative reflectance signatures for common materials such as vegetation, soil, and water. These properties make HSI particularly well suited for AiTR applications involving camouflaged targets, small manmade objects, or subtle material changes that are difficult to detect using conventional imagery.

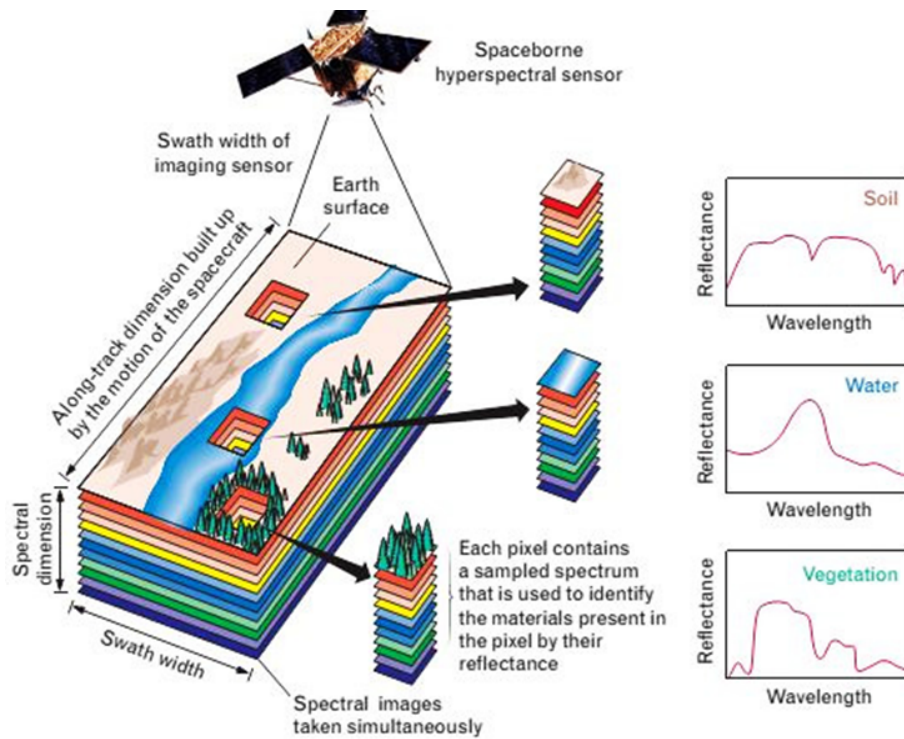


Figure 5: A generic scheme of HS imager mapping a scene identifying soil, vegetation, and water through reflectance signature [5].

However, the same characteristics that make HSI powerful also introduce significant challenges. Hyperspectral data are high-dimensional, sensitive to atmospheric effects, and

influenced by illumination conditions, sensor characteristics, and acquisition geometry. Weather conditions, time of day, cloud cover, and atmospheric absorption can degrade data quality, while sensor-to-sensor variability introduces inconsistencies across datasets. As a result, various preprocessing steps, such as noise reduction, contrast enhancement, and spectral normalization, are often required to improve machine performance and human interpretability.

A foundational concept in hyperspectral analysis is that of *endmembers*, which represent canonical spectral signatures of pure materials present in a scene [14]. Real-world pixels are frequently *mixed*, containing combinations of multiple materials. *Spectral unmixing* decomposes pixel spectra into endmember abundances, producing abundance maps that reveal material distributions and provide physically interpretable features for downstream analysis [15]. Such representations are grounded in the linear spectral mixing model and radiative transfer principles that govern hyperspectral image formation [16, 17]. These features are particularly valuable for target recognition tasks, as they offer interpretability and robustness beyond raw spectral measurements (see Fig. 6).

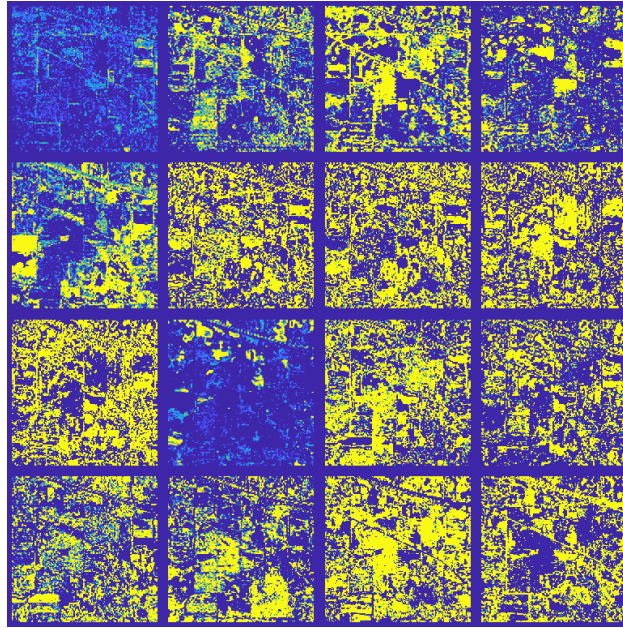


Figure 6: Abundance map for Indian Pines.

Within AiTR pipelines, *hyperspectral anomaly detection* (HS-AD) serves as a critical

early-stage cueing mechanism. HS-AD aims to identify pixels or regions whose spectral characteristics deviate significantly from the local background and may correspond to targets, foreign objects, or novel materials [13]. Because labeled hyperspectral datasets are scarce and expensive to obtain, HS-AD is often formulated as an unsupervised or weakly supervised binary classification problem (anomaly versus background) [18]. Detected anomalies can then be passed to subsequent recognition, classification, or tracking stages within the AiTR system [2, 3].

A wide range of HS-AD methods based on statistical, subspace, and learning approaches have been proposed in the literature. These approaches differ in their assumptions about background distributions, scene homogeneity, and target characteristics. In practice, no single detector consistently dominates across all environments, as performance depends strongly on scene content, background variability, and acquisition conditions [13]. This sensitivity motivates the use of *ensemble strategies*, which combine complementary detectors or feature representations to improve robustness and reduce false alarms under diverse operating conditions.

Despite these advances, robust generalization remains a central challenge. Models that perform well on one dataset frequently degrade when applied to new scenes or sensors. Addressing this challenge requires approaches that integrate physically interpretable spectral unmixing features, ensemble learning, and domain-aware modeling to maintain performance across heterogeneous domains without relying on extensive labeled data.

1.2 Hyperspectral Image Semantic Segmentation in ATR Systems

While anomaly detection provides effective cueing for potential targets, *full scene understanding* in AiTR applications often requires *semantic segmentation*. Semantic segmentation assigns a semantic label to each pixel, enabling precise localization and delineation of objects and materials across the scene. In hyperspectral imagery, segmentation leverages both spectral and spatial context to support applications such as crop mapping, land-cover analysis, and fine-grained target differentiation. However, most high-performing segmentation models are designed for RGB imagery and do not directly generalize to hyperspectral inputs, motivating the

development of HS-adaptive architectures that preserve efficiency while exploiting rich spectral information.

1.3 Challenges in Robust Hyperspectral Target Detection

1.3.1 High Dimensionality and Background Variability

HSI provides hundreds of bands with rich discriminative content, but high dimensionality increases sample complexity and degrades the stability of covariance-based detectors (e.g., RX) [19]. Real scenes exhibit heterogeneous backgrounds (soil, vegetation, water, manmade surfaces) whose reflectance varies with illumination, moisture, and atmospheric conditions [5]. Mixed pixels and nonlinear interactions further violate linear/Gaussian assumptions [20, 21]. Physically interpretable features from spectral unmixing, such as abundance maps and endmember estimates, provide material-level cues that can improve robustness and interpretability [14, 15].

1.3.2 Class Imbalance and Evaluation Instability

Anomalies are rare by definition, producing severe class imbalance that can bias training and inflate some evaluation metrics (e.g., ROC-AUC). In HS-AD, metrics emphasizing positive-class performance—such as *F1-macro* and precision–recall curves—often provide a more operationally meaningful view than ROC-AUC alone [22, 23]. Robust evaluation demands cross-scene validation, stratification, repeated folds, and statistical testing to ensure fair comparisons and stable rankings.

1.3.3 Domain Shift Across Scenes and Sensors

In hyperspectral target detection, *domain shift* arises from changes in sensing conditions, environmental factors, and background composition that alter the statistical relationship between spectral signatures observed during training and those encountered at test time. As a result, models trained on one dataset often exhibit degraded performance when applied to different

scenes, sensors, or environmental conditions. Variations in background composition, sensor spectral response, atmospheric path radiance, and flight geometry distort feature distributions and challenge fixed decision rules [24–26].

Mitigating these effects motivates the use of *transfer learning*, *domain adaptation*, *domain generalization*, and *mixture-of-experts* (MoE) designs, which aim to accommodate shifting data regimes while preserving generalization across operational conditions [27–29].

1.4 Research Objectives and Questions

This dissertation aims to advance robust HS-AD and target detection for AiTR by integrating heterogeneous ensembling and domain-adaptive deep representations under practical deployment constraints. The core objectives are:

1. **Design robust heterogeneous HS-AD stacking ensembles** that outperform individual detectors and prior ensembles across diverse backgrounds and datasets.
2. **Systematize selection** of base detectors using greedy search, reducing reliance on ad hoc expert tuning while improving accuracy.
3. **Integrate unmixing-derived features** (endmembers, abundances) and **spectral similarity** (e.g., SID) to inject material awareness and reduce false alarms, including scenarios with unknown target spectra.
4. **Improve generalization under domain shift** using transfer learning, domain adaptation, and MoE gating over dataset/sensor regimes.
5. **Extend to semantic segmentation** for fine-grained scene understanding (e.g., crop mapping) via minimally invasive HS adaptations (e.g., channel attention) to pre-trained segmentation networks with limited labels.

These objectives motivate the following research questions (RQs):

- **RQ1:** How does task-oriented image processing enhance human visibility and scene understanding, and how does HS image processing improve anomaly detection under real-world UAV constraints?
- **RQ2:** Which combinations of heterogeneous HS-AD methods and unmixing-derived features, selected greedily, yield consistently robust performance across diverse backgrounds?
- **RQ3:** Do stacking with principled weighting reliably outperform constituent detectors under F1-macro/PR-based evaluation?
- **RQ4:** How can abundance maps be leveraged when target spectra are unknown to improve detection while reducing false alarms?
- **RQ5:** Which evaluation protocols and metrics best reflect operational utility for HS-AD under severe class imbalance?
- **RQ6:** To what extent do transfer learning, domain adaptation, and mixture-of-experts (MoE) routing enhance the robustness of performance across different scenes and sensors?
- **RQ7:** How can pre-trained RGB segmentation models be minimally adapted to HS inputs with limited labels while preserving efficiency and accuracy?

1.5 Scope and Organization of the Dissertation

This dissertation focuses on anomaly and target detection in airborne hyperspectral imagery for aided and automatic target recognition (AiTR/ATR) and related applications, including agricultural monitoring. Emphasis is placed on unsupervised and weakly supervised learning paradigms due to the limited availability of labeled hyperspectral data [18], while supervised ensemble strategies are also explored when reliable ground-truth is available.

The methodological scope of this work includes:

Task-Oriented Hyperspectral Image Processing for Scene Understanding. Development of a degradation-aware image processing framework that models image degradations and enhancement strategies in a structured, task-driven manner under practical UAV acquisition constraints while improving human interpretability and visual understanding. Investigation of HS image processing to explicitly support machine scene-understanding tasks, including unmixing, anomaly detection, and semantic segmentation

Ensemble-Based Hyperspectral Anomaly Detection Architectures. Design of unified ensemble frameworks that integrate heterogeneous statistical and learning-based detectors through stacking strategies for improved robustness and detection reliability.

Physically Interpretable Spectral Feature Integration. Incorporation of physically interpretable spectral representations, including unmixing-derived abundance features and spectral similarity measures, into data-driven anomaly detection pipelines.

Robust Learning Under Domain Shift. Investigation of transfer learning, unified MoE stacking ensemble architecture, and adaptive expert modeling to address cross-sensor and cross-scene variability in hyperspectral data.

Extension to Downstream Hyperspectral Semantic Segmentation. Expansion of anomaly detection methodologies to dense prediction tasks for aided target recognition, with emphasis on agricultural crop mapping using hyperspectral data.

The dissertation builds upon and extends the author’s and collaborators’ peer-reviewed publications [30–34].

The remainder of the document is organized as follows. Chapter 2 reviews the current state of the art and the relevant literature in hyperspectral imaging, anomaly detection, and target recognition. Chapter 3 examines hyperspectral image preprocessing and its impact on machine-based detection, as well as RGB image processing for improved human interpretability. Chapter 4 introduces hyperspectral anomaly detection methods and presents ensemble strategies,

including supervised and unsupervised stacking with greedy base-model selection. Chapter 5 investigates generalization challenges and domain shift, presenting empirical results and proposed mitigation strategies. Chapter 6 addresses downstream applications, extending anomaly detection to semantic segmentation for aided target recognition, with an emphasis on agricultural hyperspectral segmentation using task adaptation and channel attention. Finally, Chapter 7 concludes the dissertation and outlines directions for future research.

1.6 Contributions

This dissertation makes the following technical and empirical contributions:

1. **Hyperspectral Image processing.** The proposed skeleton framework incorporates noise computation and channel removal in a structured, task-oriented manner to support machine scene-understanding tasks, including unmixing, anomaly detection, and semantic segmentation, under practical UAV acquisition constraints.
2. **Degradation-Aware Image Processing Framework.** A structured learning framework for RGB image processing is introduced that jointly identifies image degradations and models enhancement strategies in a task-oriented manner, thereby improving human interpretability and visual understanding compared with generic RGB image-processing approaches.
3. **Greedy Ensemble Anomaly Detection (GE-AD) for Hyperspectral Target Detection.** A systematic greedy ensemble learning framework is proposed for hyperspectral anomaly detection, jointly selecting and adaptively weighting heterogeneous statistical and machine-learning-based detectors [15, 19, 23, 35–39]. GE-AD optimizes ensemble composition under cross-validation and supports both supervised stacking (Random Forest meta-learner) and unsupervised stacking (Gaussian Mixture Model meta-learner) [30–32, 40]. Across 22 scenes from multiple datasets, supervised GE-AD achieves an average F1-macro score of 0.865 under repeated 5-fold cross-validation, while unsupervised

GE-AD attains an average ROC-AUC of 0.918 (std. 0.131), outperforming individual detectors by at least 7% under repeated 2-fold validation [31, 32]. The framework is lightweight and suitable for resource-constrained UAV/UAS deployment [2–4].

4. **Physically Interpretable Spectral Feature Integration.** Physically interpretable spectral representations, including FCLS-based abundance maps [15] and Spectral Information Divergence (SID) [41], are systematically integrated into ensemble detection. An enhanced unmixing-based anomaly formulation enables detection of small, spectrally distinct objects without requiring prior target spectra, improving robustness in unknown-target scenarios [30, 34].
5. **Unified MoE Stacking with Domain Adaptation and Mixture-of-Experts.** A unified MoE stacking ensemble architecture is proposed to improve cross-domain generalization. A pre-trained Vision Transformer serves as a deep feature extractor, while dataset-specialized ensemble experts are adaptively gated through a feedforward mixture-of-experts mechanism. Domain adaptation aligns feature distributions across sensors and scenes without retraining base detectors, achieving an average macro-F1 of 0.821 and ROC-AUC of 0.943 across seven datasets from four sensors [27, 28, 34].
6. **Evaluation Methodology Under Extreme Class Imbalance.** A comparative analysis of evaluation metrics for hyperspectral anomaly detection is conducted, demonstrating limitations of ROC-AUC under severe class imbalance. The work advocates a combination of ROC-AUC with F1-macro and precision–recall–oriented evaluation, and establishes statistically grounded repeated-validation protocols for fair and reproducible comparison [22, 23, 31].
7. **Comprehensive Multi-Dataset Empirical Validation.** Extensive experiments across benchmark hyperspectral datasets (ABU, San-Diego, Salinas, HYDICE-Urban, and Arizona) demonstrate consistent performance improvements over strong baselines and

prior ensemble methods, providing empirical insight into the interaction between ensemble learning, physically interpretable spectral features, and domain adaptation.

8. Hyperspectral-Adaptive Semantic Segmentation for Agricultural Target Recognition.

A TransUNet-based architecture is adapted to hyperspectral inputs using channel-attention mechanisms and task-adaptive transfer learning. The design preserves pre-trained knowledge while minimizing architectural and computational overhead, achieving 76.6% accuracy and 74.9% Dice score on UAV-based agricultural hyperspectral imagery [33, 42, 43].

Taken together, these contributions bridge physically interpretable spectral feature understanding and data-driven learning, delivering robust, generalizable, and deployment-ready HS target and anomaly detection within AiTR/ATR pipelines.

Chapter 2

Literature Review and Conceptual Framework

This chapter reviews the foundations of hyperspectral anomaly detection and semantic segmentation. Instead of an isolated survey, it organizes prior work according to the processing and learning pipeline used in this dissertation.

The discussion begins with hyperspectral preprocessing, which affects learning stability and performance in the presence of class imbalance. It then examines anomaly detection and hyperspectral unmixing to establish background–target modeling foundations. Based on these principles, ensemble and meta-learning strategies are reviewed, including unmixing-based ensemble approaches that inspire the proposed framework.

Since hyperspectral models are frequently deployed across sensors and scenes, the chapter further analyzes domain shift, domain adaptation, and test-time adaptation strategies, highlighting the need for robust and drift-aware learning mechanisms. Finally, hyperspectral semantic segmentation and its evaluation protocols are introduced to situate anomaly detection within the broader context of aided target recognition.

Together, these sections form the conceptual framework that informs the methodological contributions presented in subsequent chapters.

2.1 Hyperspectral Image Preprocessing

Digital image processing, a subcategory of digital signal processing, enables a broad spectrum of algorithms to be applied to the digital input data. Its ability to resolve noise and distortion during processing is a significant feature. It is in considerable demand for its applications in diverse fields such as the agriculture, defense, environment, manufacturing, and medical imaging. The evolution of digital image processing is significantly influenced by the development of mathematics, particularly the creation and enhancement of discrete mathematics theory. In this section, we will dive deeper into a select few existing state-of-the-art image

processing and enhancement solutions pertinent to this proposal. We will also review some examples followed by our proposed solution.

2.1.1 Initial HS Image Preprocessing

Radiometric calibration is the very first step in converting raw digital numbers (DNs) to radiance values. However, the sensor manufacturer supplies this solution to ensure the data represents physical measurements. We also apply geometric Correction (Orthorectification) to correct geometric distortions due to UAV movement (pitch, roll, yaw) and terrain [44].

Atmospheric Correction (AC, also referred to as Atmospheric Compensation, Characterization, or ACA) Algorithms are used in the application of remotely sensed HS imaging data to correct for the effects of atmospheric scattering, propagation, and absorption on measurements acquired by air and space-borne systems. It is usually the first step after collecting new HS data, and it removes the atmosphere's impact on an image's reflectance values [45]. FLAASH (Fast Line of Sight Atmospheric Analysis of Spectral Hypercubes) [46] is one of the standard AC techniques. The goal of AC is to convert the radiance the sensor receives at the top of the atmosphere to the radiance of the Earth's surface. We apply FLAASH as the first step in our preprocessing pipeline. Wang et al. also found that effective atmospheric correction is essential for minimizing atmospheric interference in UAV-based hyperspectral imagery [47].

Analyzing the signal-to-noise ratio (SNR) for the HS images obtained from a new sensor is essential to ensure that the sensor maintains excellent overall sensitivity and spectral resolution, which supports the capture and preservation of critical information in the scene. Noise reduction techniques can remove unwanted noises identified during SNR analysis through smoothing, blurring, and filtering. Gaussian and median blur methods are some of the standard noise reduction techniques [48]. We selected the Gaussian filter over the median for better performance in our unmixing pipeline.

Deblurring restores a sharp image from a blurred one, which can sometimes reduce noise. A Wiener low-pass filter is one of the first deblurring techniques to reduce additive random noise

in images [49]. K. Dabov et al. [50] proposed the BM3D denoising method using modified collaborative Wiener filtering. Y. Mäkinen et al. [51] introduced an improved version called BM4D with exact computation of the noise power spectrum and practical approximations of the spectrum for faster computation. We have tested BM4D and Minimum Noise Fraction (MNF) [52]; however, we have not seen any improvement in detection accuracy. We plan to test it on RGB images for better visibility.

2.1.2 Data Normalization for Machine Learning

Aggarwal [53], and Zimek et al. [54] investigated outlier detection using ensemble methods. They found using multiple methods with wide ranges of output values challenging since this often produces unbalanced ensemble results. Thus, we have investigated potential ways for normalizing anomaly detection scores. Normalization scales the feature values in a dataset to a desired similar range, often between 0 and 1, enhancing machine learning models' performance. It prevents features with larger magnitudes from dominating the learning process, allowing the model to learn more effectively from each feature. This leads to better model performance and faster convergence during training. Välikangas et al. [55] comprehensively evaluated various data normalization methods, which significantly helped our understanding and shortlisting choices. Zhao *et al.* [56] demonstrated the effectiveness of quantile normalization in their study. This method involves first ranking the gene of each sample by magnitude, calculating the average value for genes occupying the same rank, and then substituting the values of all genes occupying that particular rank with this average value. The next step is to reorder the genes of each sample in their original order [57]. It transforms the original data to remove any unwanted technical variation and preserve meaning [58]. It standardizes the distribution of each variable in a dataset, making their distribution identical across variables and allowing for easier comparison of patterns across different variables. However, this can obscure the original scale, variability, and magnitude, resulting in a loss of interpretability and original distribution characteristics.

2.1.3 Image Processing Algorithms for Visual Enhancement

Histogram equalization is a contrast adjustment technique that increases the global contrast of an image with a narrow range of intensity values [59–61]. It evenly distributes the densely populated intensity values into the full range of intensities in the histogram. It allows areas with lower local contrast to achieve a higher overall contrast, enhancing the contrast of degraded images. Adaptive histogram equalization (AHE) differs from traditional histogram equalization by calculating multiple histograms for distinct sections of an image, which helps enhance local contrast and edge definitions [62]. However, AHE can overamplify noise in relatively homogeneous regions of an image. To address this, contrast-limited adaptive histogram equalization (CLAHE) is introduced [63]. Its purpose is to limit the amplification and reduce noise, thereby improving AHE’s performance. HSV CLAHE applies CLAHE to the brightness or value part of the HSV image (hue or tint, saturation or amount of gray, value or brightness), and LAB CLAHE applies CLAHE to the lightness part of the LAB image (lightness, green-magenta color axis, yellow-blue color axis). We have evaluated various types of histogram equalization and plan to continue using them for improved visibility.

2.1.4 Image Quality Evaluation Metrics

Image quality assessment (IQA) algorithms automatically evaluate image quality relative to human observers’ perceived quality [64, 65]. Comprehensive research has evaluated image quality and assessed the performance of image enhancement algorithms. Typically, the mean squared error (MSE) and peak signal-to-noise ratio (PSNR) are used to predict noise-reduction performance. The structural similarity index measure (SSIM) assesses the improvement achieved through enhancement. Blind IQA methods, such as the Natural Image Quality Evaluator (NIQE), are useful when no reference images are available for comparison. These methods typically evaluate distortion- and blur-reduction algorithms.

2.2 Classical Hyperspectral Target and Anomaly Detection Methods

Table 1 shows various types of individual HS-AD Algorithms. The Reed–Xiaoli (RX) Anomaly Detector algorithm characterizes the HS image’s background using the hypercube’s mean and covariance [19, 66]. RX calculates the Mahalanobis distance between the background and the pixel under test. The pixel will be declared an anomaly if the distance exceeds a predefined threshold [67]. The RX algorithm is currently the benchmark for performance comparison among most anomaly detection algorithms. It has gained popularity by enabling the detection of anomalies without the need for scarcely available labeled data. However, the RX algorithm does not account for the influence of anomaly targets on background computation. The Kernel-RX Algorithm (KRX) [35] treats HS image data as complex and nonlinear. It uses the Gaussian Radial Basis Function (RBF) kernel trick to compute global normal statistics, project them into a linear model in a higher-dimensional feature space, and apply the RX algorithm in that space. Thus, KRX produces better results than the RX algorithm for a complex scene. One of the main limitations of KRX is that it is computationally intensive. If the dimensions of the HS image are large, then KRX would also require a large amount of memory.

Table 1: Examples from various types of HS-AD algorithms.

Type of HS-AD	Examples
Statistics based	RX, MD-RX, WIN-RX
Subspace based	LSUNRSORAD, CSD, SSRX
Cluster-based	CBAD, FCBAD
Spatial–spectral based	AED
ML-based	SVM, iForest
Kernel-based	KRX, GM-RX, KIFD

Schaum et al. [68] in SSRX removed several high-variance Principal Components (PCs) as background clutter identified in principal component analysis (PCA) to improve the RX performance. They later introduced CSD [69], in which the highest-variance PCs define the background subspace, and the remaining PCs (the complementary subspace) define the target subspace.

The Gaussian Mixture RX Anomaly Detector (GMRX) [36] fits the Gaussian Mixture Model (GMM), assuming the entire image is the background, and assigns pixels to the highest posterior probability mixture component. Then, it applies the RX algorithm to those pixels. However, GMRX fails if it assumes an incorrect number of components.

The Complementary Subspace Detector (CSD) [37] considers HS image data to be linear, assumes that the background and target are complementary subspaces of principal component analysis (PCA), and applies the RX algorithm in those subspaces. PCA is a linear dimensionality reduction technique that uses the data's singular value decomposition (SVD) to project it to a lower dimensional space.

Cluster-Based Anomaly Detection (CBAD) [70] considers each pixel to be strictly a member of one cluster in k-means clustering. It computes the pixel's Mahalanobis distance to the component mean. Fuzzy Cluster-Based Anomaly Detection (FCBAD) [71] is a novel extension of the CBAD and GMRX algorithms. It assumes that each pixel can have several possible fuzzy logic membership functions in fuzzy c-means clustering and computes each pixel's Mahalanobis distance to the component mean. The weakness of any clustering-based outlier detection is that the efficacy largely depends on the clustering method used, and these methods are hard to optimize for outlier detection. The model is not more flexible as it needs to be adjusted according to the distribution characteristics of the different datasets. It fails if it assumes an incorrect number of components. Finally, the application of clustering techniques to large datasets is usually expensive.

HS images provide spatial and spectral data about objects of interest. Spatial information can improve the reliability and stability of anomaly detection if the resolution is high enough. For example, HS-AD with Attribute and Edge-Preserving Filters (AED) [23] consists of two stages. This algorithm assumes that anomalies tend to be smaller and possess unique reflectance signatures and that the pixels belonging to the same class would have a high correlation in the spatial domain. First, it utilizes principle component analysis to reduce the data's dimensionality. Then, a predefined attribute filter is used along with boolean-based fusion to create the initial

detection map. The detection map is then refined using edge-preserving filters to reduce the false positive outliers. AED uses a predefined filter, which may not be suitable for various datasets. Moreover, AED assumes that the anomalies must be small objects, making it unsuitable for detecting the large anomalies in a scene.

Du et al. [72] proposed Local Summation Anomaly Detection (LSAD), which utilizes the second-order Mahalanobis distance and a window filter to create the local summation. LSAD represents the correlation between backgrounds via a correlation matrix. The matrix inverse is used during mathematical computation, making the process computationally intensive. Thus, Tan et al. [39] proposed Local Summation Unsupervised Nearest Regularized Subspace with an Outlier Removal Anomaly Detector (LSUNRSORAD) to improve the process by utilizing a linear combination using multiplication and addition. The general idea of the LSUNRSORAD algorithm is the ability to represent background pixels by its neighbors. In addition, LSUNRSORAD uses two thresholds in the outlier removal process to improve the accuracy; one is called the minimum threshold, and the other is called the maximum threshold. If the value of a pixel is higher than the maximum threshold or lower than the minimum threshold, it would be considered an anomaly. The main disadvantage of this algorithm is that it assumes the background obeys a Gaussian distribution. The probability of degraded performance is increased in scenarios where this assumption does not hold.

The isolation forest (iForest) [73] constructs multiple binary trees by random sub-sampling to isolate anomalies. The iForest assumes that in an HS image, a background pixel is hard to isolate from the neighboring pixels whereas anomalous pixel separation should be relatively straightforward. As a result, in the constructed trees, regular pixels reside in a deeper part of the trees, while anomalous pixels reside in the most shallow parts of the trees. The HS-AD with Kernel iForest (KIFD) [38] algorithm assumes that anomalies are easily distinguishable and do not occur frequently. First, it maps the data into the kernel space in order to reduce scene complexity. It then applies the PCA to choose most important features. Next, it uses the iForest algorithm to construct the trees, which outputs an initial detection map that is further

improved using locally constructed iForest over sliding windows. The depth of a pixel in the tree determines if the pixel is regular or anomalous.

2.3 Hyperspectral Unmixing for Background and Target Modeling

Higher-resolution multispectral images enable whole-pixel classification but are constrained by limited spectral bands and the need to unmix materials within pixels [15, 15]. By contrast, HS images contain significantly more spectral bands than multispectral images, enabling finer spectral discrimination. However, HS images typically suffer from lower spatial resolution. As a result, an HS image pixel may be a pure pixel consisting of the spectral characteristics of a single material, or a mixed pixel representing a mixture of two or more distinct materials, with mixed pixels occurring more frequently. This mixing may arise from sensor limitations that reduce spatial resolution or from homogeneous mixtures of materials within the scene [15].

Hyperspectral unmixing identifies and quantifies the individual components within a mixed pixel in an HS image. It decomposes the measured pixel spectra into a set of constituent spectral signatures, referred to as endmembers, and their corresponding fractional contributions, known as abundances [15, 74]. Due to the lower spatial resolution of HS imagery, unmixing plays a critical role in identifying materials and objects using their physical reflectance properties.

The first step in the spectral unmixing process is estimating the number of unique spectra, or endmembers, present in the HS image. This can be accomplished using algorithms such as the noise-whitened Harsanyi–Farrand–Chang (HFC) method [75]. Once the number of endmembers is known, algorithms such as N-FINDR can be used to estimate the individual endmember spectra [76]. Figure 7a shows the endmembers estimated using the N-FINDR algorithm for the ABU-Airport-I hyperspectral image. In this dataset, 14 endmembers are identified, corresponding to 14 unique spectral signatures.

Once the endmembers have been estimated, the Fully Constrained Least Squares (FCLS) method can be applied to estimate the abundance maps of the unique spectra in the HS image [15]. This method unmixes an HS image into the fractional abundance of each material at

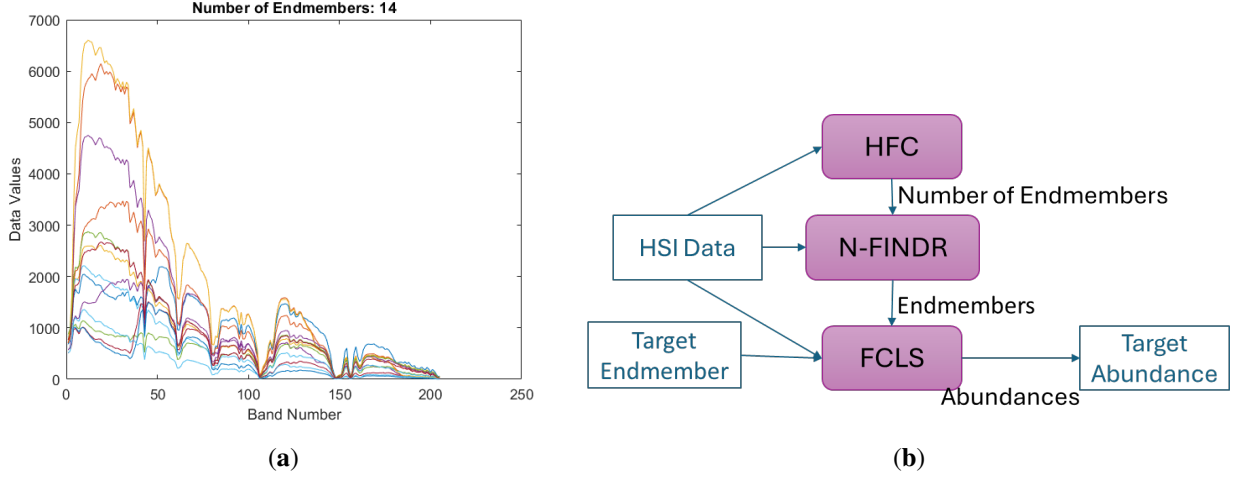


Figure 7: **(a)** Endmembers extracted from the ABU-Airport-I dataset. **(b)** The complete process for obtaining target abundance using a target endmember [30]. A target endmember is a reference spectrum extracted from a reference HS image. The process of extracting target abundance is similar to object detection.

every pixel location. The resulting abundance maps have the same spatial dimensions as the input HS image, with one abundance map corresponding to each estimated endmember.

Further analysis of HS imagery can be performed using the Spectral Information Divergence (SID) algorithm, which measures spectral similarity between pixels and specific reference spectra [41]. The reference spectra may be extracted from the same image or from a separate reference HS image. In this case, the process is similar to object detection, as illustrated in Figure 7b. The FCLS method assigns a probability distribution to each class in the form of an abundance map. The class with the highest probability is normalized to one and classified as the corresponding pure material spectrum. This normalized target abundance is used in this work.

2.3.1 Role of Abundance Information in Anomaly Detection

Abundance representations derived from hyperspectral unmixing provide a material-level description of the scene, which can reduce background variability and highlight compositional deviations associated with anomalies. Unlike raw spectral features, abundance maps encode fractional material contributions, enabling anomaly detection methods to operate in a physically

interpretable and lower-dimensional space.

Early works incorporated unmixing into anomaly detection by modeling the background as a low-rank mixture of endmembers and identifying anomalies through sparse residuals. For example, Qu *et al.* [77] combined spectral unmixing, dictionary learning, and low-rank decomposition to separate background structure from anomalous components. Similarly, Ravel *et al.* [78] addressed the presence of small targets whose endmembers appear in only a few pixels, proposing Non-negative Matrix Factorization (NMF) based strategies to recover rare endmembers that may otherwise be absorbed into the background model.

More recent approaches explicitly exploit abundance structure to enhance anomaly–background separation. Shang *et al.* [79] formulated background modeling in a tensor framework, representing the scene as a low-rank abundance tensor coupled with an endmember matrix while enforcing sparsity on anomaly components. Zhang *et al.* [80] further interpreted anomalies as deviations in endmember composition and integrated manifold-constrained sparse unmixing with spatial consistency analysis to jointly leverage pixel-level and subpixel-level information.

Beyond purely linear unmixing formulations, hybrid and learning-based frameworks have also incorporated physical mixing constraints. Chenot *et al.* [81] constructed a supervised linear background unmixing model to detect unknown targets through deviations from the estimated background energy distribution. Jiang *et al.* [82] proposed a low-rank embedded network that integrates physical mixing properties into a deep latent representation, combining low-rank modeling with density estimation for unsupervised anomaly detection. In addition, Shang *et al.* [83] explored anomaly–background separation from a spectral band selection perspective, designing criteria that prioritize bands enhancing abundance-level discrimination between background and anomalous materials.

Overall, these studies demonstrate that abundance-based representations provide a principled mechanism for suppressing dominant background structure while exposing rare or compositionally distinct materials, making them particularly effective for anomaly detection

in mixed-pixel hyperspectral imagery. In a similar spirit, recent works such as HUE-AD [30] and GE-AD (Section 4.3) further explore abundance-guided modeling strategies for anomaly detection. HUE-AD leverages hyperspectral unmixing to enhance background–target separability through abundance-driven voting refinement, while GE-AD extends this paradigm by incorporating ensemble-based modeling to improve robustness and generalization across diverse scenes. Together, these developments reinforce the importance of the abundance information as a physically meaningful and discriminative representation for hyperspectral anomaly detection.

2.4 Ensemble Learning for Hyperspectral Anomaly Detection

Classic HS anomaly detection algorithms usually make assumptions, including background distributions and the frequency of anomalies. However, real-world scenarios are much more complex than this, which is why many anomaly detectors fail to detect anomalies. Various techniques are proposed to solve these anomaly detection methods’ instability problems, and ensemble anomaly detection methods are one of them. This section reviews several ensemble learning algorithms related to our work and various approaches to automating the selection and aggregation of the base models for ensemble learning. In Table 2, we summarize all these methods’ similarities and dissimilarities.

The detectors derived from the Collaboration Representation Method (CRD) are classic HS-AD methods. CRD represents each test query as a linear combination of all the training samples from all the classes for image classification [90]. It has a high computational cost due to the use of a dual sliding window to explain a complex scene through a linear combination. Lu et al. [84] proposed an Ensemble and Random CRD (**ERC RD**) method that utilizes random sub-sampling on homogeneous CRD (RCRD) base models to reduce the computational complexity of each RCRD method. It then uses summation to aggregate the detection results from the RCRD methods. Although ERC RD reduces the computational complexity of each RCRD method, the ensemble process still brings its own computational burden without the assurances of complex detection.

Table 2: Feature comparison of ensemble learning algorithms.

Literature	Base Models	Type of Base Models	Base Model Optimization	Base Model Selection	Aggregation
ERCRD [84]	homogeneous	unsupervised	random sampling	N/A	bagging, summation
UE-kPCA [85]	homogeneous	unsupervised	random sampling, gradient descent optimization	N/A	bagging, average
ERRX MFs [86]	heterogeneous, mixed with passthrough	unsupervised	random sampling	not mentioned	bagging, average
SED [87]	only one method or heterogeneous, mixed with passthrough	unsupervised	not mentioned	based on ensemble performance	stacking, unsupervised meta method
Fatemifar et al. [88]	heterogeneous	unsupervised	not mentioned	genetic algorithm	stacking, unsupervised meta method
Nalepa et al. [89]	heterogeneous	supervised	not optimized	not investigated	stacking, supervised meta-learner
HUE-AD [30]	heterogeneous	unsupervised	not optimized	based on individual performance	bagging, voting
GE-AD	heterogeneous	unsupervised	not optimized	greedy search	stacking, supervised meta-learner

Compared to principal component analysis (PCA), kernel PCA (kPCA) can carry out non-linear projections for more complex data. Merrill et al. [85] proposed Unsupervised Ensemble kPCA (**UE-kPCA**) by suggesting an anomaly scoring function using reconstruction error in kPCA feature space. To reduce computational costs, the authors utilized sub-sampling and an ensemble of homogeneous kPCA base models to refine the final anomalies through averaging. Furthermore, they introduced a novel loss function and automated the kernel parameter selection using batch gradient descent. Despite these advancements, UE-kPCA remains a complex method that requires finding parameters for multiple kPCAs, which can be

computationally expensive.

Yang et al. [86] proposed an Ensemble and Random RX with Multiple Features (**ERRX MFs**) anomaly detector. Three features, Gabor, Extended Morphological Profile (EMP), and Extended Multiattribute Profile (EMAP), were computed using heterogeneous base models mixed with the original HS image (passthrough). They utilized random sub-sampling to compute multiple results for each feature using the RX algorithm. The results for each feature are fused using a voting average ensemble. Then, the four ensembles (for the three features and the original HS image) produce the final results. This approach avoids the problem of mixing diverse features in an ensemble. However, overrepresented features or noise can dictate the anomaly detection results without standardization.

Wang et al. [87] proposed a subfeature ensemble called **SED**. They removed the noisy bands, normalized the remaining ones, randomly sub-sampled several subfeature sets, and used six different base models in the ensemble. Unlike the ERRX MFs algorithm proposed by Yang et al. [86], SED evaluates the six base models to choose the best-performing methods and uses them to obtain an enhanced feature set. This enhanced feature set is combined with the original HS image (passthrough) to become the input for the PTA algorithm (meta-model). However, some methods used the same distribution model of the background. Due to background interference, feature redundancy, and noise, the feature enhancement process for these processes may become irrelevant to the final anomaly map. Additionally, SED uses a feature ensemble method rather than an ensemble of different algorithms, as the first-level algorithms aim to obtain the enhanced feature map.

Fatemifar et al. [88] proposed a stacking ensemble for face spoofing anomaly detection that consisted of 63 base classifiers and a Gaussian Mixture Model (GMM) as the meta-classifier for the second stage. The proposed stacking ensemble utilized all 63 classifiers by assigning weights ranging from -8 to 8 . While the application for this ensemble is different, it is worth noting that assigning values for each of the 63 weights, even those not performing well, might not result in the most optimized algorithm, especially for the larger dimensional images.

Nalepa et al. [89] proposed a deep learning ensemble for data classification and unmixing that utilized convolutional base models and model augmentation to create new modified models by adding Gaussian noise to the weights of the base models and a second level fuser. The fuser could be as simple as an averaging process, a majority vote, or a supervised second-level fuser. While the proposed method performs well for the tasks of classification and unmixing, it is unclear how effective adding Gaussian noise to the base model weights is when used for detecting anomalous objects.

Ensemble methods yield superior outcomes as they combine the predictions of multiple methods [91]. Aggarwal et al. [92] delved into the effects of using homogeneous weak outlier detectors in the bagging ensemble method in their research. On the other hand, we found that combining heterogeneous weak detectors in an ensemble creates a better understanding of the background and provides better prediction results. Thus, we explored this direction using unsupervised heterogeneous weak detectors in our previous work [30]. However, we found that the process of selecting these algorithms from many options and assigning them weights is not systematic or efficient. In this study, we propose the Greedy Ensemble Anomaly Detection (GE-AD) method, which employs a greedy search to address the above problem.

2.4.1 *Stacking and Meta-Learning Strategies*

The three main classes of ensemble learning methods are **bagging**, **stacking**, and **boosting** [93]. Stacking uses another ML model to aggregate the base models' predictions, whereas bagging uses voting, averaging, or summation, and boosting uses a sequence of models to correct the predictions of prior models.

Drawing inspiration from Feature Weighted Linear Stacking (FWLS) [40, 94], we developed a two-stage weighted stacking ensemble with unsupervised weak HS-AD methods in the first stage and a supervised machine learning (ML) method in the second stage to compute the weights.

Figure 8 shows the flowchart of a stacked model [91], which uses output predictions from

multiple base models trained on the same dataset to train another ML model (*meta-model*) to make the final predictions [95]. Please note that the meta-model will assign different weights to the base models to maximize the performance of the proposed ensemble method.

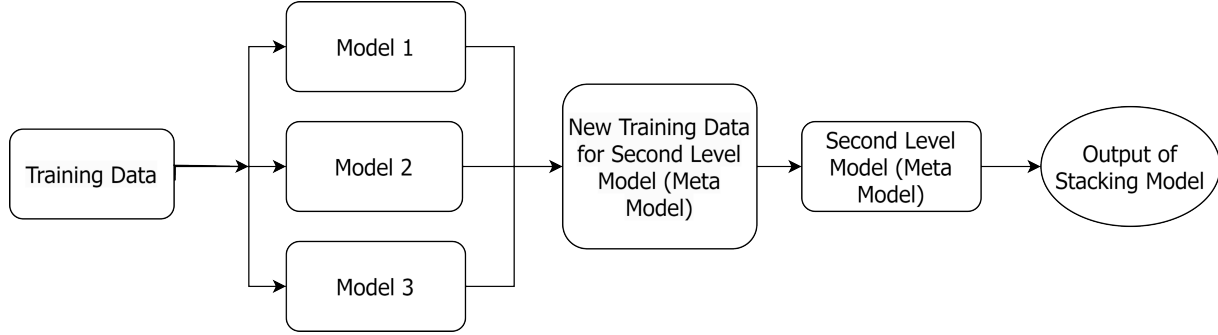


Figure 8: The architecture of a stacked model, where the new training data consist of the outputs of the first-level models.

2.5 Hyperspectral Unmixing-Based Voting Ensemble Anomaly Detector (HUE-AD)

To address the shortcomings of the statistical HS-AD methods, Younis et al. combined them into an ensemble. HUE-AD [30] is an equal-weight voting method that combines four detectors (Abundance, AED, KIFD, and LSUNRSORAD) to identify anomalies. These methods were selected based on our domain knowledge and promising precision scores. We used the unmixing method N-FINDR and FCLS to compute the abundances (see Section 2.3). FCLS assigns a probability distribution to each class, representing it as an abundance map. The class with the highest probability is identified as that pure material spectrum, and its value is normalized to 1.

HUE-AD only takes in the binary vote (a vote for a pixel means the pixel is a detected anomaly). We created binarized results by thresholding the four base models' results. Based on the three-sigma rule of thumb, a 99.7th percentile threshold can be used to identify the anomalies in normally distributed data. However, Chebyshev's inequality [96] suggests that at least 88.8% of cases should fall within properly calculated three-sigma intervals, even for variables that are not normally distributed. We conducted tests to improve the F1-macro score and determined that

a 97th- percentile threshold would be more appropriate for our purpose.

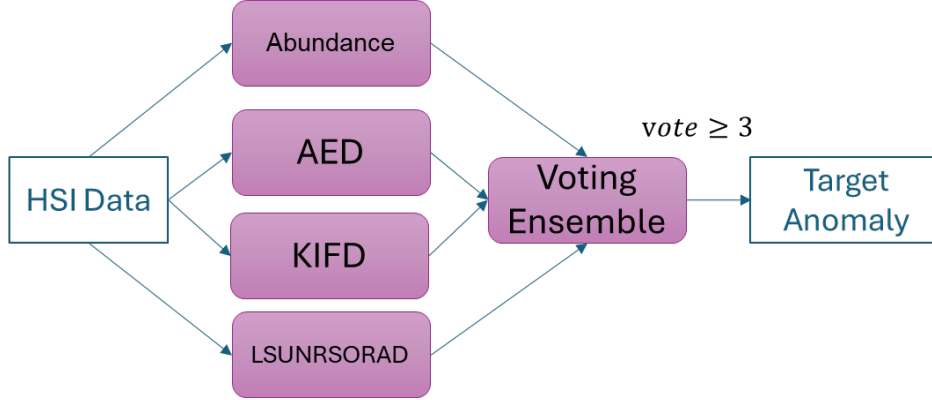


Figure 9: Design for Hyperspectral Unmixing-Based Voting Ensemble Anomaly Detector (HUE-AD).

During our evaluation, we found that systematically assigning weights to inputs improved the accuracy compared to the traditional equal-weight HUE-AD algorithm. We also found that threshold selection can be a significant source of information loss, leading to complete system failure in extreme cases. Thus, we developed our model-stacking solution, GE-AD, to learn those weights using a supervised meta-learner and a normalization method, thereby leveraging the complete information generated by the base AD algorithms without introducing bias.

2.6 Hyperspectral Anomaly Detection Datasets

We evaluated the effectiveness of the anomaly detection algorithms using seven datasets drawn from five benchmark collections: the Airport–Beach–Urban (ABU) dataset [23, 97], the San-Diego dataset [98, 99], the Salinas dataset [100], the HYDICE-Urban dataset [101], and the Arizona dataset [44].

The ABU dataset contains three subsets: Airport, Beach, and Urban. Following the preprocessing pipeline described by Kang *et al.* [23], noisy bands were identified using a noise-estimation method [102], additional preprocessing steps were applied, and unusable bands were discarded across all subsets. The ABU-Airport subset includes four airport scenes of size 100×100 pixels, extracted from larger AVIRIS images [103]. The AVIRIS sensor captures spectral radiance across 224 spectral bands covering wavelengths from 370 to 2510 nm. Three

scenes have a spatial resolution of 7.1 m per pixel, while one scene has a resolution of 3.4 m per pixel. These differences arise from variations in flight height during data acquisition. In this subset, anomalies correspond to airplanes, while the background mainly consists of roofs, shadows, and aircraft parking aprons. The ABU-Beach subset consists of four 100×100 pixel beach scenes acquired using AVIRIS and Reflective Optics System Imaging Spectrometer (ROSIS-03) sensors at varying spatial resolutions. Anomalies in these scenes include ships and other foreign objects in the sea. The ABU-Urban subset comprises five 100×100 pixel urban scenes collected with the AVIRIS sensor, each at a different spatial resolution. Anomalies in these images are primarily man-made structures located in semi-developed urban areas.

The San-Diego dataset, collected using AVIRIS, contains two 100×100 pixel images with a spatial resolution of 3.5 m per pixel. After removing low-SNR and water-vapor absorption bands, 189 spectral bands remain. Specifically, bands affected by water-vapor absorption and low signal-to-noise ratio were removed following [98, 99]. Airplanes in the scenes are considered anomaly targets.

The Salinas dataset, another AVIRIS subset, has dimensions of 512×217 pixels with a spatial resolution of 3.7 m per pixel. From the original 224 bands, 20 water-absorption bands (108–112, 154–167, and 224) were removed prior to analysis.

The HYDICE-Urban dataset consists of 80×100 pixel images captured by the Hyperspectral Digital Imagery Collection Experiment (HYDICE) sensor [104], covering 162 bands from 400 to 2500 nm. The data originate from the US AG Center [105], and bands affected by atmospheric conditions were discarded. These include bands corrupted by dense water vapor and atmospheric effects. In this scene, cars and roofs constitute the anomalies.

Finally, the Arizona dataset introduced by Watson *et al.* [44] contains five scenes captured by a UAV equipped with a Pika-L sensor [106]. The data comprise 281 spectral bands with a spatial resolution of 0.1 m, and several bands affected by atmospheric conditions were discarded. The scenes were collected at the Santa Rita Experimental Range and include varying image sizes. Anomalies include vehicles and other foreign objects embedded in the backgrounds of soil,

vegetation, and water.

2.7 Performance Evaluation under Class Imbalance in HS-AD

Hyperspectral anomaly detection (HS-AD) is inherently characterized by extreme class imbalance, where anomalous pixels typically constitute only a small fraction of the scene. In many benchmark datasets, anomaly prevalence is below 5%, and in some cases, below 1%. Across the datasets evaluated in this study, anomaly prevalence is approximately 3%. Under such conditions, conventional classification metrics may lead to misleading interpretations of detection performance.

Traditional evaluation metrics such as the receiver operating characteristic (ROC) curve are widely used to assess binary classifiers [107]. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) across varying decision thresholds. The area under the ROC curve (ROC-AUC) [108–110] summarizes performance across all thresholds and can be interpreted as the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative sample.

However, in highly imbalanced settings, the FPR may appear deceptively small even when the absolute number of false positives is large. Because hyperspectral imagery typically contains orders of magnitude more background pixels than anomaly pixels, even a small FPR may correspond to thousands of false alarms in large scenes. Consequently, ROC-AUC may overestimate a detector’s practical effectiveness in operational scenarios.

Precision–recall-based metrics provide a more informative evaluation under such an imbalance. Precision–recall (PR) AUC summarizes the precision-recall trade-off across thresholds and is more informative than ROC-AUC, as it focuses on the performance of the minority (anomaly) class.

The true positive rate at a fixed false-positive rate (TPR@FPR) evaluates detection performance under strict false-alarm constraints by measuring TPR at a fixed FPR. This is particularly important in hyperspectral anomaly detection, where even a low false-positive rate

can result in a large number of false alarms.

Precision measures the proportion of predicted anomalies that are truly anomalous, while recall measures the proportion of actual anomalies that are correctly detected:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \quad (1)$$

Here, TP denotes true positives, FP false positives, and FN false negatives. Unlike ROC-based evaluation, precision explicitly penalizes excessive false alarms, which is critical in highly imbalanced HS-AD tasks.

The F1 score [111] combines precision and recall through their harmonic mean:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

The harmonic mean penalizes extreme imbalance between precision and recall, thereby ensuring that both missed detections and false alarms are considered jointly.

To further mitigate class imbalance effects, we report the macro-averaged F1 (F1-macro):

$$F1_{\text{macro}} = \frac{\sum_{n=1}^N F1_n}{N} \quad (3)$$

where N is the number of classes (two in our case: background and anomaly), and $F1_n$ denotes the F1 score computed independently for each class. By assigning equal weight to both classes, F1-macro prevents the dominant background class from overwhelming the evaluation, thereby providing a more balanced assessment of anomaly detection performance.

Accuracy is defined as the ratio of correctly predicted samples (both positive and negative) to the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Although intuitive, Accuracy can be misleading in highly imbalanced settings, as it may

be artificially inflated when the background class dominates. To address this limitation, we additionally report Cohen’s Kappa (κ):

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (5)$$

where p_o is the observed agreement (i.e., the Accuracy metric) and p_e represents the agreement expected by chance. Unlike Accuracy, Kappa accounts for random agreement and penalizes trivial majority-class predictions, making it more informative for imbalanced HS-AD scenarios.

Following the findings of Saito and Rehmsmeier [112], precision–recall metrics are more informative than ROC curves in imbalanced detection problems. Since HS-AD operates under severe class imbalance and emphasizes reliable anomaly localization, we report F1-macro alongside ROC-AUC to provide a more comprehensive and balanced evaluation. Specifically, Precision–Recall and F1 emphasize anomaly retrieval performance; ROC AUC measures class separability; Accuracy reflects overall correctness; and Kappa assesses reliability beyond chance in highly imbalanced settings.

2.7.1 Limitations of ROC-AUC in Hyperspectral Detection

Although ROC-AUC is threshold-independent and widely adopted in detection literature, it does not explicitly account for anomaly prevalence. In hyperspectral imagery, where background pixels vastly outnumber anomalous pixels, the false positive rate (FPR) can remain numerically small even when the absolute number of false positives is operationally unacceptable.

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

For example, in a scene containing one million pixels with 1% anomalies, an FPR of only 1% would generate nearly 10,000 false alarms. Such a detection map would be impractical for

downstream analysis or human interpretation, despite potentially achieving a high ROC-AUC score.

Furthermore, ROC-AUC evaluates ranking quality across all thresholds, including operating points that may not be relevant for practical deployment. In anomaly detection, the operating region of interest typically emphasizes low false alarm rates while maintaining high precision. ROC curves do not directly reflect this trade-off.

Therefore, while ROC-AUC is reported for completeness and comparison with prior work, it does not fully capture the operational constraints of hyperspectral anomaly detection.

2.7.2 F1-Based Metrics and Thresholding Strategies

Unlike ROC-AUC, F1-based metrics require a specific decision threshold to convert anomaly scores into binary labels. Threshold selection therefore plays a critical role in determining detection performance.

In many hyperspectral anomaly detection methods, thresholding is performed using fixed percentile rules or adaptive statistical criteria. Improper threshold selection may significantly degrade precision or recall, particularly in imbalanced settings where small changes in threshold can lead to large variations in false positives.

In ensemble-based frameworks such as GE-AD, thresholding also influences how individual detector outputs are combined. Naïve binarization of detector outputs prior to fusion may result in substantial information loss, as continuous anomaly scores contain richer ranking information than binary decisions. For this reason, the proposed framework operates on continuous score distributions before applying final thresholding, thereby preserving complementary information from multiple detectors.

To ensure fair comparison across methods, thresholds in our experiments are selected consistently using percentile-based strategies, and performance is evaluated using F1-macro to reflect balanced detection of both background and anomaly classes.

2.8 Domain Shift, Generalization, and Transfer Learning in Hyperspectral Imaging

Domain shift is a fundamental challenge in hyperspectral imaging due to variations in sensors, acquisition geometry, atmospheric conditions, and scene composition. These variations can induce shifts in the data distribution between training (source) and deployment (target) environments, potentially degrading model performance. This section reviews prior work on domain adaptation, generalization, and test-time adaptation, with emphasis on their relevance to hyperspectral anomaly detection.

2.8.1 *Cross-Scene and Cross-Sensor Variability*

Distribution shift in hyperspectral imaging has been widely documented across scenes, sensors, and acquisition conditions. Variations in illumination, atmospheric absorption, viewing geometry, spatial resolution, and sensor spectral response functions induce substantial changes in the statistical distribution of spectral signatures, even when the underlying material classes remain unchanged.

Spectral variability across sensors has been particularly emphasized in prior hyperspectral studies, where differences in calibration, band placement, and signal-to-noise characteristics lead to measurable domain discrepancies. For example, datasets acquired by AVIRIS and more recent compact sensors exhibit different spectral resolutions and radiometric properties, complicating direct model transfer. Such cross-sensor variability has been shown to degrade classification and detection performance when models trained on one dataset are deployed on another without adaptation [113, 114].

Cross-scene variability further exacerbates this issue. Changes in land-cover composition, background materials, and environmental context alter background statistics, which is particularly critical for anomaly detection methods that rely on background modeling assumptions. Classical anomaly detectors assume that background pixels dominate the scene and follow a stable distribution; however, when deployed across scenes, these assumptions may no longer hold.

Importantly, in hyperspectral anomaly detection, the semantic definition of anomaly typically remains constant while the background distribution shifts. This setting differs from standard classification-based domain adaptation, where label spaces may differ across domains. Instead, the primary challenge lies in preserving anomaly contrast while mitigating background distribution mismatch. These characteristics motivate the exploration of domain adaptation and generalization strategies tailored to hyperspectral anomaly detection.

2.8.2 *Domain Adaptation and Representation Alignment*

Domain Adaptation (DA) aims to reduce distribution mismatch between source and target domains. Wang et al. [115] proposed a unified framework bridging unsupervised and zero-shot learning scenarios, assuming no access to target-domain data. In contrast, our setting assumes access to unlabeled target data, aligning more closely with unsupervised DA.

Meegahapola et al. [113] introduced M3BAT, an adversarial-based unsupervised DA method employing a shared encoder and dual discriminators to align source and target representations. Such adversarial alignment methods are powerful but introduce significant architectural and optimization complexity.

Semi-supervised DA approaches, such as the Subspace Alignment (SA) method explored by Pillai et al. [114], align statistical properties of source and target feature subspaces. SA provides a lightweight alternative to adversarial alignment by directly minimizing geometric discrepancies between domains. Their sub-sampling strategy for selecting related source examples is particularly appealing, as it promotes data efficiency and domain robustness. Given its lower complexity compared to GAN-based models like M3BAT, SA represents a promising intermediate solution for our framework.

Beyond feature alignment, Mancini et al. [116] proposed an ensemble learner with a domain predictor that estimates the probability of a sample belonging to each domain. This domain-aware weighting mechanism is particularly relevant to our ensemble-based GE-AD design, where domain-specific experts may benefit from adaptive weighting. Similarly, Ryu et

al. [117] investigated generalization strategies for random forests, offering insights applicable to our random-forest-based meta-model.

Muhammad et al. [118] addressed cross-dataset generalization in face Presentation Attack Detection (PAD) using synthetic data generation and deep ensemble learning. Their evaluation protocol—training on multiple datasets and testing on a held-out dataset using HTER and ROC-AUC—provides a rigorous cross-domain assessment strategy that we adopt in our experimental design. However, in contrast to their complex synthetic augmentation pipeline, our current framework prioritizes architectural simplicity, incorporating controlled data augmentation and lightweight ensemble learning before exploring more computationally intensive generative approaches.

2.8.3 *Test-time Domain Adaptation*

Test-time domain adaptation (TTA) aims to improve model robustness under distribution shifts at test time, where the test data (target domain) differ from the training data (source domain) [119]. Unlike traditional domain adaptation, TTA operates without access to labeled source data during deployment and typically adapts only a subset of model parameters using unlabeled test samples.

A common approach in TTA relies on normalization-based adaptation. Batch Normalization (BN) [120], which normalizes features using population-level statistics accumulated during training (source statistics), has been widely used as an adaptation handle by updating its running statistics or affine parameters at test time (target statistics). Methods such as TENT [121] minimize predictive entropy by adapting normalization parameters, while DUA [122] updates BN statistics to align test-time feature distributions with those observed during training.

Several methods rely on normalization-based adaptation mechanisms, which are effective for convolutional architectures that employ BN. However, Transformer-based architectures predominantly employ Layer Normalization (LN) [123], which normalizes features at the

individual sample level and does not maintain population statistics. This architectural difference limits the direct applicability of BN-based TTA methods to Transformer-style models. As discussed above, entropy-minimization approaches such as TENT [121] adapt normalization affine parameters and are therefore applicable to Layer Normalization as well. CoTTA [124] further demonstrates normalization-based adaptation in Transformer backbones by updating affine parameters while introducing temporal ensembling to improve stability. These methods suggest that normalization-based adaptation remains effective for LN-based models, although they do not explicitly distinguish between population-level and representation-level drift.

Moreover, normalization-based test-time adaptation presents additional challenges in hyperspectral imaging. Spectral bands are highly correlated, acquisition conditions vary across sensors and scenes, and test-time batch sizes are often small, violating the assumptions underlying Batch Normalization [120]. As a result, indiscriminate normalization updates can lead to unstable parameter shifts, distorted spectral signatures, or degraded anomaly contrast, particularly when distribution shifts are mild or absent. Prior studies have reported such negative transfer effects, motivating selective or constrained adaptation strategies [125].

More recent TTA approaches aim to improve robustness by limiting adaptation to specific model components or by introducing regularization mechanisms. SHOT [126] adapts only the classifier while freezing the feature extractor, whereas CoTTA [124] introduces temporal ensembling and stochastic restoration to stabilize adaptation over time. While these methods reduce catastrophic drift, they do not explicitly assess whether adaptation is necessary for a given test sample or distribution.

In contrast to prior work, our approach introduces a drift-aware mechanism that explicitly measures distributional deviation before enabling adaptation, which is particularly important in high-dimensional hyperspectral settings where unnecessary updates can degrade anomaly-detection performance.

2.9 Mixture-of-Experts Architectures

The Mixture-of-Experts (MoE) framework was originally introduced to decompose a complex problem into sub-problems handled by specialized models, with a gating network responsible for expert selection [127]. In this formulation, each expert is optimized using a separate loss, and the gate imposes hard assignment constraints to route each input to a single expert.

Sparsely-Gated Mixture-of-Experts (SG-MoE) further improved scalability by activating only the top-(k) experts per input, enabling large increases in model capacity with limited computational overhead [128]. A key contribution was the load-balancing loss, which prevents expert collapse in token-based routing scenarios common in natural language processing.

MoE architectures have since been widely adopted in language and sequence modeling tasks. Zoph *et al.* [129, 130] demonstrated that soft routing with top-(k) expert collaboration improves robustness and generalization across diverse linguistic domains. More recently, Feng *et al.* [131] proposed an MoE framework for autonomous driving that mitigates performance degradation under ambiguous guidance by weighting multiple experts rather than relying on a single network.

Unlike these prior works, which primarily focus on token-level or trajectory-level routing, our work applies MoE principles to hyperspectral anomaly detection, where domain shifts arise from sensor and dataset heterogeneity rather than semantic ambiguity. This distinction motivates a different routing strategy and loss design.

2.10 Hyperspectral Semantic Segmentation

Semantic Segmentation (SS) is a computer vision technique that classifies each pixel in an image into semantic categories using a deep learning (DL) algorithm. In applications that require precise object localization and boundary delineation—such as military surveillance—SS identifies and classifies every pixel within an image. It enables us to better understand a scene by distinguishing objects at a granular level, a capability that is also highly valuable

in other fields, including autonomous driving, agricultural monitoring, industrial inspection, and medical imaging. In agriculture, SS assists with crop classification, weed detection, and disease identification. It enables stakeholders to differentiate between various crops and identify unhealthy plants at a finer level.

Recent studies have examined various methods for agricultural image segmentation. For example, Luo et al. [132] reviewed recent advancements in conventional and DL approaches for segmenting agricultural images, focusing on crop analysis and pest identification. Our research builds upon TransUNet [43], a model initially developed for medical imaging that we modified for agricultural use.

U-Net is a versatile semantic segmentation network among deep networks, supporting various data types and architectures [132, 133]. Multiple researchers have included transformers [42, 43] and attention networks [134, 135] in the U-NET. Figure 10 shows the U-Net-like Visual Transformer that Chen et al. [43] had proposed. TransUNet replaced the UNet’s encoder bottleneck layer with ViTs. We specifically improved their encoder’s contracting path and network bottleneck to better enable gradient flow, ensuring the model learns effectively from hyperspectral images. Compared to TransUNet, Isensee et al. [136] argue for using vanilla U-Net and developed nnU-Net (no new net) for medical imaging. This adaptive framework can automatically configure itself for various datasets and segmentation tasks. We utilized their open-source framework and configurations as the foundation for our model development.

HSI-TransUNet [42] provided the UAV-HSI-Crop Dataset, introduced the attention module [137] in TransUNet, and tested it on this new dataset in an agricultural context. Unlike their work, we designed separate attention and residual pathways. In contrast to TransUNet, Liu et al. [137] proposed a multiscale global attention module (MGA) instead of multi-head attention at a single scale within the encoder bottleneck layer of U-Net to perform semantic segmentation of agricultural RGB images.

When there is insufficient data or a change in dataset distribution, transfer learning solves this problem by using a pre-trained model as a starting point and leveraging knowledge

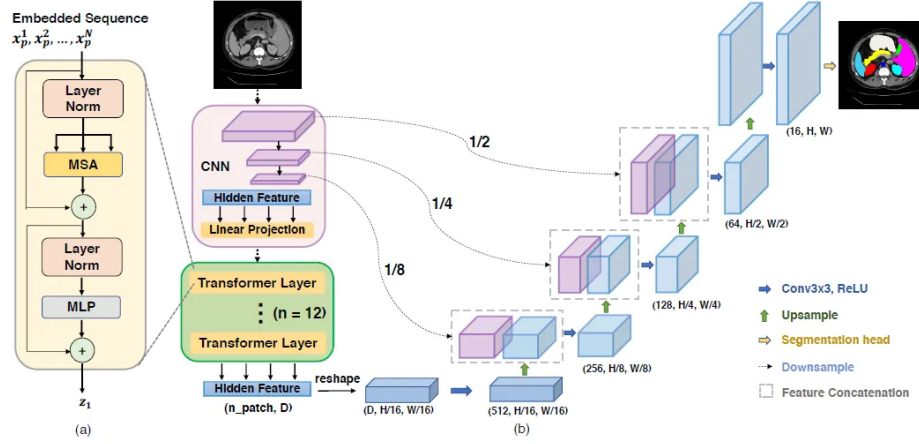


Figure 10: Overview of the TransUNet framework [43]: (a) schematic of the Transformer layer; (b) architecture of the TransUNet.

from an initial task to improve performance. Karimi et al. [138] presented a critical assessment of the role of transfer learning in training fully convolutional networks (FCNs) for medical image segmentation and found that transfer learning improved results significantly when the segmentation task was more challenging, and the target training data was smaller. We adopted transfer learning as our target dataset was small and complex, containing only 363 images for training and thirty classes. In comparison, the "Automatic Cardiac Diagnosis Challenge" (ACDC) dataset [139], which both nnUNet and TransUNet evaluated in their study, contains 100 cardiac MRI recordings for training and has five classes.

Schlemper et al. [140] proposed an Attention-gated U-Net. Figure 11 shows their proposed U-Net. Sadly, these deep SS models cannot achieve their full potential when trained from scratch on HS image datasets, because there are many large RGB datasets but very few large HS datasets for various tasks [141].

Hu et al. [142] first proposed the attention module, which they called the squeeze-and-excitation block. Wang et al. [143] later updated it with an adaptive bottleneck in their Channel Attention Module (CAM). We also used an adaptive bottleneck in the squeeze-and-excitation block to improve spectral information extraction, enabling better object

differentiation in our proposed model.

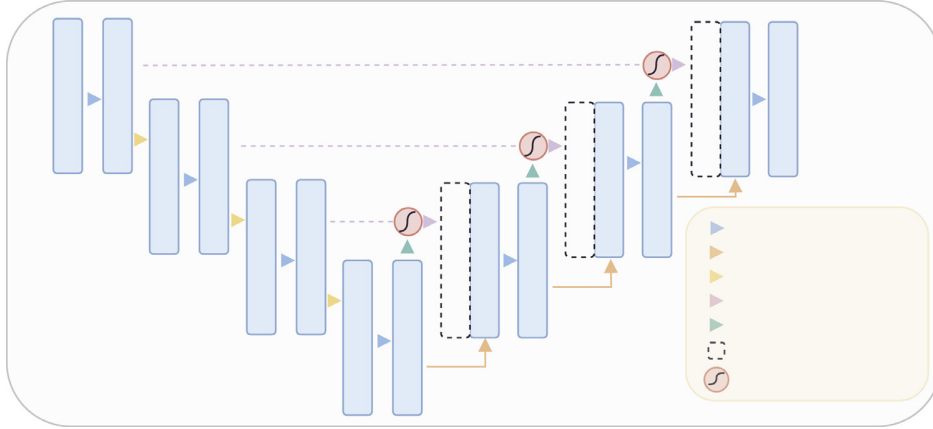


Figure 11: A block diagram of the Attention U-Net [140] segmentation model. Input image is progressively filtered and downsampled by a factor of 2 at each scale in the encoding part of the network (e.g., $H_4 = H_1 / 8$). N_c denotes the number of classes. Attention gates (AGs) filter the features propagated through the skip connections.

Compared to greyscale and RGB images, HS images contain hundreds of channels. Wang et al. [143] proposed a Channel Attention Module (CAM). Figure 12 shows a basic structure of their CAM. Leveraging this spectral information in the CAM can help our new U-Net better differentiate objects. However, HS sensors collect HS data differently, and the same sensor may produce a different number of usable channels due to noise, hindering domain adaptation. Thus, unmixing to select the most informative channels [144] ensures that our model always uses the same number of channels with the best spectral information.

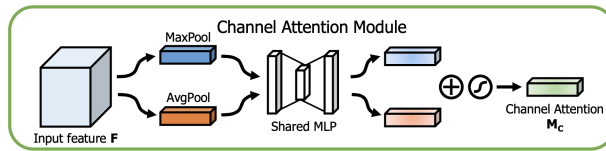


Figure 12: Channel Attention Module [143].

2.11 Hyperspectral Semantic Segmentation Dataset

The authors of UAV-HSI-Crop Dataset [42] initially used random sampling in the original dataset, which resulted in one class being absent in the training dataset, some others missing in

the test and many others missing in the validation. It posed a challenge in evaluating our model accurately. To overcome this, we performed stratified random sampling on the total dataset with a 70-15-15 split for training, validation, and test datasets. This method ensured that every class was present in all three splits: training, validation, and test, thereby enhancing the efficacy of our model evaluation. However, only one image had okra, and two had bok choy. We kept okra in the training dataset; otherwise, the model would never see okra. We distributed the bok choy between the training and validation datasets.

2.12 Performance Evaluation in HS-SS

Unlike hyperspectral anomaly detection, which primarily focuses on identifying rare anomalous pixels, hyperspectral semantic segmentation (HS-SS) aims to assign a class label to every pixel in the scene while preserving spatial coherence. Therefore, segmentation performance must be evaluated not only by classification accuracy but also by the spatial overlap between predicted and ground-truth regions.

To assess region-level agreement, we employ the Dice coefficient and the Jaccard index (also known as Intersection over Union, IoU). These metrics quantify the degree of overlap between predicted segmentation masks and ground-truth annotations.

The Dice coefficient is defined as Eq. 7

$$Dice_{coef} = \frac{2|A \cap B|}{|A| + |B|} \quad (7)$$

where A represents the set of predicted pixels for a given class and B denotes the corresponding ground-truth pixels. A Dice score of 1 indicates perfect overlap, whereas a score of 0 indicates no overlap. Dice is particularly sensitive to small target regions and is widely used in segmentation tasks involving class imbalance.

The Jaccard index (IoU) is closely related to the Dice coefficient and is commonly used to quantify the similarity and diversity of sample sets. It is defined as Eq. 8.

$$Jaccard_{index} = \frac{|A \cap B|}{|A \cup B|} \quad (8)$$

Although closely related to the Dice coefficient, IoU provides a stricter measure of overlap and penalizes over-segmentation and under-segmentation more strongly. Reporting both metrics enables a comprehensive assessment of spatial agreement.

In addition to overlap-based metrics, we evaluate pixel-wise classification performance using overall precision, recall, F1-score, accuracy, and Cohen’s kappa (κ). While overall accuracy reflects the proportion of correctly classified pixels, it can be biased toward dominant classes in imbalanced scenes. Cohen’s kappa accounts for chance agreement and provides a more robust measure of classification reliability across multiple classes.

We used multiple metrics to evaluate our model’s classification performance. Accuracy measures the proportion of instances for which a machine learning (ML) classification model is correct. Whereas, Cohen’s kappa (κ) is a statistical measure used to assess inter-rater reliability, particularly for qualitative (categorical) items, and it accounts for the level of agreement that could occur by chance.

By combining spatial overlap metrics (Dice and IoU) with classification-based metrics (F1, accuracy, and κ), we ensure that the HS-SS model is evaluated with respect to both semantic correctness and structural consistency.

Chapter 3

Hyperspectral Image Preprocessing

This chapter examines various image processing techniques applied initially to hyperspectral (HS) data, before proceeding to anomaly detection and segmentation, which may improve data quality and increase the training data for machine learning and deep learning models. These image-processing techniques enhance human visibility and understanding when applied to 1- or 3-channel digital images. These digital images can be 1-channel gray-scale and NIR images or 3-channel RGB and NIR-RG images.

3.1 Process HS images to Improve Machine Detection Accuracy

3.1.1 Preprocessing Methods to Improve Hyperspectral AD Algorithms for Arizona Dataset

When dealing with HS images from a new sensor, it is crucial to analyze the SNR to ensure that the sensor has an excellent overall sensitivity and spectral resolution. This supports capturing and preserving the most important information in the scene and increases the efficacy of pre-processing techniques such as geo-rectification [44]. We have computed the SNR for all bands using the Avhyas plugin [145] in QGIS. Many HS images contain noisy bands, so further analysis of the SNR is needed. Figure 13a displays the SNR for Image 5 from the Arizona dataset. Analysis of the figure reveals that the hypercube tends to be noisy for input wavelengths close to 400 nm and those higher than 860 nm. This suggested a procedure to remove those noisy wavelengths/bands, evaluate the anomaly detector on the new hypercube containing less noisy wavelengths, and create a new hypercube consisting of the bands with signal-to-noise ratios above our predefined threshold. The results of the noisy band removal are depicted in Figures 13a and 13b. Once all noisy bands are removed, we selected the most informative bands using the method proposed by Du and Yang [144] to reduce the hypercube's dimensionality and improve the processing times without degrading performance.

Analysis of data contained Figures 13a and 13b supports the conclusion that Figure 13b

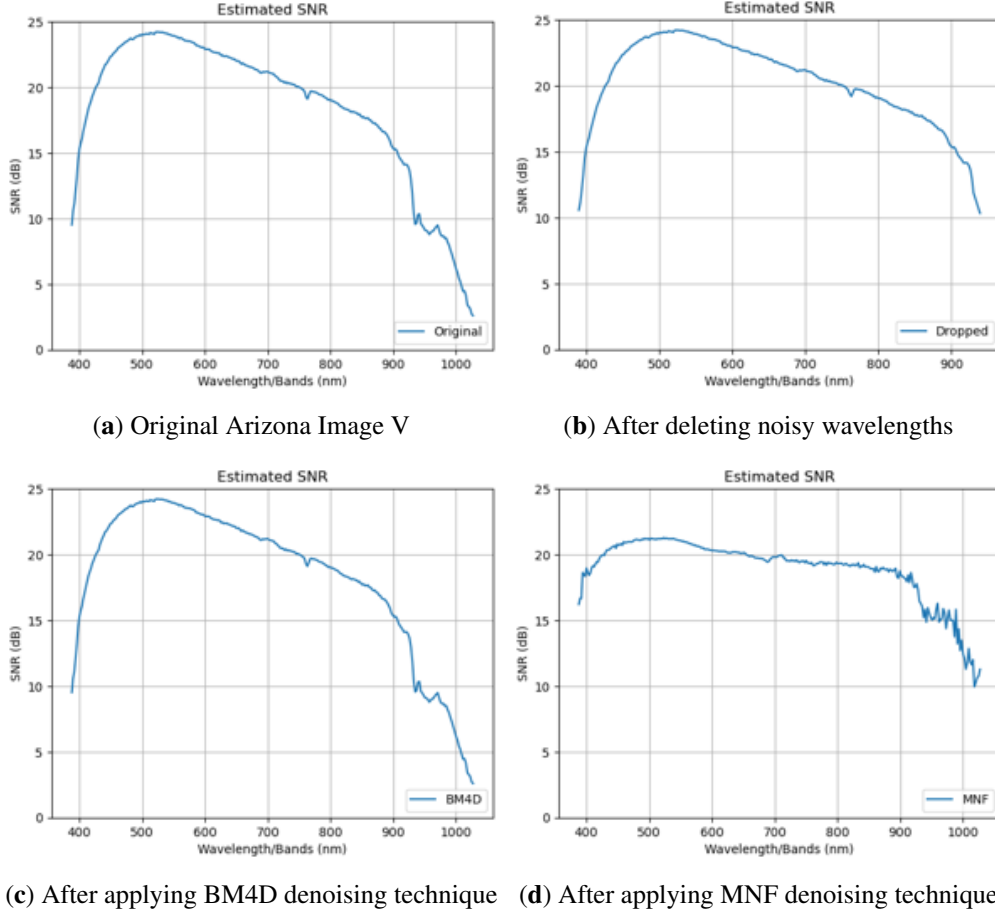


Figure 13: SNR in [dB] showing the impact of removing noisy bands and other denoising techniques from Image V in the Arizona dataset.

has a much better minimum hypercube SNR and does not result in any significant change to the input data characteristics. However, aggressively removing the noisy bands can reduce the spectral content and cause information loss, resulting in worse anomaly detection performance.

3.1.2 Removal of Noisy Bands

Previously, we have removed noisy bands below a fixed threshold (10 SNR) or below a certain percentage (97%) [31, 146]. Figure 14 shows various bands with noise. We may recover sensor noise before 400 nm and after 1000 nm, weak atmospheric water absorptions at 700 nm and 660 nm, and slightly stronger atmospheric water absorptions at 730, 820, and 910 nm.

However, the strong atmospheric water absorptions at 904 and 1140 nm are not possible to recover.

Bands located in water absorption regions and those exhibiting an extremely low signal-to-noise ratio were removed prior to training. This preprocessing step follows standard hyperspectral analysis practice, as these bands contain limited physical information and are often dominated by atmospheric noise [16, 147]. Retaining such bands may introduce artificial spectral variability and negatively impact cross-domain feature alignment. Therefore, only high-quality spectral bands were preserved to ensure robust anomaly detection and domain generalization.

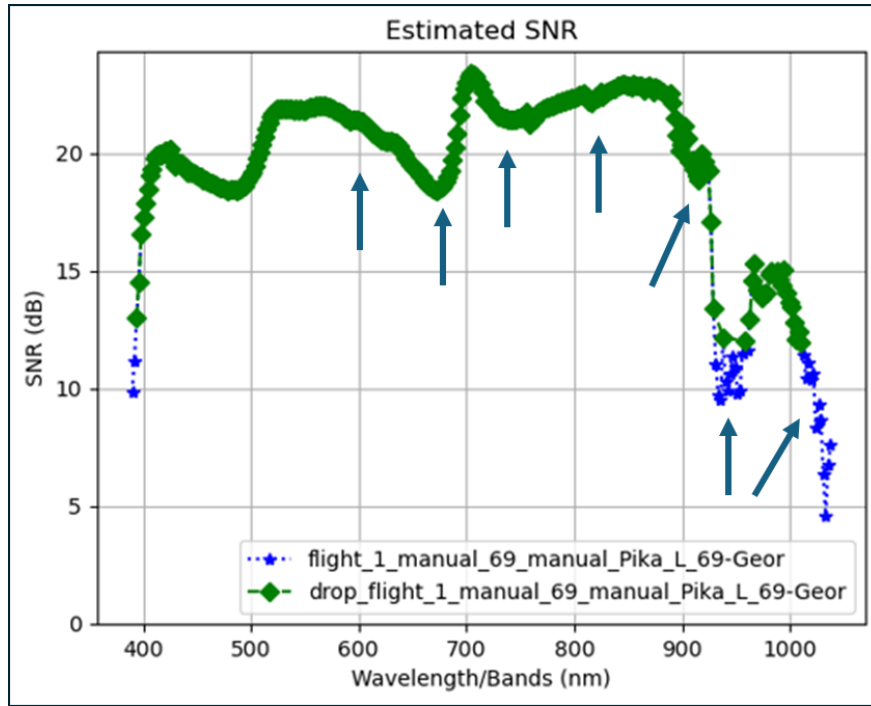


Figure 14: Visual comparison showing quantitative improvement of removing noisy bands from Florida Image 1. The green line shows SNR in [dB] after deleting noisy bands, whereas the blue line shows the dropped channels

Although aggressively removing noisy channels may lead to partial loss of spectral information, retaining severely corrupted bands can introduce artificial variability and degrade anomaly separability. In cross-domain settings, such bands may amplify domain-specific artifacts and negatively affect feature alignment. Therefore, a conservative band-selection strategy was adopted to preserve physically meaningful spectral information while minimizing noise-induced

bias.

3.1.3 Normalization of Anomaly Detection Results

We have diverse output values representing anomalies from different AD algorithms. Quantile normalization (quantile) can convert these results to a new distribution without losing relative patterns (see Section 2.1.2). Figure 15 shows how quantile normalization transformed the irregular results from KIFD into a normal distribution.

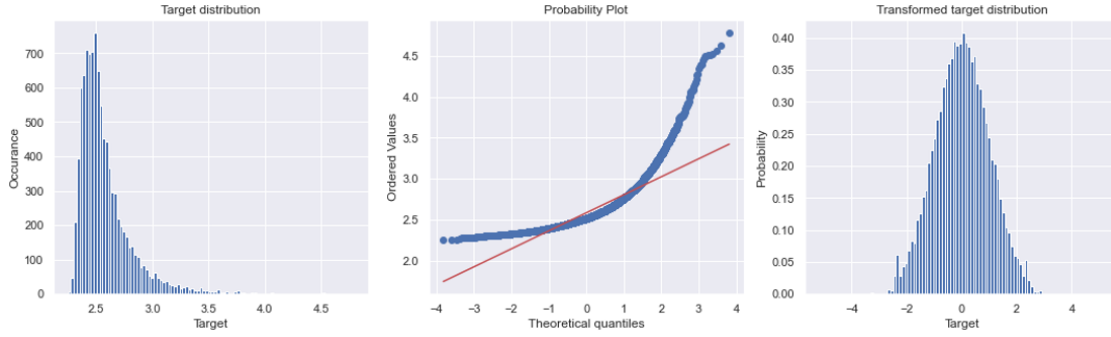


Figure 15: Irregular results from kernel isolation forest anomaly detection are transformed into a normal distribution. (a) Distribution of Original data. (b) Irregularities. (c) Transformed normal distribution.

When aggregating the output values, some meta methods, e.g. logistic regression, require normalization of the values, while others, e.g., Random Forest, inherently use information gain, which is not negatively impacted by scaling. Nevertheless, we applied normalization to ensure fair evaluation between various models. Figure. 16 shows a positive impact of the quantile normalization on the final results. We have also tested Z-score normalization, also known as standardization. It adjusts data values to have a mean of zero and a standard deviation of one. Figure 16 shows that quantile normalization boosts performance for tricky ABU-IV data.

3.1.4 Unmixing-based Anomaly Detection

Endmembers are unique spectra representing the materials in a hyperspectral image. However, the spectral signature of a pixel can contain a linear combination of multiple endmembers [14]. Spectral unmixing algorithms such as Harsanyi–Farrand–Chang (HFC) [75],

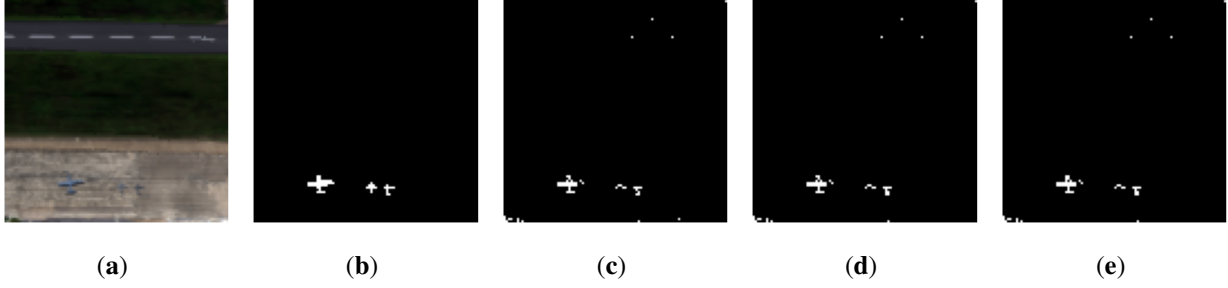


Figure 16: Normalization can give a boost in performance for tricky data like ABU-IV. (a) RGB, (b) ground-truth, (c) GE-AD with no normalization (F1 = 0.727), (d) GE-AD with z score normalization (F1 = 0.744), (e) GE-AD with quantile normalization (F1 = 0.757).

N-FINDR [76], and Fully constrained least-squares (FCLS) [15] decompose the measured spectra into a collection of endmembers, determine the proportion of each endmember in the pixel, and produce the *abundance map* using the identified endmembers. Younis *et al.* [30] used spectral information divergence (SID) [41] to measure spectral similarity between pixels and reference spectra, analogous to hyperspectral target detection. We can obtain these reference spectra from datasets such as the ECOSTRESS spectral library [148, 149] or extract them from annotated HS images within the same dataset or from other datasets. However, this method is unusable when target spectra are unknown. Thus, we propose a method that yields a set of anomalous abundances representing fewer than a small percentage (p) of pixels in a scene, since anomalies are typically rare, and use it as one of the input base methods. We found that $p < 20\%$ provides a stable trade-off between sparsity and coverage across datasets [150]. Unlike SID-based methods that require known target spectra, our formulation operates without prior target knowledge and derives anomalous references directly from the scene. Algorithm 1 summarizes the proposed target-free anomalous abundance generation procedure, which replaces predefined target spectra with a rarity-based selection of abundance responses.

In Alg. 1, $a_{i,k}$ denotes the abundance of endmember k at pixel i , and rarer endmember classes are emphasized by assigning larger weights, ensuring that the most infrequent spectral responses contribute more strongly to the anomalous abundance map.

Algorithm 1 Anomalous Abundance Generation via Spectral Unmixing

Require: Hyperspectral image cube $\mathbf{X} \in \mathbb{R}^{H \times W \times B}$, wavelengths λ , unmixing method (FCLS), rarity threshold $p = 20\%$

Ensure: Anomalous abundance map $\mathbf{A}_{\text{anom}} \in \mathbb{R}^{H \times W}$

- 1: Ensure/Construct hyperspectral cube $\mathbf{X}$ using spectral bands λ
 - 2: Estimate number of endmembers K using HFC with noise whitening
 - 3: Extract endmembers $\mathbf{E} \in \mathbb{R}^{B \times K}$ via N-FINDR
 - 4: **for all** pixels $\mathbf{x}_i \in \mathbf{X}$ **do**
 - 5: Estimate abundance vector $\mathbf{a}_i$ using FCLS
 - 6: **end for**
 - 7: Assign dominant endmember class map $c_i = \arg \max_k a_{i,k}$
 - 8: Compute pixel percentage for each endmember class
 - 9: Select rare classes $\mathcal{C}_{\text{anom}} = \{k \mid \text{percentage}(k) < p\}$
 - 10: Sort $\mathcal{C}_{\text{anom}}$ in ascending order of percentage
 - 11: Initialize $\mathbf{A}_{\text{anom}} \leftarrow \mathbf{0}$
 - 12: Initialize weighting factor $w \leftarrow 1.0$
 - 13: **for all** endmember indices $k \in \mathcal{C}_{\text{anom}}$ **do**
 - 14: **for all** pixels i such that $c_i = k$ **do**
 - 15: $\mathbf{A}_{\text{anom}}(i) \leftarrow w \cdot a_{i,k}$
 - 16: **end for**
 - 17: $w \leftarrow w - 0.05$
 - 18: **end for**
 - 19: **return** $\mathbf{A}_{\text{anom}}$
-

Final Anomaly Map:

$$A_{\text{anom}}(i) = \sum_{k \in \mathcal{C}_{\text{anom}}} w_k a_{i,k}, \quad \mathcal{C}_{\text{anom}} = \{k \mid \text{coverage}(k) < p\} \quad (9)$$

For example, in a 100×100 scene, abundance vectors are first estimated for all pixels; endmember classes representing fewer than 20% of pixels (2000) are then retained, yielding sparse abundance maps that highlight small, spectrally distinct objects. The resulting anomalous abundance map then serves as a base anomaly detector and is combined with other base methods within the proposed ensemble framework.

3.1.5 Discussion and Summary

This section demonstrates that improving machine detection accuracy in hyperspectral anomaly detection is not solely a matter of algorithmic sophistication but also of disciplined signal stewardship. The quality of anomaly detection is fundamentally constrained by the physical fidelity of the input hypercube. Sensor noise, atmospheric absorption, and scale inconsistencies introduce distortions that propagate through subsequent learning stages. Therefore, preprocessing should be viewed not as a peripheral step but as an integral component of the detection pipeline.

Band selection based on SNR analysis and atmospheric absorption removal reduces artificial spectral variability while preserving physically meaningful information. Importantly, this process reflects a balance: overly aggressive pruning may discard subtle discriminative cues, whereas retaining severely corrupted bands amplifies domain-specific artifacts and degrades separability. The adopted conservative strategy prioritizes robustness and cross-domain stability over maximal spectral retention.

Normalization further illustrates that anomaly detection performance depends not only on spatial-spectral representation but also on statistical comparability. By aligning score distributions across heterogeneous base detectors, quantile normalization enhances ensemble coherence and improves the stability of meta-level learning. The observed improvements, particularly in challenging scenes such as ABU-IV, indicate that distributional alignment can be as influential as architectural design.

The proposed unmixing-based anomalous abundance formulation extends this philosophy from signal cleaning to representation refinement. Rather than relying on predefined target spectra, anomalies are characterized by spectral rarity within the scene. This shift from target-dependent detection to rarity-driven representation increases generality and supports deployment in unknown environments where prior target knowledge is unavailable.

These preprocessing strategies unify physical signal quality, statistical consistency, and representation sparsity to enhance anomaly detection. Instead of changing the core detector,

structured preprocessing improves accuracy and generalization, serving as an inductive bias guiding models toward physically plausible, statistically stable, and domain-resilient hyperspectral data interpretations.

3.2 Image Enhancement Methods for Human Understandability

Figure 17 illustrates the challenge of improving human interpretability through digital image processing. The middle image shows automatic brightness adjustment, which shifts pixel intensities toward higher values. The right image is produced by HSV CLAHE (Contrast Limited Adaptive Histogram Equalization), which redistributes intensities to better utilize the display range and enhance local contrast.

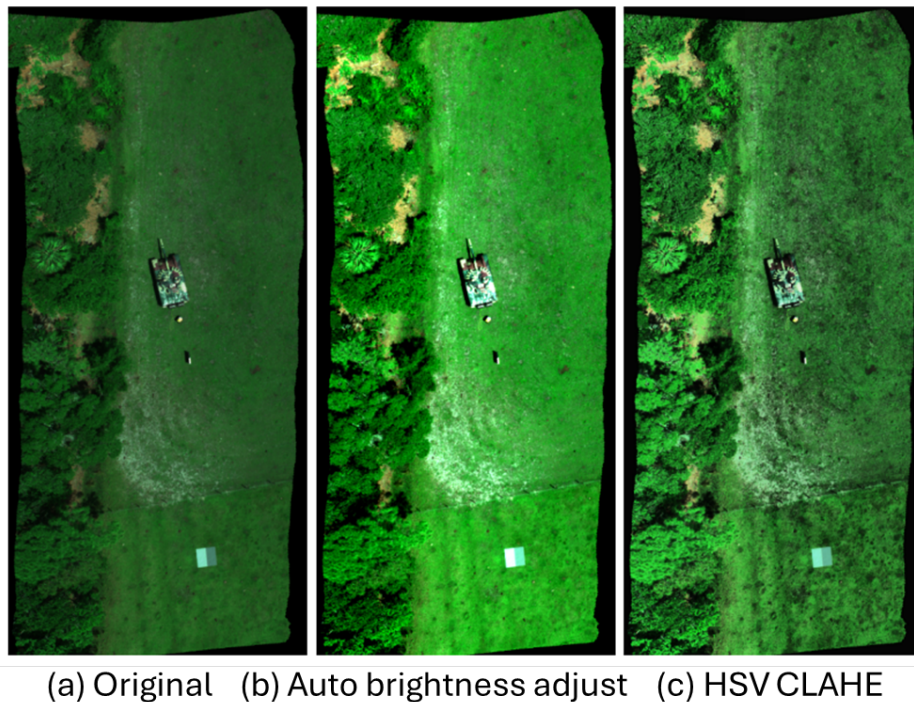


Figure 17: Example of digital image processing. From left to right: original image, auto-brightness adjustment (SSIM 0.685), and HSV CLAHE (SSIM 0.677).

While the processed images contain more visible detail, increased information does not necessarily improve anomaly perception. In some cases, the original image may better support anomaly detection. Thus, the effectiveness of image enhancement for human analysis is

inherently subjective. In this work, the goal is not to design a new enhancement algorithm, but to select the most appropriate enhancement method for a given degraded image in order to support human understandability.

3.2.1 Motivation

This study investigates enhancement method selection for improved human interpretability and aided target recognition. Rather than optimizing a single enhancement pipeline, multiple enhancement methods are applied to each degraded image. A classifier then predicts the most suitable method based on latent image features (Fig. 18).

As user-provided ground-truth preferences were unavailable, labels were generated by manually degrading clean images and selecting the enhancement method that best recovered the reference image. This transforms enhancement selection into a supervised learning problem with empirically defined targets.

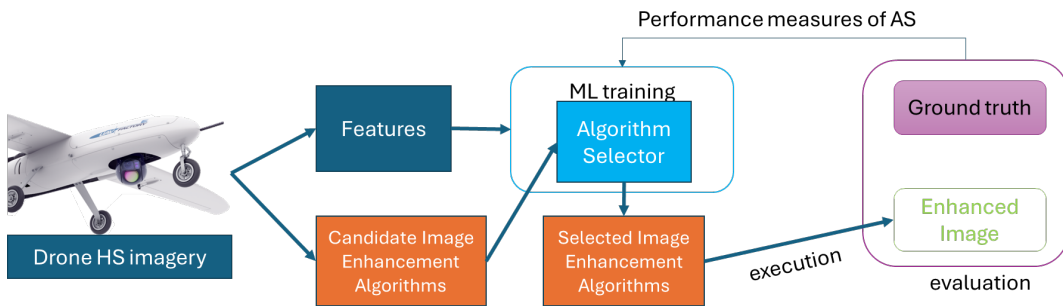


Figure 18: Previously proposed enhancement method selection framework.

3.2.2 Dataset Generation

The dataset includes nine different representative degradations that model realistic acquisition and post-processing artifacts. Blur-based distortions include *Gaussian blur* [151] and *motion blur*. *Gaussian noise* models additive sensor noise [152], while *JPEG compression* introduces quantization artifacts [153]. Resolution degradation is simulated via *down-up scaling*. Photometric distortions include *low contrast*, *gamma degradation*, *histogram clipping*, and

illumination gradient. Together, these operators span structural, statistical, and illumination perturbations.

Enhancement methods include global histogram equalization (*hist-eq*), local contrast enhancement via *CLAHE* and *clahe-lab* [63], edge-preserving *bilateral* filtering [154], *unsharp* masking [151], and *auto-brightness* contrast stretching.

Hyperspectral cubes (mentioned in Section 2.6) were converted to pseudo-RGB images using three visible bands (Red 667.5 nm, Green 540 nm, Blue 470 nm). From these, a total of $629 \times 128 \times 128$ patches were extracted, degraded, restored, and compared to their original counterparts.

Recovery quality is quantified using a composite score that balances *structural fidelity* and *contrast enhancement*. We used the degraded and restored images to compute these metrics.

Structural Quality. Structural quality is defined as a weighted combination of PSNR, SSIM [155], and Dice similarity:

$$S_{\text{struct}} = w_p \frac{\text{PSNR}}{50} + w_s \text{SSIM} + w_d \text{Dice}, \quad (10)$$

where $w_p = 0.4$, $w_s = 0.4$, and $w_d = 0.2$. PSNR is normalized by 50 to align its scale with the other metrics.

Contrast Quality. contrast quality is defined as a weighted combination of Contrast-SIM, Michelson Contrast [156], and Shannon Entropy [49]:

$$S_{\text{contrast}} = w_c \text{CSIM} + w_m \text{MSIM} + w_e \text{Entropy}, \quad (11)$$

where $w_c = 0.3$, $w_m = 0.3$, and $w_e = 0.4$.

Final Composite Score. The overall final enhancement score is computed as a convex combination:

$$S_{\text{final}} = \alpha S_{\text{struct}} + (1 - \alpha) S_{\text{contrast}}, \quad (12)$$

where $\alpha \in [0, 1]$ controls the balance between structural and perceptual quality. In our experiments, we set $\alpha = 0.5$.

This evaluation left us with four enhancement methods, *CLAHE*, *bilateral* filtering, *unsharp* masking, and *auto-brightness* contrast stretching.

3.2.3 Methodological Design Choice

Enhancement selection is formulated as a supervised classification problem, where the ground-truth label corresponds to the enhancement method that maximizes reconstruction quality relative to the reference image. This converts the selection process into a well-defined empirical risk minimization task with stable supervision.

Algorithm selection has traditionally been studied using performance-based labels under supervised or meta-learning settings [157, 158]. By explicitly computing the optimal enhancement choice across all degradation–method combinations, the framework avoids the instability associated with implicit or bandit-style selection strategies [159].

3.2.4 Proposed Method

Unlike traditional enhancement datasets consisting of paired original–enhanced images, the proposed dataset links degraded images to preferred enhancement strategies. The objective is not pixel reconstruction but rather the enhancement method classification.

We formulate the automatic enhancement method selection as a supervised multi-task problem (Fig. 19). Given a degraded image patch, the model jointly predicts:

1. its degradation type
2. the empirically determined best enhancement method

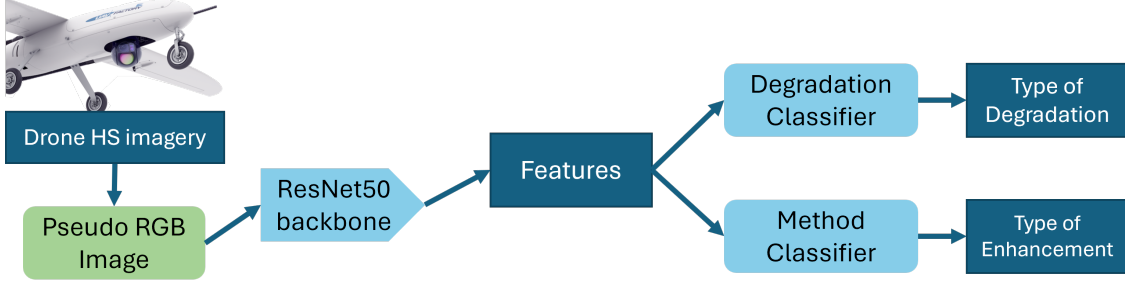


Figure 19: Image Enhancement Method Selection using ML Model

Multi-Task Model Architecture

Let $x \in \mathbb{R}^{3 \times 224 \times 224}$ denote a degraded RGB patch. Features are extracted using a pretrained ResNet50 backbone (ImageNet 2012 [160, 161]), producing a 2048-dimensional representation. The backbone is frozen during training.

The extracted feature vector is fed into two independent classification heads:

- A degradation classifier with C_d classes ($C_d = 9$).
- A enhancement method classifier with C_m classes ($C_m = 4$).

Each head comprises a linear layer, a ReLU activation, dropout, and a final linear projection. Training uses a multi-task cross-entropy objective:

$$\mathcal{L} = \mathcal{L}_{deg} + \mathcal{L}_{method} \quad (13)$$

where both $\mathcal{L}_{deg}$ and $\mathcal{L}_{method}$ are standard cross-entropy losses computed directly from the output logits.

Training Strategy

Training uses AdamW [162] (learning rate 10^{-3}), batch size 32, and 10 epochs. Only the classification heads are updated. The dataset is split using stratified sampling over joint degradation–method labels. Images are resized to 224×224 and normalized with ImageNet

statistics.

Performance is evaluated using overall and per-class accuracy for both tasks.

3.2.5 Evaluation and Results

The model achieved an overall degradation classification accuracy of 89%, with macro-averaged precision and recall of 0.89 and 0.89, respectively (see Table 3). Perfect classification accuracy was obtained for *motion blur*, *JPEG compression*, *low contrast*, *histogram clipping*, and *illumination gradient*. Lower accuracy was observed for *gaussian blur* (43%), likely due to overlap with other smoothing artifacts. The F1-macro of 0.89 indicates stable multi-class discrimination.

Table 3: Quantitative performance of the proposed model on overall degradation classification.

	Precision	Recall	F1-score	Accuracy
Gaussian blur	0.60	0.43	0.50	0.43
Motion blur	1.00	1.00	1.00	1.00
Gaussian noise	1.00	0.93	0.96	0.93
JPEG compression	1.00	1.00	1.00	1.00
Down-up scaling	0.56	0.71	0.62	0.71
Low contrast	0.93	1.00	0.97	1.00
Gamma degradation	1.00	0.93	0.96	0.93
Histogram clipping	1.00	1.00	1.00	1.00
Illumination gradient	0.93	1.00	0.97	1.00
Average	0.89	0.89	0.89	0.89

For enhancement method classification, overall accuracy reached 89% (see Table 4). *CLAHE* achieved perfect recall, while *auto-brightness* exhibited low recall (0.43), reflecting ambiguity between global intensity adjustments and other contrast-based methods. The F1-macro of 0.81 indicates good multi-class discrimination.

Table 4: Quantitative performance of the proposed model on overall enhancement method classification.

	Precision	Recall	F1-score	Accuracy
CLAHE	1.00	1.00	1.00	1.00
Bilateral filtering	1.00	0.85	0.92	0.85
Unsharp masking	0.88	0.95	0.91	0.95
Auto-brightness	0.43	0.43	0.43	0.43
Average	0.83	0.81	0.81	0.89

3.2.6 Discussion and Summary

This work formulates the selection of enhancement methods as a structured multi-task learning problem, in which a shared visual backbone extracts degradation-sensitive features, and task-specific heads predict both the degradation type and the optimal enhancement strategy. The results demonstrate that discriminative representations can be learned even with a frozen backbone, enabling reliable classification across heterogeneous degradations while significantly reducing training complexity and computational cost.

The framework establishes a practical approach to data-driven enhancement selection, grounded in empirical measures of reconstruction quality. Although currently limited to predefined degradation and enhancement methods, it provides a scalable foundation for adaptive image processing pipelines that support human interpretability. Future work may extend this framework toward domain adaptation for improved robustness under distribution shifts and toward meta-learning-based strategy selection for handling novel or mixed degradations. Additional extensions include incorporating uncertainty estimation and exploring expanded enhancement heads within the multi-task architecture to further enhance real-world applicability.

Chapter 4

Ensemble-Based Hyperspectral Anomaly Detection

This chapter focuses on ensemble-based approaches for hyperspectral anomaly detection (HS-AD), where anomalies are defined as pixels whose spectral responses deviate from the dominant background without prior knowledge of target characteristics. Traditional HS-AD methods rely on fixed statistical or structural assumptions about background behavior, which can limit their robustness across diverse scenes. To address these limitations, this chapter examines how combining multiple complementary anomaly detectors can improve detection reliability. We first discuss the motivation for ensemble learning in HS-AD and then introduce the proposed Greedy Ensemble Hyperspectral Anomaly Detection (GE-AD) framework, which automatically selects and fuses base detectors using probabilistic stacking. Extensions of this framework to unsupervised ensemble fusion (UGE-AD) and the incorporation of spectral information (mGE-AD) are also presented, providing a unified view of ensemble design choices for hyperspectral anomaly detection.

4.1 Background

Classical statistical HS-AD methods typically model the background using predefined assumptions regarding its distribution, such as Gaussianity, local homogeneity, or low-rank structure. While these assumptions enable mathematically tractable detectors, they implicitly constrain what constitutes "normal" background behavior.

In practice, hyperspectral scenes exhibit substantial variability due to differences in material composition, illumination conditions, spatial structure, and sensor characteristics. As a result, a background model that performs well in one scene may fail to generalize to another when its underlying assumptions are violated. This sensitivity manifests as unstable anomaly responses, increased false alarms, or missed detections, particularly in complex or heterogeneous environments.

These limitations highlight a fundamental challenge in HS-AD: no single detector or background model is universally optimal across all scenes. Instead, detector performance is strongly coupled to how well its embedded assumptions align with the statistical and spectral characteristics of a given dataset. Addressing this challenge requires moving beyond reliance on a fixed detector or handcrafted background model and toward mechanisms that can adaptively leverage multiple, complementary detection strategies.

4.2 Motivation

To mitigate the sensitivity of individual HS-AD methods, ensemble-based approaches have been proposed as a means of aggregating complementary detectors. One such approach, HUE-AD, incorporates domain knowledge to manually select a set of base anomaly detectors and spectral unmixing methods deemed suitable for a specific dataset. The ensemble output is then obtained by combining the binary anomaly decisions of the selected base methods through majority voting.

While this strategy improves robustness compared to single-detector methods, it introduces two critical limitations. First, the selection of base detectors relies on expert knowledge and dataset-specific tuning, which restricts transferability to new scenes. Second, majority voting treats all base detectors as equally reliable, ignoring variations in detector confidence and discriminative power across pixels and scenes.

To address these issues, we propose replacing manual ensemble construction with a greedy, data-driven selection process that automatically identifies effective combinations of base anomaly detectors and unmixing methods for a given dataset. This approach eliminates the need for prior domain knowledge in detector selection and allows the ensemble composition to be optimized directly from data.

Furthermore, instead of relying on hard voting, we investigate the use of continuous anomaly scores and probabilistic outputs to combine base detectors. Initial experiments using logistic regression demonstrate that weighting detector outputs by their predictive reliability

yields improved performance compared with unweighted voting. Building on this insight, we adopt a Random Forest–based stacking strategy to model nonlinear interactions among detector outputs, yielding further performance gains.

This progression culminates in the proposed Greedy Ensemble Hyperspectral Anomaly Detection (GE-AD) framework, which integrates greedy base-method selection with supervised ensemble fusion. To extend this framework to settings where labeled data are unavailable, we subsequently introduce an unsupervised variant based on Gaussian Mixture Models (UGE-AD), which enables probabilistic score fusion without supervision.

These observations motivate a unified ensemble-based framework that (i) automatically selects complementary anomaly detectors, (ii) combines their outputs using probabilistic fusion rather than hard voting, and (iii) supports both supervised and unsupervised ensemble learning. The remainder of this chapter presents this framework and its extensions.

4.3 Greedy Ensemble Hyperspectral Anomaly Detection (GE-AD)

4.3.1 Ensemble Learning Formulation

Ensemble learning is a machine learning technique that combines multiple individual *base models* or weak learners to create a stronger, more accurate model with better predictions. The individual base models in an ensemble can be different algorithms (heterogeneous) or the same algorithm (homogeneous), and they can be unsupervised or supervised. Note that, even if the base models are the same algorithm, they may be trained on different subsets (random sampling) of the training data. Furthermore, selecting the best parameter values using various approaches, such as cross-validation, i.e., *hyperparameter tuning*, can optimize the individual base model’s performance. The three main classes of ensemble learning methods are *bagging*, *stacking*, and *boosting*. Stacking uses another ML model to aggregate the base models’ predictions, whereas bagging uses voting, averaging, or summation, and boosting uses a sequence of models to correct the predictions of prior models.

4.3.2 *Weighted Stacking with Supervised Meta-Models*

We previously proposed HUE-AD [30], an equal-weight voting ensemble of four heterogeneous detectors. Our empirical analysis showed that a weighted voting approach outperforms HUE-AD. However, determining the voting system’s optimal weights requires a training dataset. Drawing inspiration from Feature Weighted Linear Stacking (FWLS) [40, 94], we developed a two-stage weighted stacking ensemble with unsupervised weak HS-AD methods in the first stage and a supervised machine learning (ML) method in the second stage to compute the weights. For the first stage, as there are limited annotated HS image datasets, a supervised ML model may suffer from under-fitting, so we opted to use unsupervised HS-AD methods. For the second stage, as learning the appropriate weights to assign to multiple inputs requires much less training data than learning features from images, we can utilize a supervised ML method. Initially, we explored Logistic Regression [163] and decision trees [164] to determine the weights in the second stage. Ultimately, we opted for random forest [165] due to its promising results. A random forest is a meta-estimator that fits several decision tree classifiers on various sub-samples of the dataset. These sub-samples are generated by randomly drawing from the dataset with replacement. It uses averaging over the predictions from the decision trees to improve the predictive accuracy and control over-fitting.

4.3.3 *Greedy Base-Model Selection Strategy*

Additionally, our empirical analysis showed that the combination of base models that yields the best outcomes varies based on the given scenario. As a result, we focus on identifying a more systematic and efficient method for base model selection in a weighted voting ensemble. Figure 20 shows the overall flowchart of the GE-AD algorithm, and our proposed greedy search is depicted in Figure 21.

The GE-AD algorithm uses consistent stratified 2-fold cross-validation. K -fold cross-validation is a widely used technique to assess the accuracy of a model [166–168]. It splits the training dataset into K subsets or folds. The model is trained using $K - 1$ of the

folds as training data, and the resulting model is validated on the remaining part of the data. The model is trained and evaluated K times, with each fold serving as the validation set in turn and being used for validation exactly once. The performance metrics obtained from each fold are then averaged to estimate the model's generalization performance. In our approach, 2-fold cross-validation divides the pixels in our data into 50–50 split training and testing datasets and evaluates the performance of our GE-AD algorithm twice. Our evaluation found that getting stuck in a local maxima can hinder searching for the best methods. We solved this issue by running the search algorithm five times. We included this 5×2 -fold cross validation as a part of the GE-AD algorithm to ensure that we find the best combinations. Out of these ten sets of results, GE-AD identifies which combinations survived the most. If there is a tie, it uses first the average test score and then the average training score to break the tie. For our ensemble fusion, we have considered one HS unmixing algorithm, FCLS, to generate an abundance map (see Section 2.3) and nine AD algorithms, including AED, CSD, FCBAD, GMRX, KIFD, KRX, LSUNRSORAD, Median AD, and RX.

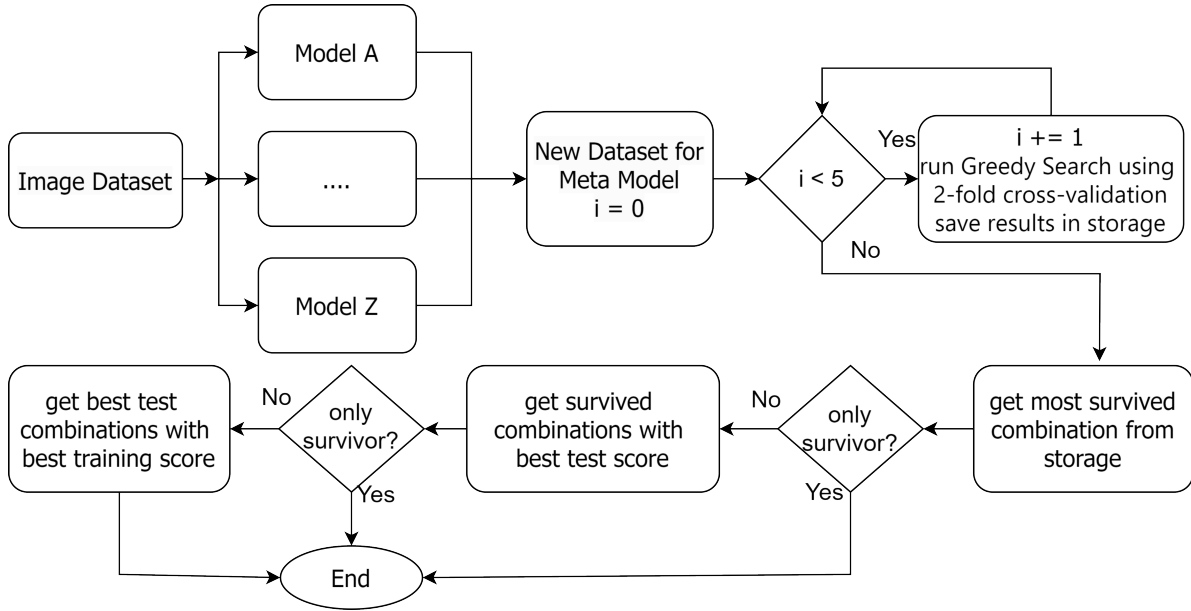


Figure 20: Overall flowchart of GE-AD to find the best AD methods for ensemble fusion.

In Figure 21, the k_best variable defines the highest number of base models to be used in combination. We varied the k_best variable to determine how many AD methods to

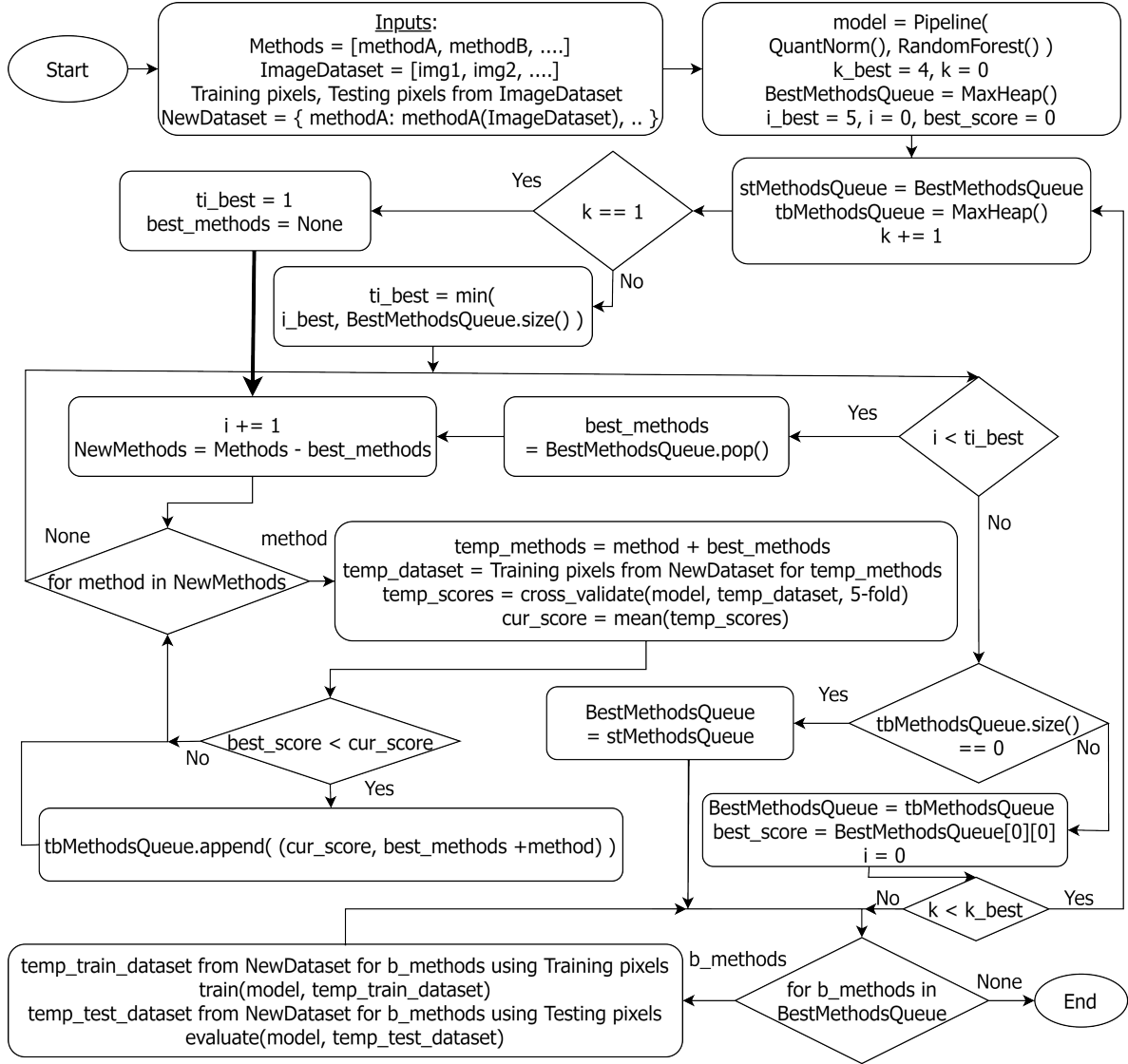


Figure 21: Flowchart of the greedy search to find the best AD methods for ensemble fusion.

include in the ensemble. We found that using four AD methods provides the best results within the constraints of computation complexity. Thus, we set four as the value for k_best . We first applied the AD algorithms to the image data to get each method's raw anomaly detection output. Then, we used these detection outputs as input features to train the ML model pipeline, which contains quantile normalization and random forest. The greedy search uses 5-fold cross-validation to compute the average F1-macro score of the trained ML model which fuses the output from the selected methods and stores the scores in a Priority Queue (Max Heap). Initially,

the top i_best AD methods are selected individually from the priority queue. We varied the i_best variable to determine how many combinations to pick up from the priority queue. We found that using the top five unique performers is enough to provide better results. Thus, we set i_best to be five. Then, in each subsequent round, one more method is added to the set found in the previous iteration. The methods (or combinations of them) that perform better than the best score of the last round are stored again in the queue, and iteration continues. As the number of methods reaches k_best or no better result is found, the greedy algorithm stops its search. We then train the ML pipeline on the training dataset and evaluate the performance using the testing dataset (both datasets are created from the selected methods' outputs). As GE-AD model trained on one dataset may differ from that trained on other datasets, we may identify each of these distinct GE-AD models by a different name.

4.3.4 Computational Complexity Analysis

Now, we analyze the time complexity of the greedy search algorithm. The algorithm begins by evaluating the performance of the entire set of n AD methods. The priority queue greedily selects only the best-performing combinations. Because ML model training time is orders of magnitude higher than the priority queue's enqueueing and dequeuing time, we do not consider the latter in the complexity analysis.

Thus, the runtime of the first round is $\mathcal{O}(n)$ as n AD methods are evaluated. In the second round, i_best best-performing methods were dequeued, each combined with one of the remaining $(n - 1)$ methods for training and evaluation.

Thus, the runtime of the second round is $\mathcal{O}(i_best \times (n - 1))$ for model training computation as up to $i_best \times (n - 1)$ AD methods are evaluated and enqueued into the priority queue. Then, in the third round, the i_best best-performing combinations of two methods are dequeued, each combined with one of the remaining $(n - 2)$ methods for training and evaluation. A table with all combinations is kept for time optimization and to avoid recomputation. Eventually, k_best number of methods are chosen. Then, $n + i_best \times (n - 1) + \dots + i_best \times (n - k_best + 1)$ training and

evaluations are performed resulting in time complexity of

$$\mathcal{O}(n + i_{best} \times (n - 1) + i_{best} \times (n - 2) + \dots + i_{best} \times (n - k_{best} + 1))$$

or, $\mathcal{O}(n)$, when i_{best} and k_{best} are small constants.

On the other hand, a brute-force search trains and evaluates $C(n, 1) + C(n, 2) + \dots + C(n, k_{best})$ combinations, with a time complexity of

$$\mathcal{O}(C(n, 1) + C(n, 2) + \dots + C(n, k_{best}))$$

or, $\mathcal{O}(C(n, k_{best})) = \mathcal{O}(\frac{n!}{(n - k_{best})!k_{best}!}) \sim \mathcal{O}(n^{k_{best}})$, when k_{best} is a small constant.

4.4 Unsupervised Extension of GE-AD (UGE-AD)

While GE-AD leverages limited labeled data to learn an effective stacking function, many hyperspectral anomaly detection scenarios lack any form of annotation. To address this constraint, we extend the GE-AD framework to an unsupervised setting by replacing the supervised meta-model with an unsupervised probabilistic model.

We have utilized the same greedy search 4.3.3 approach to find suitable HS-AD methods as the base model. We considered the Reed–Xiaoli (RX) [19, 66] detector, Attribute and Edge-Preserving Filters (AED) [23], Cluster-based anomaly detector (CBAD) [70], Complementary Subspace Detector (CSD) [69], Fuzzy clustering-based anomaly detector (FCBAD) [71], Gaussian mixture RX (GM-RX) [36] anomaly detector, Kernel iForest (KIFD) [38] anomaly detector, Local Summation Unsupervised Nearest Regularized Subspace with an Outlier Removal Anomaly Detector (LSUNRSORAD) [39], Median RX (MD-RX) anomaly detector, Subspace-RX (SSRX) [68], Windowed RX (WIN-RX).

We considered a similar idea to Wang et al. [87] and Fatemifar et al. [88] and investigated multiple unsupervised methods as meta-model. We used the Gaussian mixture model for the meta-model. We also investigated Forest; however, it did not perform well.

We have considered the proposed methodology of Yang et al. [86]. Instead of directly passing through the HSI data, we utilized PCA to find ten PCs as a pass-through to the meta-model to improve the ROC-AUC score.

For our first evaluation, we replaced the Random Forest with an unsupervised Mixture Model (GMM) as the meta-model.¹ In our evaluation, we found that it has an AUC ROC score of 0.943 ± 0.088 and F1-macro score of 0.548 ± 0.048 among all datasets (see Section 4.7). We plan to investigate other unsupervised models in the future. We also plan to apply all options we used in improving our supervised GE-AD to this new unsupervised GE-AD (UGE-AD) model and investigate their impacts.

4.5 Incorporating Spectral Information into GE-AD (mGE-AD)

As GE-AD uses only the results from base models, the spatial and spectral information from an HS image is not visible. GE-AD can compute different weights for different datasets based on features from the input data. However, we cannot feed all the channels to the our model due to two concerns: (a) this approach may hinder our model from learning essential features from the base models; and (b) the curse of dimensionality leads to increased computational complexity, overfitting, and spurious correlations. As an initial investigation, we modified GE-AD (mGE-AD) to take thirty random spectral channels and four other base methods as input. In our evaluation, we found that it has an AUC ROC score of 0.593 ± 0.070 and F1-macro score of 0.548 ± 0.048 among all datasets (see Section 4.7). These results indicate that naive inclusion of raw spectral channels does not necessarily improve ensemble performance, underscoring the need for principled dimensionality reduction and feature selection strategies.

4.6 Evaluation of the Proposed GE-AD Framework

This section presents a comprehensive experimental evaluation of the proposed ensemble-based hyperspectral anomaly detection frameworks, with a primary focus on validating

¹We also tested the isolation tree, which did not work.

the effectiveness of the Greedy Ensemble Hyperspectral Anomaly Detection (GE-AD) method. The experiments are designed to assess whether greedy base-method selection and probabilistic ensemble fusion can consistently improve detection performance over individual anomaly detectors and existing ensemble approaches.

We first evaluate GE-AD against its constituent base anomaly detectors and a knowledge-guided ensemble baseline to quantify the benefits of weighted stacking and greedy selection. Performance is analyzed using both threshold-dependent and threshold-independent metrics, complemented by visual inspection of detection maps and formal statistical significance testing. We then extend the evaluation to compare GE-AD with state-of-the-art ensemble anomaly detection methods reported in the literature.

To examine the framework’s behavior under limited or no supervision, we further evaluate the unsupervised extension of GE-AD (UGE-AD) across multiple benchmark datasets, analyzing its stability, generalization characteristics, and sensitivity to base-method selection. Finally, we report results for the modified GE-AD variant incorporating spectral information (mGE-AD), highlighting both its potential and limitations. Collectively, these experiments provide insight into the strengths, trade-offs, and practical applicability of greedy ensemble design for hyperspectral anomaly detection.

4.6.1 Comparison with Individual AD Methods

We evaluate our GE-AD algorithm using consistent stratified 5×2 -fold cross-validation. 2-fold cross-validation randomly selects 50% of the pixels as training data and the other 50% for testing.

In this evaluation with ABU-Airport dataset, the GE-AD algorithm selected the Abundance, AED, FCBAD, and KRX methods as the base models for the ensemble. Subsequently, the individual predictions from these selected algorithms were supplied to the meta-model for the final ML model training and detection. We trained a model with seed 842, which achieved a training F1-macro score of 0.81 and a test F1-macro score of 0.86, a higher

F1-macro score than any of the methods used in the ensemble. We are calling this model trained on the ABU dataset GE-AD model-1. GE-AD model-1 produces fewer false positives and creates a clear detection map with all the targets at least partially detected, as shown in Figures 22 from the ABU dataset.

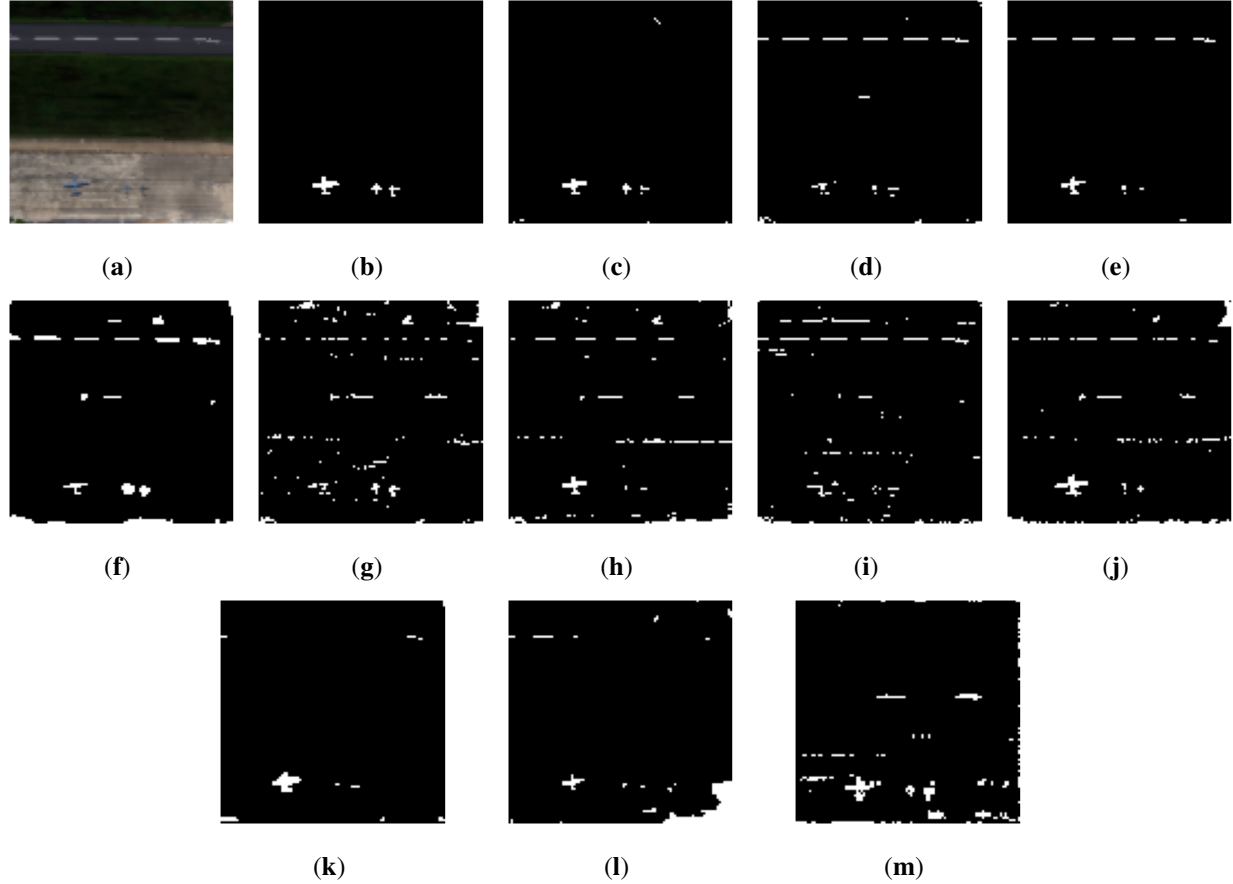


Figure 22: Visual comparison showing better performance of our proposed ensemble method (GE-AD) compared to other methods using ABU-IV data. (a) RGB, (b) ground-truth, (c) GE-AD ($F1 = 0.856$), (d) HUE-AD ($F1 = 0.649$), (e) Abundance ($F1 = 0.735$), (f) AED ($F1 = 0.610$), (g) KIFD ($F1 = 0.587$), (h) KRX ($F1 = 0.604$), (i) LSUNRSORAD ($F1 = 0.569$), (j) FCBAD ($F1 = 0.631$), (k) ERRX MF ($F1 = 0.666$), (l) ERCRD ($F1 = 0.545$), and (m) SED ($F1 = 0.592$). Please note that, as the implementations of algorithms for Figures (k–m) were not publicly available, we binarized the original outputs from the authors’ papers [84, 86, 87], respectively.

The null hypothesis of the Wilcoxon signed-rank test [169] is that the rank sums of the control algorithm **are not significantly less** than those of the other base models used in the ensemble. If the p -value is less than the significance level (0.05), the null hypothesis can be rejected, and the result is statistically significant. We ran our tests ten times and used all the ABU

results to run significance tests. In Table 5, GE-AD has the lowest rank, and all the methods have a p -value less than 0.05; thus, the null hypothesis can be rejected, and our process is better and differs significantly from the other methods.

For comparative analysis, we applied the Friedman rank test [170], which measures the significance of the performance differences among the detection methods considered. The null hypothesis of the Friedman rank test is that the rank sums **do not differ** significantly between the three or more “paired” samples, and they are identical. As in the case of the Wilcoxon test, if the p -value is less than the significance level (0.05), the null hypothesis can be rejected, and the result is statistically significant. The Friedman test for F1-macro is one-tailed with a p -value of 0.05, and ROC AUC is two-tailed [171–173] with a p -value of $0.05/2$. Table 5 shows our method is better and differs significantly from the other methods for the F1-macro metric, with the best average rank of 1.07 (lower is better), and the p -value is 3.397^{-26} (numerical zero), which is less than 0.05.

In addition to the F1-macro metric, we also determined the ROC AUC scores [108] for each algorithm as the authors of other papers evaluated their work using this metric. The results are summarized in Figure 23. Note that even though GE-AD model-1 resulted in outputs that are visually closer to the ground-truth data, it failed to achieve a higher ROC AUC score compared to all the methods used in the ensemble. This demonstrates that a single metric is not sufficient to assess the performance.

As shown in Figures 22 and 23, the proposed method outputs results that are visually closer to the ground-truth data with a lower ROC AUC score; we still conducted the statistical tests on the ROC AUC score and obtained an interesting result. Our evaluation shows that our method is statistically significant for the ROC AUC score over the ten runs. The p -value of AED in the one-sided Wilcoxon signed-rank test is 0.038011, much higher than the others but still less than the significance level (0.05). Table 6 shows that the rank of the proposed method is the lowest, and the p -value in the Friedman rank test is 3.9^{-22} , which is less than $0.05/2$.

Table 5: Comparison of F1-macro scores showing our proposed ensemble method (GE-AD) performs better compared to HUE-AD, and individual methods using the ABU dataset.

Method	ABU-I	ABU-II	ABU-III	ABU-IV	Avg. F1 -macro Score	Wilcoxon's p -Value [169]	Avg. Rank	Friedman's p -Value [170]
GE-AD [31]	0.887	0.878	0.875	0.769	0.852 ± 0.087		1.07	
HUE-AD [30]	0.793	0.770	0.751	0.649	0.741 ± 0.058	1.9262×10^{-8}	2.62	
Abundance [15]	0.670	0.741	0.760	0.736	0.727 ± 0.038	3.033×10^{-8}	3.10	
AED [23]	0.781	0.646	0.697	0.610	0.683 ± 0.065	1.7847×10^{-8}	4.40	3.397×10^{-26}
FCBAD [71]	0.606	0.603	0.591	0.621	0.605 ± 0.014	1.7847×10^{-8}	6.75	
KIFD [38]	0.694	0.682	0.695	0.587	0.664 ± 0.046	1.7847×10^{-8}	5.05	
KRX [35]	0.559	0.658	0.659	0.604	0.620 ± 0.043	1.7847×10^{-8}	6.42	
LSUNRS ORAD [39]	0.695	0.608	0.661	0.568	0.633 ± 0.049	1.7847×10^{-8}	6.58	

Table 6: Performance results of all methods regarding AUC metric (%) along with Friedman rank using the ABU-airport dataset.

Method	ABU-I	ABU-II	ABU-III	ABU-IV	Avg. AUC Score	Wilcoxon's p -Value [169]	Avg. Rank	Friedman's p -Value [170]
GE-AD	0.987	0.980	0.975	0.978	0.980 ± 0.018		2.00	
Abundance	0.620	0.733	0.834	0.805	0.748 ± 0.085	1.7847×10^{-8}	7.15	
AED	0.992	0.980	0.975	0.932	0.970 ± 0.025	3.8011×10^{-2}	2.48	3.901×10^{-22}
FCBAD	0.900	0.896	0.889	0.978	0.916 ± 0.038	1.4139×10^{-7}	4.28	
KRX	0.914	0.971	0.951	0.975	0.953 ± 0.025	2.7606×10^{-5}	3.55	

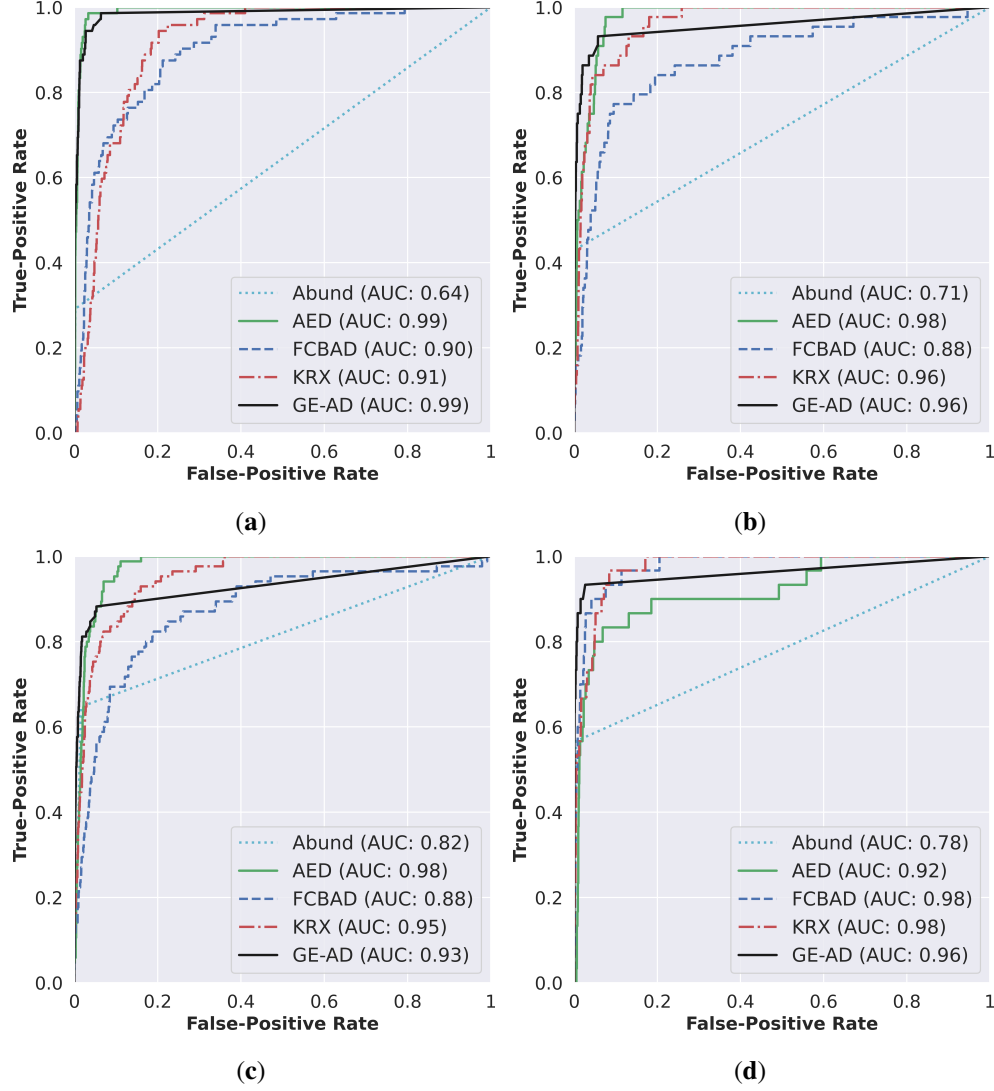


Figure 23: Detection accuracy evaluation through ROC curve and AUC scores (shown in the legend) of (a) ABU-I, (b) ABU-II, (c) ABU-III, and (d) ABU-IV.

4.6.2 Comparison with Baseline SOTA Ensemble Methods

We broaden our comparative analysis by elaborating on the performance results of several existing ensemble AD methods. Our GE-AD model-1 struggled to achieve a higher ROC AUC score compared to the other individual methods. Still, it achieved a considerably similar ROC AUC score to the other SOTA ensemble anomaly detection algorithms, as shown in Table 7. However, our previous work [30] discussed how the F1-macro score is a better-suited metric

to evaluate anomaly detection tasks as a small number of incorrect or correct predictions can significantly change the AUC score. Additionally, the AUC score does not consider the impact of false positives. Hence, this study has included the ROC AUC score, the F1-macro score, and their corresponding visual comparisons. Our method outperforms the other ensemble methods both visually and in terms of the F1-macro score, as shown in Table 7 and Figure 22, while it provides comparable scores for the ROC AUC.

Table 7: ABU-IV quantitative performance (as reported in the authors’ papers [84, 86, 87]) comparison between our proposed ensemble method (GE-AD) and other ensemble methods. Here, the best scores are shown in bold.

	GE-AD	ERRX MF [86]	ERCRD [84]	SED [87]
ROC AUC score	0.9632	0.9970	0.9533	0.9980
F1-macro score	0.856	0.666	0.545	0.592

4.6.3 Public Dataset Benchmark Evaluation: Hydice Urban

In this evaluation, the GE-AD algorithm selected the FCBAD and LSUNRSORAD methods as the parallel components for the ensemble. Subsequently, the individual predictions from these selected algorithms were supplied to the meta-model for the final ML model training and detection. We trained a model with seed 577 and achieved a training F1-macro score of 0.81 and a test F1-macro score of 0.94, a higher F1-macro score than any of the methods used in the ensemble. This model produced fewer false positives and created a clear detection map with all the targets at least partially detected, as shown in Figure 24 from the HYDICE-Urban dataset. We ran our tests ten times using the HYDICE-Urban dataset and reported the F1-macro scores in Table 8. Our method performed better than any of the methods used in the ensemble.

4.6.4 Public Dataset Benchmark Evaluation: Salinas

In this evaluation, the GE-AD algorithm selected the LSUNRSORAD methods as the only component for the ensemble. Subsequently, the individual predictions from these selected algorithms were supplied to the meta-model for the final ML model training and detection.

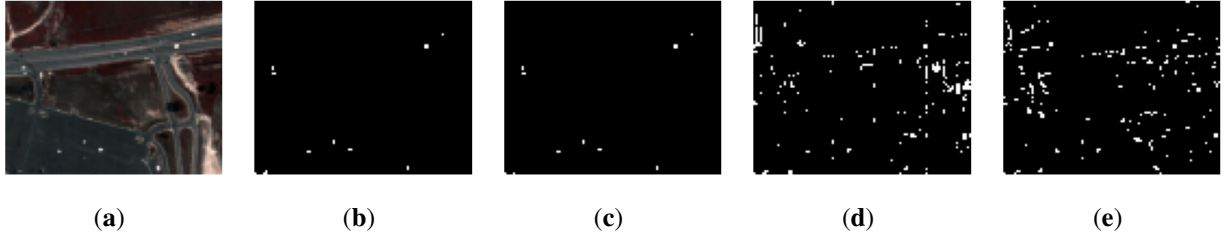


Figure 24: Visual comparison showing better performance of our proposed ensemble method (GE-AD) compared to other methods using HYDICE-Urban data. (a) RGB, (b) ground-truth, (c) GE-AD (F1 = 0.944), (d) FCBAD (F1 = 0.575), and (e) LSUNRSORAD (F1 = 0.567).

Table 8: Comparison between our proposed ensemble method (GE-AD) and individual methods using the HYDICE-Urban dataset.

Method	Avg. F1-macro Score	Wilcoxon's p -Value [169]	Avg. Rank
GE-AD	0.797 ± 0.048		1.00
FCBAD	0.570 ± 0.012	9.7656×10^{-4}	2.20
LSUNRSORAD	0.566 ± 0.006	9.7656×10^{-4}	2.80

We trained a model with seed 529 and achieved a training F1-macro score of 0.98 and a test F1-macro score of 0.98, a higher F1-macro score than any of the methods used in the ensemble. This model produced fewer false positives and created a clear detection map with all the targets partially detected, as shown in Figure S10. We ran our tests using the Salinas dataset ten times and reported the F1-macro scores in Table 9. Our method performed better than any of the methods used in the ensemble.

Table 9: Comparison between our proposed ensemble method (GE-AD) and individual methods using the Salinas dataset.

Method	Avg. F1-macro Score	Wilcoxon's p -Value [169]	Avg. Rank
GE-AD	0.917 ± 0.108		1.00
LSUNRSORAD	0.548 ± 0.003	9.7656×10^{-4}	2.00

4.6.5 Public Dataset Benchmark Evaluation: San-Diego Airport

In this evaluation, the GE-AD algorithm selected the Abundance, GMRX, KIFD, and KRX methods as the parallel components for the ensemble. Subsequently, the individual predictions from these selected algorithms were supplied to the meta-model for the final ML model training and detection. We trained a model with seed 306 and achieved a training F1-macro score of 0.86 and a test F1-macro score of 0.89, a higher F1-macro score than any of the methods used in the ensemble. This model produced fewer false positives and created a clear detection map with all the targets at least partially detected, as shown in Figures S11 and S12. We ran our tests using the San-Diego airport dataset ten times and reported the F1-macro scores in Table 10. Our method performed better than any of the methods used in the ensemble.

Table 10: Comparison between our proposed ensemble method (GE-AD) and individual methods using the San-Diego Airport dataset.

Method	San-Diego-01	San-Diego-02	Avg. F1 -macro Score	Wilcoxon's p -Value [169]	Avg. Rank
GE-AD	0.883	0.910	0.896 ± 0.050		1.00
Abundance	0.790	0.655	0.722 ± 0.068	9.5367×10^{-7}	2.20
GMRX	0.552	0.497	0.524 ± 0.028	9.5367×10^{-7}	5.00
KIFD	0.774	0.642	0.708 ± 0.067	9.5367×10^{-7}	2.80
KRX	0.714	0.627	0.671 ± 0.046	9.5367×10^{-7}	4.00

4.6.6 Private Dataset Benchmark Evaluation: Arizona

The signal-to-noise ratio (SNR) analysis of the hyperspectral images obtained from a new sensor is necessary to ensure the sensor has excellent overall sensitivity and spectral resolution to support capturing and preserving the most critical information in the scene. It increases the efficacy of preprocessing techniques such as geo-rectification [44]. Our project collected the Arizona dataset with a new sensor; therefore, custom preprocessing and initial SNR analyses were required. We applied these preprocessing methods before using them for unmixing and anomaly detection. The details of these preprocessing techniques are discussed in Supplement-I.

We trained a separate GE-AD model using the Arizona dataset. In this evaluation, the GE-AD algorithm selected the FCBAD, KIFD, KRX, and LSUNRSORAD methods as the parallel components for the ensemble. Subsequently, the individual predictions from these selected algorithms were supplied to the meta-model for the final ML model training and detection. We trained a model with seed 319 and achieved a training F1-macro score of 0.81 and a test F1-macro score of 0.86, a higher F1-macro score than any of the methods used in the ensemble. This model produced fewer false positives and created a clear detection map with all the targets at least partially detected, as shown in Figures 25 and S13–S16 from the Arizona dataset. We ran our tests using the Arizona dataset ten times, and the comparison between our proposed ensemble and other individual methods using the Arizona dataset is shown in Figure 25 and Table 11. Our method performed better than any of the methods used in the ensemble.

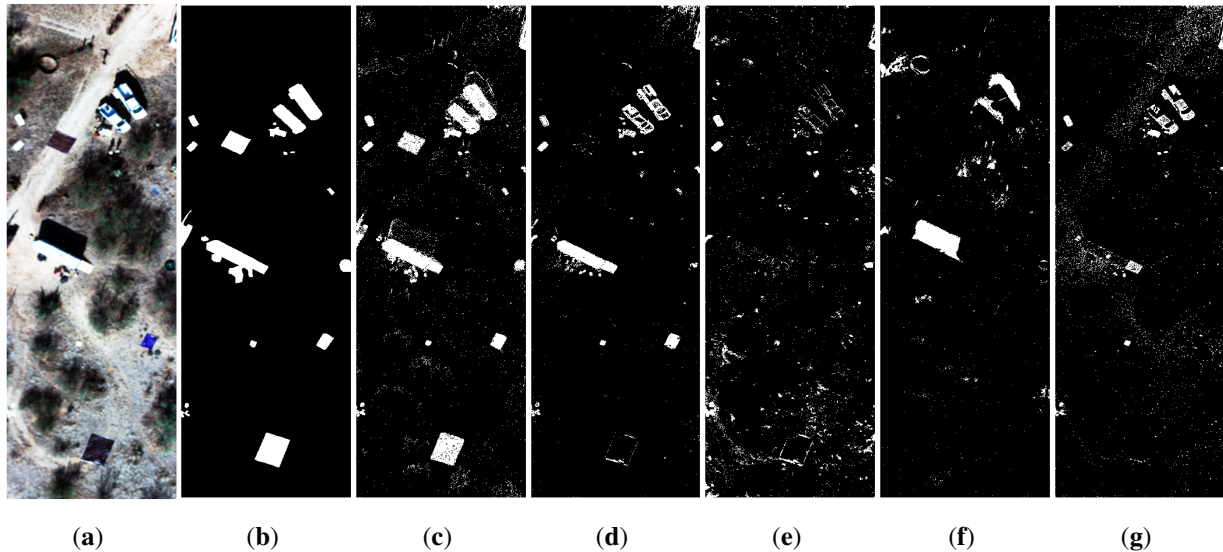


Figure 25: Visual comparison showing better performance of our proposed ensemble method (GE-AD) compared to other methods using Arizona image-III data. (a) RGB, (b) ground-truth, (c) GE-AD ($F1 = 0.858$), (d) FCBAD ($F1 = 0.773$), (e) KIFD ($F1 = 0.583$), (f) KRX ($F1 = 0.494$), and (g) LSUNRSORAD ($F1 = 0.639$).

4.6.7 Discussion

The experimental results demonstrate that the proposed GE-AD framework can effectively leverage complementary information from multiple unsupervised hyperspectral

Table 11: Comparison between our proposed ensemble method (GE-AD) and individual methods using the Arizona dataset.

Method	AZ-I	AZ-II	AZ-III	AZ-IV	AZ-V	Avg. F1 -macro Score	Wilcoxon's p -Value [169]	Avg. Rank	Friedman's p -Value [170]
GE-AD	0.770	0.799	0.882	0.816	0.859	0.825 ± 0.047		1.00	
FCBAD [71]	0.587	0.614	0.774	0.542	0.812	0.666 ± 0.107	3.7785×10^{-10}	2.40	
KIFD [38]	0.570	0.569	0.584	0.505	0.578	0.561 ± 0.029	3.7785×10^{-10}	3.96	3.465×10^{-29}
KRX [35]	0.633	0.675	0.494	0.514	0.503	0.564 ± 0.075	3.7785×10^{-10}	3.40	
LSUNRSO -RAD [39]	0.520	0.571	0.640	0.498	0.572	0.560 ± 0.049	3.7785×10^{-10}	4.24	

anomaly detection methods through greedy base-model selection and supervised ensemble fusion. By utilizing limited labeled data to guide ensemble construction, GE-AD consistently identifies combinations of base detectors that outperform individual methods across multiple evaluation scenarios. These results support the central motivation of this work: that no single anomaly detector is universally optimal and that aggregating diverse background modeling assumptions can yield more reliable detection performance.

The evaluation further shows that the performance of individual HS-AD methods varies substantially across datasets and scenes, reflecting their sensitivity to underlying background characteristics (Tables 5 and 8–11). In contrast, GE-AD exhibits more stable behavior by combining outputs from multiple detectors, thereby reducing reliance on any single modeling assumption. Statistical significance tests conducted on the ABU dataset confirm that the observed improvements achieved by GE-AD over its constituent base methods are statistically meaningful under the selected evaluation criteria.

All base detectors considered in this study are well-established classical HS-AD methods with publicly available implementations and clearly defined assumptions. GE-AD operates by fusing information from these methods rather than replacing them, producing a combined detection result that consistently exceeds the performance of any individual input detector. As a result, improvements in the performance of individual anomaly detection algorithms are

expected to translate directly into further gains for GE-AD, provided that the detectors contribute complementary information.

The generalization capability of GE-AD was examined by applying a model trained on the ABU dataset to the San-Diego airport dataset. The resulting F1-macro scores of 0.822 and 0.826 for the two San-Diego scenes indicate that a GE-AD model trained on one dataset can remain effective when applied to a different dataset with similar scene characteristics and sensor properties. These results suggest that GE-AD can generalize across related scenarios when limited labeled data are available.

An ablation study confirms that the greedy search procedure plays a critical role in GE-AD by adaptively selecting base methods that maximize the chosen scoring function. However, because the greedy search in this work optimizes the F1-macro score, GE-AD does not necessarily achieve the highest ROC-AUC scores relative to all baseline methods. This behavior reflects the inherent trade-offs between threshold-dependent and threshold-independent metrics and highlights the importance of aligning the greedy search objective with the intended operational requirements. The scoring function used by the greedy algorithm can be modified as needed to emphasize alternative performance criteria.

Noise robustness experiments indicate that GE-AD performs reliably under low to moderate noise levels but degrades when base detectors respond inconsistently to severe noise corruption. This limitation arises from the dependence of ensemble performance on the reliability of its constituent detectors. Although no data augmentation or noise-aware training strategies were employed in this study, such approaches represent promising directions for improving robustness in future work. For the Arizona dataset, manual preprocessing was applied to mitigate noise effects; automating this preprocessing decision remains an open challenge.

Finally, the evaluation highlights limitations of relying on a single metric to assess hyperspectral anomaly detection performance. ROC-AUC can be insensitive to false positives in highly imbalanced settings, while F1-macro may provide overly optimistic assessments when only a subset of anomalies is detected. These observations motivate future investigation

into alternative evaluation metrics and objective functions for greedy search, such as Matthews Correlation Coefficient (MCC) [174], which may better balance these competing considerations.

During the greedy search process, we observed that GE-AD can become trapped in local optima when using a single instance of 2-fold cross-validation. Repeating the cross-validation multiple times and selecting base-method combinations that consistently perform well across runs mitigated this issue. In this work, five repetitions of 2-fold cross-validation were sufficient to obtain stable results, and we recommend this practice for reliable ensemble construction.

The GE-AD framework employs an ML pipeline that integrates Quantile normalization with a Random Forest meta-classifier. This design simplifies training, evaluation, and deployment by encapsulating preprocessing and model inference within a unified framework. Although the normalization step may be redundant for some models, its inclusion ensured consistent evaluation across different meta-models explored during development.

Overall, the results demonstrate that GE-AD can dynamically adapt its ensemble composition based on dataset-specific characteristics when limited labeled data are available. Different datasets led to different base-method selections, reflecting changes in background structure and scene content. This adaptive behavior enables GE-AD to construct scenario-specific ensembles that better capture the underlying characteristics of the data.

4.7 Evaluation of the Unsupervised GE-AD Extension

This section evaluates the proposed Unsupervised Greedy Ensemble Anomaly Detection (UGE-AD) framework under settings where no labeled data are available for training a supervised meta-model. Consistent with the goal of assisting human-in-the-loop decision-making, the evaluation emphasizes accurate identification of anomalous pixels while minimizing missed detections.

4.7.1 *Evaluation Protocol and Metrics*

Performance is primarily assessed using the area under the receiver operating characteristic curve (ROC-AUC), a threshold-independent metric commonly used in hyperspectral anomaly detection and suitable for highly imbalanced data scenarios [108–110]. Although F1-macro scores are reported for completeness in later comparisons, ROC-AUC is used as the primary optimization criterion for the greedy search in the unsupervised setting.

4.7.2 *Single-Dataset Greedy Search Evaluation*

In the first evaluation, a single dataset was used during the greedy search to identify suitable base anomaly detectors for UGE-AD. The greedy search was configured to maximize the ROC-AUC score on the selected dataset, after which the resulting ensemble was evaluated across multiple datasets.

Table 12 reports the average ROC-AUC scores obtained using this procedure. The first column indicates the dataset used during the greedy search, while the second column lists the selected base methods. Entries along the diagonal correspond to training performance, whereas off-diagonal entries reflect cross-dataset generalization performance. Figure 26 provides a visual comparison of these results.

4.7.3 *Multi-Dataset Greedy Search Evaluation*

To evaluate the robustness of UGE-AD under more diverse conditions, a second experiment was conducted using twenty-two scenes drawn from six public and one private benchmark datasets. In this setting, datasets were randomly assigned to the greedy search such that at least 50% of the available scenes were used for base-method selection.

Across five runs with 2-fold cross-validation, the greedy search optimized the ROC-AUC score and consistently selected AED, FCBAD, GM-RX, and KIFD as base methods. These methods, along with ten principal components of the hyperspectral data, were used as inputs to a Gaussian Mixture Model (GMM) meta-model. The resulting UGE-AD achieved an average

Table 12: Average ROC-AUC scores of our newly proposed Unsupervised ensemble UGE-AD over various datasets. The first column shows the dataset used in the greedy search to find suitable base methods, and the second column shows those base methods. Where the dataset names in rows and columns match, those cells show the training score after the greedy search.

		HYDICE -Urban	Salinas	San -Diego	ABU -Airport	ABU -Beach	ABU -Urban	Arizona
HYDICE -Urban	WIN-RX	0.997	0.951	0.962	0.943	0.963	0.964	0.802
Salinas	LSUNRSORAD, RX	0.995	0.999	0.965	0.949	0.971	0.971	0.780
San -Diego	GM-RX	0.997	0.997	0.964	0.934	0.961	0.962	0.847
ABU -Airport	AED, GM-RX, KIFD, LSUNRSORAD	0.995	1.000	0.971	0.962	0.985	0.966	0.712
ABU -Beach	KIFD, LSUNRSORAD, RX, WIN-RX	0.996	0.997	0.965	0.960	0.973	0.976	0.766
ABU -Urban	AED, KIFD, LSUNRSORAD, RX	0.996	1.000	0.970	0.962	0.985	0.971	0.706
Arizona	CBAD, FCBAD, GM-RX, KIFD	0.994	0.999	0.972	0.948	0.967	0.966	0.779

ROC-AUC score of 0.943 ± 0.088 across seven datasets, as summarized in Table 13.

Table 13 summarizes the average and standard deviation of ROC-AUC scores across all evaluated datasets, providing a global view of detection performance. For completeness, Table 14 reports the corresponding F1-macro scores for ensemble-based methods, highlighting differences in threshold-dependent behavior.

4.7.4 Per-Dataset Quantitative Results

Detailed quantitative results are reported to examine performance variability across datasets and individual scenes. Tables 15–19 present per-scene ROC-AUC scores for individual

Table 13: Average and standard deviation of ROC-AUC scores of various methods computed over the ABU, Arizona, HYDICE-Urban, Salinas, and San-Diego datasets.

Methods	HYDICE -Urban	Salinas	San -Diego	ABU -Airport	ABU -Beach	ABU -Urban	Arizona	Average score
CBAD	0.970	0.390	0.611	0.753	0.681	0.466	0.566	0.634 ± 0.178
CSD	0.982	0.788	0.906	0.877	0.962	0.979	0.451	0.849 ± 0.175
FCBAD	0.994	0.841	0.964	0.910	0.973	0.904	0.508	0.870 ± 0.156
GM-RX	0.957	0.546	0.388	0.599	0.901	0.651	0.497	0.648 ± 0.194
MD-RX	0.986	0.817	0.923	0.869	0.961	0.977	0.453	0.855 ± 0.174
RX	0.985	0.807	0.915	0.886	0.962	0.979	0.451	0.855 ± 0.175
SSRX	0.747	0.848	0.802	0.758	0.925	0.848	0.589	0.788 ± 0.099
WIN-RX	0.749	0.798	0.705	0.807	0.934	0.857	0.541	0.770 ± 0.116
AED	0.979	0.956	0.988	0.979	0.983	0.923	0.392	0.886 ± 0.203
KIFD	0.830	0.995	0.991	0.964	0.948	0.962	0.652	0.906 ± 0.116
LSUNRSORAD	0.994	1.000	0.945	0.964	0.989	0.973	0.353	0.888 ± 0.219
baseline HUE-AD (AED, KIFD, LSUNRSORAD)	0.962	1.000	0.988	0.980	0.986	0.987	0.411	0.902 ± 0.201
mGE-AD (GM-RX, KIFD, LSUNRSORAD, MD-RX), RF	0.927	0.997	0.939	0.907	0.965	0.970	0.472	0.882 ± 0.170
UGE-AD (AED, FCBAD, GM-RX, KIFD), GMM	0.995	0.999	0.970	0.960	0.980	0.968	0.731	0.943 ± 0.088

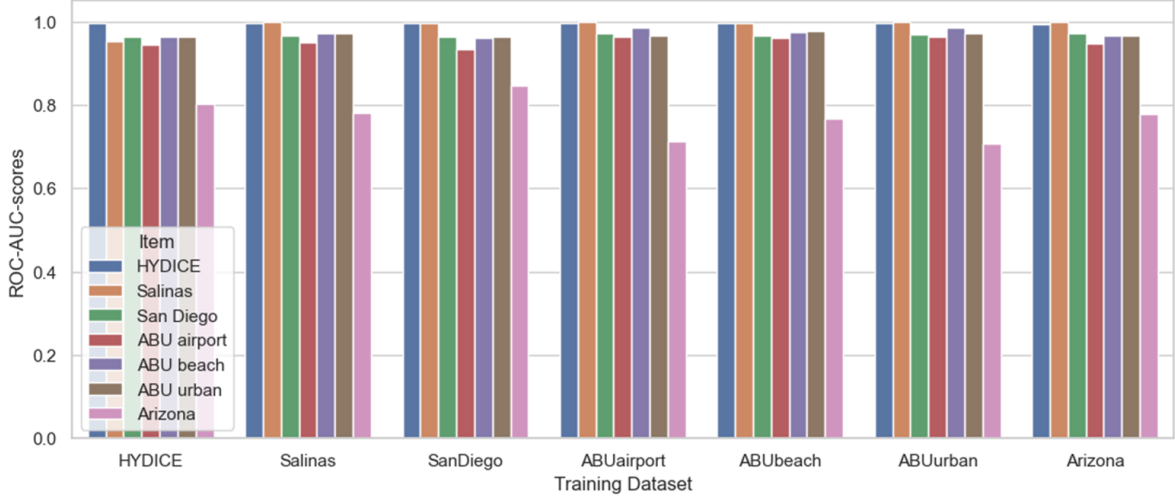


Figure 26: Average ROC-AUC scores of our UGE-AD over various datasets showing consistent performance against various training datasets. The X-axis shows the dataset used in the greedy search to find suitable base methods. The Y-axis shows the testing ROC AUC scores. Except where the dataset names in rows and columns match, those bars show the training score after the greedy search.

Table 14: Average and standard deviation of F1-macro scores of our ensemble methods computed over the ABU, Arizona, HYDICE-Urban, Salinas, and San-Diego datasets.

Methods	HYDICE -Urban	Salinas	San -Diego	ABU -Airport	ABU -Beach	ABU -Urban	Arizona	Average score
baseline HUE-AD	0.519	0.512	0.619	0.629	0.564	0.663	0.541	0.578 ± 0.055
mGE-AD	0.525	0.729	0.560	0.584	0.628	0.624	0.503	0.593 ± 0.070
UGE-AD	0.500	0.480	0.559	0.566	0.514	0.605	0.613	0.548 ± 0.048

anomaly detectors, baseline ensemble methods, and the proposed ensemble approaches.

Table 15 reports results for the HYDICE-Urban, Salinas, and San-Diego datasets. Tables 16, 17, and 18 present scene-level results for the ABU-Airport, ABU-Beach, and ABU-Urban datasets, respectively. Results for the Arizona dataset are summarized in Table 19. These tables allow direct comparison of ensemble behavior relative to individual detectors under diverse background conditions.

Table 15: Comparison of ROC-AUC scores between various methods using the HYDICE-Urban, Salinas, and San-Diego datasets

Methods	HYDICE-Urban	Salinas	San-Diego-01	San-Diego-02
CBAD	0.970	0.390	0.674	0.547
CSD	0.982	0.788	0.941	0.871
FCBAD	0.994	0.841	0.969	0.958
GM-RX	0.957	0.546	0.557	0.219
MD-RX	0.986	0.817	0.945	0.900
RX	0.985	0.807	0.940	0.889
SSRX	0.747	0.848	0.768	0.836
WIN-RX	0.749	0.798	0.656	0.753
AED	0.979	0.956	0.985	0.991
KIFD	0.830	0.995	0.992	0.989
LSUNRSORAD	0.994	1.000	0.980	0.910
baseline HUE-AD	0.962	1.000	0.993	0.982
mGE-AD	0.927	0.997	0.916	0.962
UGE-AD	0.995	0.999	0.976	0.964

4.7.5 Discussion

The experimental results indicate that UGE-AD can effectively aggregate complementary anomaly detectors in the absence of labeled data, producing competitive ROC-AUC performance across multiple datasets. Performance varies depending on the selected base methods and dataset characteristics, underscoring the sensitivity of unsupervised ensemble construction to the greedy search configuration.

Across several datasets, individual detectors such as AED, KIFD, and LSUNRSORAD frequently achieve strong performance, which is reflected in their repeated selection during greedy search. While the baseline HUE-AD ensemble performs competitively in many scenarios, UGE-AD demonstrates improved robustness in challenging cases such as the Arizona dataset,

Table 16: Comparison of ROC-AUC scores between various methods using the ABU-Airport dataset

Methods	ABU-airport-1	ABU-airport-2	ABU-airport-3	ABU-airport-4
CBAD	0.858	0.633	0.633	0.886
CSD	0.824	0.826	0.922	0.934
FCBAD	0.877	0.880	0.897	0.986
GM-RX	0.674	0.590	0.680	0.453
MD-RX	0.813	0.841	0.925	0.896
RX	0.822	0.840	0.929	0.953
SSRX	0.639	0.857	0.892	0.644
WIN-RX	0.774	0.797	0.949	0.708
AED	0.969	0.994	0.957	0.995
KIFD	0.940	0.979	0.961	0.975
LSUNRSORAD	0.972	0.959	0.966	0.959
baseline HUE-AD	0.970	0.989	0.973	0.988
mGE-AD	0.859	0.916	0.920	0.931
UGE-AD	0.933	0.971	0.948	0.986

where several individual detectors fail.

The results also show that access to additional information, such as principal components of the hyperspectral data, can help mitigate failure modes when base detector outputs are insufficient. However, UGE-AD generally underperforms compared to supervised GE-AD variants, reflecting the inherent limitations of unsupervised ensemble fusion.

Finally, the comparison between ROC-AUC and F1-macro scores highlights that no single metric fully captures anomaly detection performance under class imbalance. While ROC-AUC is suitable for evaluating ranking quality, F1-macro provides complementary insight into threshold-dependent behavior. These observations motivate future investigation into alternative evaluation metrics and objective functions that better align with operational requirements.

Table 17: Comparison of ROC-AUC scores between various methods using the ABU-Beach dataset

Methods	ABU-beach-1	ABU-beach-2	ABU-beach-3	ABU-beach-4
CBAD	0.571	0.603	0.664	0.885
CSD	0.982	0.911	1.000	0.956
FCBAD	0.987	0.955	1.000	0.948
GM-RX	0.986	0.701	1.000	0.916
MD-RX	0.980	0.910	1.000	0.955
RX	0.981	0.911	1.000	0.954
SSRX	0.980	0.906	0.982	0.831
WIN-RX	0.957	0.962	1.000	0.815
AED	0.997	0.955	1.000	0.979
KIFD	0.990	0.989	0.998	0.813
LSUNRSORAD	0.994	0.986	1.000	0.975
baseline HUE-AD	0.992	0.992	1.000	0.958
mGE-AD	0.962	0.980	0.981	0.935
UGE-AD	0.995	0.946	1.000	0.980

4.8 Summary

This chapter evaluated the proposed GE-AD framework through comprehensive experiments against individual anomaly detectors and baseline ensemble methods. The results demonstrate that greedy base-model selection combined with supervised ensemble fusion enables GE-AD to achieve consistent performance improvements under class imbalance. Statistical significance testing and cross-dataset evaluation further validate the effectiveness and stability of the proposed approach when limited labeled data are available. These findings establish GE-AD as a robust ensemble framework for hyperspectral anomaly detection and motivate its extensions and broader analysis presented in subsequent chapters.

Table 18: Comparison of ROC-AUC scores between various methods using the ABU-Urban dataset

Methods	ABU-urban-1	ABU-urban-2	ABU-urban-3	ABU-urban-4	ABU-urban-5
CBAD	0.391	0.155	0.613	0.598	0.573
CSD	0.991	0.994	0.950	0.989	0.969
FCBAD	0.992	0.851	0.941	0.847	0.891
GM-RX	0.760	0.549	0.778	0.765	0.402
MD-RX	0.991	0.994	0.939	0.989	0.970
RX	0.991	0.995	0.951	0.989	0.969
SSRX	0.928	0.989	0.748	0.786	0.791
WIN-RX	0.927	0.909	0.783	0.787	0.881
AED	0.998	0.796	0.998	0.840	0.984
KIFD	0.938	0.917	0.992	0.978	0.984
LSUNRSORAD	0.996	0.972	0.988	0.928	0.983
baseline HUE-AD	0.993	0.972	0.996	0.985	0.990
mGE-AD	0.980	0.935	0.985	0.968	0.981
UGE-AD	0.977	0.976	0.962	0.973	0.952

Table 19: Comparison of ROC-AUC scores between various methods using the Arizona dataset

Methods	Arizona 1	Arizona 2	Arizona 3	Arizona 4	Arizona 5
CBAD	0.510	0.374	0.556	0.742	0.648
CSD	0.335	0.410	0.601	0.264	0.645
FCBAD	0.367	0.432	0.662	0.350	0.727
GM-RX	0.576	0.271	0.457	0.760	0.423
MD-RX	0.334	0.410	0.602	0.271	0.648
RX	0.334	0.410	0.602	0.262	0.647
SSRX	0.564	0.578	0.661	0.458	0.682
WIN-RX	0.534	0.507	0.629	0.437	0.596
AED	0.329	0.326	0.492	0.307	0.504
KIFD	0.580	0.567	0.671	0.663	0.780
LSUNRSORAD	0.208	0.369	0.501	0.187	0.499
baseline HUE-AD	0.248	0.465	0.551	0.216	0.573
mGE-AD	0.427	0.290	0.622	0.382	0.639
UGE-AD	0.671	0.695	0.925	0.458	0.905

Chapter 5

Generalization of Hyperspectral Anomaly detection

This chapter addresses generalization limitations, in which the GE-AD ensemble demonstrated strong in-domain performance but degraded under cross-sensor and cross-scene shifts. Classical ensemble weighting implicitly assumes stationary feature distributions, resulting in suboptimal performance when applied to unseen datasets.

To overcome this limitation, we propose a unified MoE stacking ensemble architecture that integrates domain-specialized experts with domain-invariant representation learning. The framework combines (i) dataset-specialized FFNN-AD experts, (ii) a hyperspectral vision transformer backbone for shared feature extraction, (iii) a soft-gated Mixture-of-Experts routing mechanism, (iv) domain-adversarial training to promote invariant representations, and (v) a drift-aware test-time adaptation strategy.

The objective is to develop a generalizable hyperspectral anomaly detector that operates without domain labels at inference time while maintaining competitive performance across heterogeneous sensors and scenes.

5.1 Limitations of Classical Ensemble Methods

Generalization refers to a model’s ability to maintain predictive performance when applied to unseen data distributions. In hyperspectral anomaly detection, many approaches implicitly assume generalization by adapting to the test image during inference. However, such evaluation does not reflect realistic deployment conditions where models must operate without prior exposure to target-domain data.

Following the evaluation protocol adopted in [118], we assess generalization by training on a source dataset and testing directly on a different target dataset without retraining.

5.1.1 Cross-Scene Transfer: ABU-Airport to San-Diego

In our previous evaluation [31], GE-AD trained and tested on the San-Diego dataset achieved F1-macro scores of 0.883 and 0.910 on two images.

To evaluate cross-domain performance, we trained GE-AD solely on the ABU-Airport dataset and tested it on the San-Diego dataset without adaptation. The model achieved F1-macro scores of 0.822 and 0.826 (Fig. 27).

Although slightly lower than in-domain performance, these results demonstrate that GE-AD generalizes reasonably well between airport scenes captured under similar environmental and structural conditions.

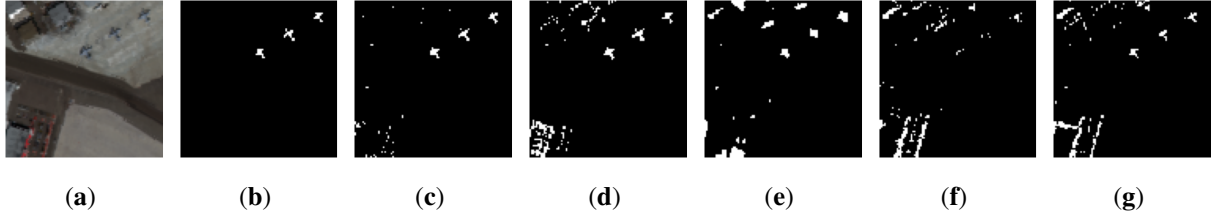


Figure 27: The visualization shows the generalization of our proposed ensemble method (GE-AD) trained on the ABU dataset compared to other methods using San-Diego-02 data. (a) RGB, (b) ground-truth, (c) GE-AD (F1 = 0.826), (d) Abundance (F1 = 0.655), (e) AED (F1 = 0.636), (f) FCBAD (F1 = 0.536), and (g) KRX (F1 = 0.627).

5.1.2 Cross-Sensor Transfer: ABU-Airport to Arizona

We further evaluated generalization on the Arizona dataset, which differs significantly in scene characteristics and sensor properties.

When trained directly on Arizona data, GE-AD previously achieved F1-macro scores up to 0.859 [31]. However, when the model trained only on ABU-Airport was tested on Arizona-V, the F1-macro score dropped dramatically to 0.490 (Fig. 28).

Interestingly, some individual base detectors (e.g., FCBAD) achieved higher performance than the ensemble itself. This indicates that the meta-learner’s weighting strategy, optimized for ABU-Airport characteristics, fails to transfer under substantial domain shift.

Feature-importance analysis (Fig. 29) using Mean Decrease Impurity (MDI) [165] further reveals substantial discrepancies in detector relevance across datasets. The learned importance distribution on ABU-Airport does not align with the required relevance pattern for Arizona scenes. This mismatch suggests that GE-AD implicitly learns dataset-specific correlations rather than domain-invariant anomaly representations.

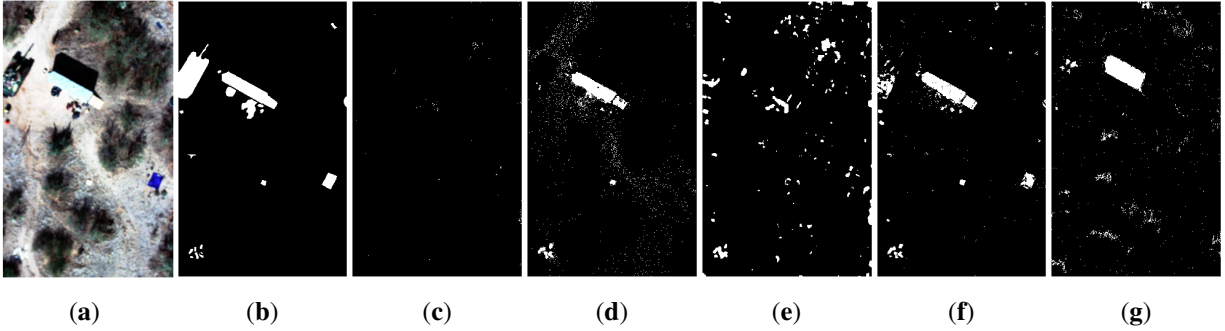


Figure 28: The visualization shows the performance of our proposed ensemble method (GE-AD) trained on the ABU dataset compared to other methods using Arizona-V data. (a) RGB, (b) ground-truth, (c) GE-AD (F1 = 0.490), (d) Abundance (F1 = 0.701), (e) AED (F1 = 0.584), (f) FCBAD (F1 = 0.810), (g) KRX (F1 = 0.503).

These results indicate that classical ensemble-based HS-AD systems may generalize across visually similar scenes but struggle under significant cross-sensor and cross-environment shifts. The failure arises from distribution-dependent feature weighting and the absence of explicit domain alignment mechanisms.

Therefore, a more principled approach is required—one that learns domain-invariant

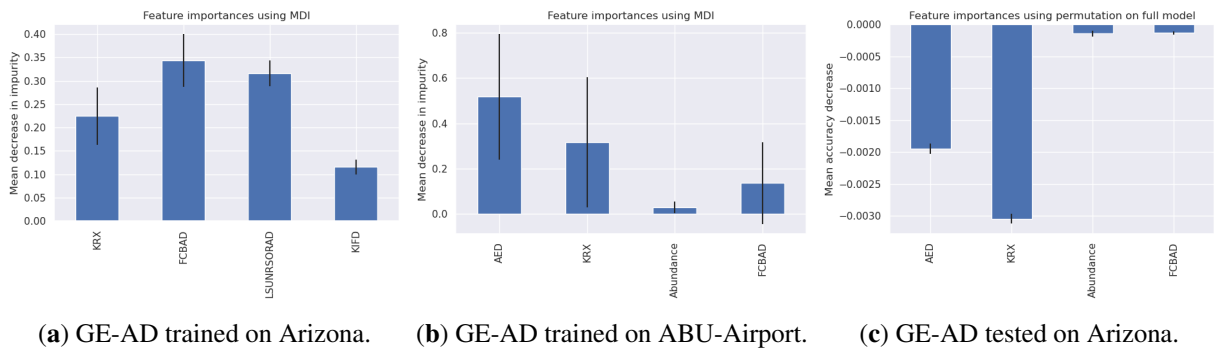


Figure 29: Visualization of difference of feature importance between models.

representations while preserving anomaly discriminability. This motivates the development of domain-adaptive neural architectures, introduced in the following sections.

5.1.3 Problem Formulation

Let $\{\mathcal{D}_1, \dots, \mathcal{D}_N\}$ denote N source domains, where each domain corresponds to a dataset-specific distribution characterized by sensor properties, spectral resolution, acquisition conditions, and scene content. Each domain $\mathcal{D}_i$ provides labeled samples (x, y) for anomaly detection.

The objective is to learn a function

$$f : x \mapsto \hat{y}$$

that:

1. Accurately detects anomalous pixels,
2. Generalizes to unseen target domains $\mathcal{D}_T$ whose distribution differs from all source domains,
3. Operates without requiring domain labels at inference time.

Unlike conventional HS-AD models trained and evaluated on a single dataset, this formulation explicitly addresses cross-domain deployment, where both sensor characteristics and scene statistics may shift.

5.2 Unified MoE Ensemble framework for HS-AD

5.2.1 System Overview

Classical ensemble-based HS-AD methods assign fixed importance weights to base detectors learned from a single dataset. Because these weights implicitly encode dataset-specific statistical structure, performance degrades when the underlying spectral distribution shifts across

sensors or scenes.

To address this limitation, we design a unified MoE framework for HS-AD, which extends ensemble-based detection to a cross-domain setting and combines adaptive expert specialization with domain-invariant representation learning. Figure 30 illustrates the overall design of the proposed Generalized Anomaly Detector. The system follows a left-to-right processing pipeline consisting of base anomaly detectors, a unified feature-extraction backbone, stacking ensembles, and a MoE soft-gate router.

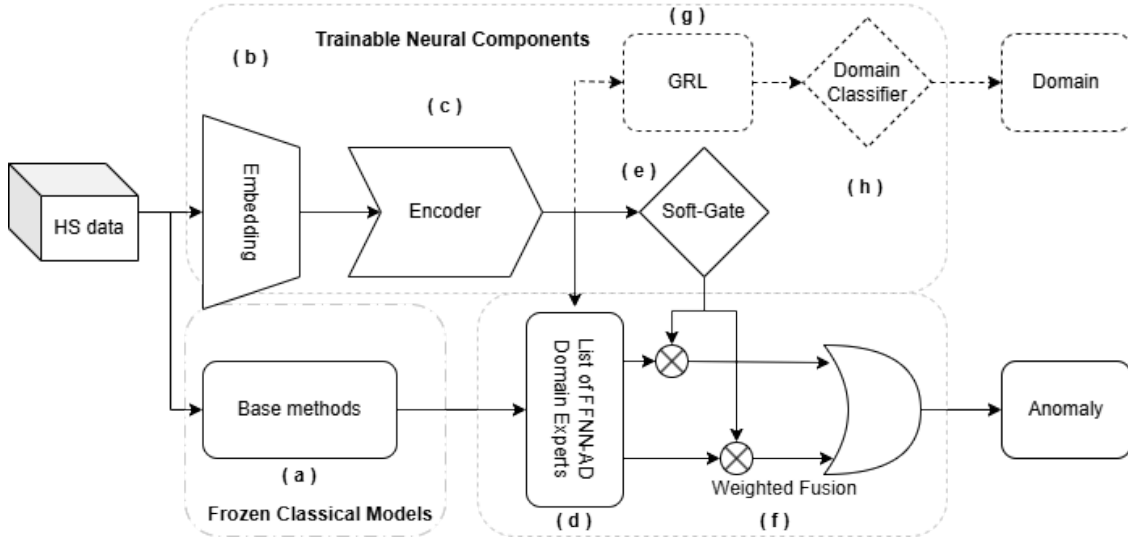


Figure 30: The overall system design of the proposed Generalized Anomaly Detector consists of the following components: (a) base methods that produce initial anomaly detection maps; (b) HyperSL embedding including Universal Spectral Tokenizer and Spectral Wavelength Positional Encoding; (c) HyperSL encoder, which, together with (b), forms the unified feature-extraction backbone; (d) a group of FFNN-AD experts; (e) soft-gate network that predicts expert weights from latent features; (f) weighted fusion module to produce the final anomaly prediction; (g) Gradient Reversal Layer (GRL); and (h) domain classifier, together with (g), enforces domain-invariant feature learning during training (Note that the dashed connections indicate training-only paths. The GRL (g) and domain classifier (h) are removed at inference.).

(a) Base Methods Each input hyperspectral image is processed by multiple base HS-AD methods to produce anomaly response maps. The outputs are concatenated into a feature vector b , which serves as input to the FFNN-AD experts. We used the following statistical HS-AD methods: Attribute and Edge-Preserving Filters (AED) [23], Fuzzy Cluster-Based Anomaly Detection (FCBAD) [71], HS-AD With Kernel iForest (KIFD) [38], Kernel-RX Algorithm

(KRX) [35], and Local Summation Unsupervised Nearest Regularized Subspace With An Outlier Removal Anomaly Detector (LSUNRSORAD) [39]. We also used one unmixing-based AD, and this abundance generation is described in Sec. 3.1.4.

(b-c) Feature Embedding and Encoder (Backbone) In parallel, the hyperspectral input is passed through the pre-trained foundation model HyperSL [175]. HyperSL is selected due to its spectral-aware architecture and strong cross-dataset generalization capability, which are desirable for domain-robust anomaly detection (see Sec. 5.2.2 for further discussion). The HyperSL embedding stage (b), composed of the Universal Spectral Tokenizer and Spectral Wavelength Positional Encoding, together with the HyperSL encoder (c), constitutes the HyperSL backbone and collectively produces the latent feature representation f_l . These latent features are also consumed by our new extended FFNN-AD experts. They are also used for expert routing (Sec. 5.2.4) and domain-invariant feature learning (Sec. 5.2.6).

(d) FFNN-AD Experts A set of FFNN-AD experts is employed, meta-model implemented as a lightweight two-layer FFNN that takes both the base-method feature vector b and the latent feature f_l as input. The experts are soft-specialized, where multiple experts collaborate to solve the AD problem across domains. All experts are trained jointly in an end-to-end manner with the rest of the framework (Sec. 5.2.3).

(e) Soft-Gate Network A lightweight gating network takes f_l as input and produces a probability distribution over experts (Sec. 5.2.4). These probabilities indicate the relevance of each expert for the input and are used as fusion weights.

(f) Weighted Fusion The final anomaly score is computed as a weighted sum of the top- k expert outputs (Sec. 5.2.4). This allows multiple high-confidence experts to contribute proportionally to the final prediction.

(g-h) Domain-Adversarial Path During training, the latent feature f_l is passed through (g) a GRL and (h) a domain classifier. This branch predicts dataset-domain labels while

reversing gradients during backpropagation, encouraging the backbone to learn domain-invariant representations (Sec. 5.2.6). During inference, this branch would be removed; the model does not accept domain labels as input and only outputs the anomaly score.

5.2.2 Foundation Model as Backbone

We use a hyperspectral vision transformer (ViT)-based foundation model called HyperSL [175] as our feature-extraction backbone because of its unified spectral representation. This allows transfer across different datasets and sensors with minimal changes to the architecture. HyperSL was pre-trained on about 200 million spaceborne hyperspectral samples from two sensors, EnMAP (the Environmental Mapping and Analysis Program) and DESIS (the DLR Earth Sensing Imaging Spectrometer). This pre-training transfers knowledge from related source domains to improve generalization in target domains with limited labeled data.

The backbone comprises a Universal Spectral Tokenizer, Spectral Wavelength Positional Encoding, and a three-block Transformer encoder. The tokenizer maps diverse spectral vectors into a shared token space, while the wavelength-aware positional encoding aligns features across spectral ranges. As illustrated in Fig. 30, these components form the embedding stage (component b) and the encoder (component c), which produce latent features. The resulting representations are provided to the Soft-gate module for routing and to the GRL and domain classifier for domain-adversarial learning.

During training, the HyperSL encoder is fine-tuned jointly with the downstream modules, while the tokenizer and positional encoding components remain fixed.

5.2.3 Experts

Let $\{E_1, E_2, \dots, E_N\}$ denote a pool of N FFNN-based AD experts, as shown in component d of Fig. 30. Each expert produces an anomaly prediction:

$$a_i = E_i(b, f_l), \quad i = 1, \dots, N,$$

where b is the output of the base methods and f_l is the latent feature representation from the backbone for an input hyperspectral pixel x .

Figure 31 shows the outline of our proposed FFNN-AD expert. During training, all FFNN-AD experts are jointly optimized using samples from all datasets (domains). The MoE framework [127] learns a soft routing function that assigns input-dependent weights to experts, enabling specialization. As a result, different experts capture diverse background distributions and anomaly characteristics across domains, while the gating mechanism selects and combines the most relevant experts for each input.

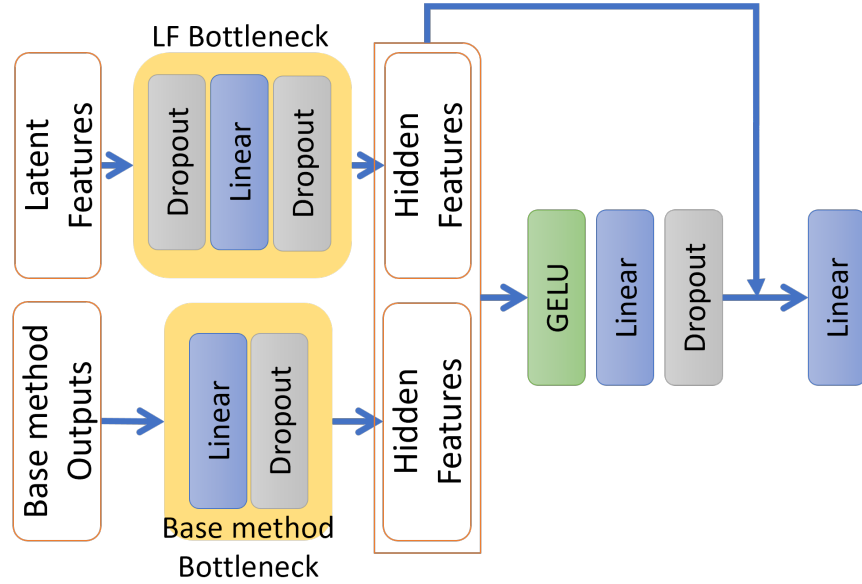


Figure 31: Outline of our proposed FFNN-AD expert

5.2.4 Mixture-of-Experts (MoE) Based Framework

We propose a unified Mixture-of-Experts (MoE) framework for HS-AD, which extends ensemble-based detection to a cross-domain setting. Instead of relying on dataset-specific models, the proposed framework learns a set of neural experts that implicitly specialize across heterogeneous data distributions through joint training on pooled multi-domain data. Given an input hyperspectral sample, a soft gating network adaptively selects and combines a subset of experts to produce the final anomaly score. Although domain labels are utilized during training

to guide experts, encourage feature alignment, and expert diversity, they are not required at inference time. As a result, expert selection is performed solely via the gating network based on the input features, enabling domain-agnostic inference.

Soft-Gate Network

Figure 32 shows the outline of our proposed lightweight feedforward Soft-Gate network $G(\cdot)$. It takes the latent feature f_l along with simple spectral statistics—mean, standard deviation, and maximum—calculated from the input hyperspectral pixel. These serve as a lightweight spectral fingerprint that complements the learned latent representation. This spectral fingerprint offers a compact summary of the spectral distribution, providing additional low-level cues about the data source and its domain that might not be fully captured by the learned latent representation, especially during cross-domain spectral shifts. Soft-Gate network then outputs a set of non-negative weights (see Fig. 30 (e)):

$$\mathbf{w} = G(f_l), \quad \mathbf{w} \in \mathbb{R}^N, \quad \sum_i w_i = 1$$

These weights represent the relevance of each expert for the given input.

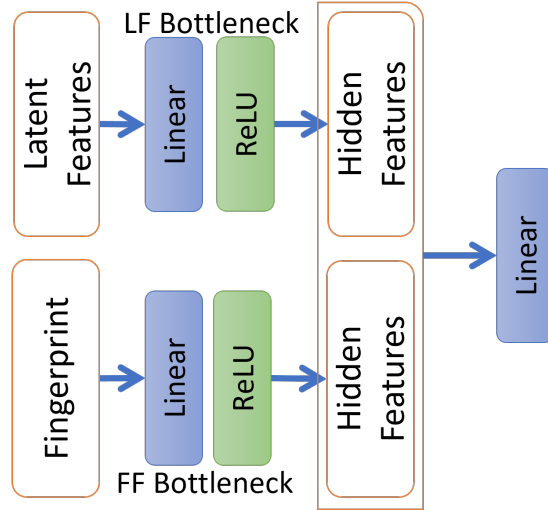


Figure 32: Outline of our proposed Soft-Gate Network

Top-k Expert Selection

To reduce interference from weakly related domains, only the top- k gating weights are retained [129, 130]:

$$\mathcal{K} = \text{TopK}(\mathbf{w}, k)$$

Experts outside $\mathcal{K}$ are ignored for a given input to avoid performance degradation caused by unrelated experts.

In our implementation, we set $k = 2$ to balance specialization and collaborative robustness. Using only the top-1 expert reduces the model to hard routing, increasing sensitivity to domain misclassification and amplifying routing errors. Conversely, activating all experts introduces contributions from weakly related domains, leading to over-smoothing of anomaly scores and reduced discriminative contrast.

Zoph *et al.* [130] recommend $k = 2$. We also observed that $k = 2$ provides stable performance across heterogeneous datasets and helps avoid routing collapse (i.e., dominance by a single expert). They also report small gains from top- $(n + 1)$ over top- n routing; however, we found this improvement to be negligible.

Unlike token-based MoE models, where load balancing prevents experts from collapsing due to stochastic routing, expert specialization is explicitly tied to dataset and sensor characteristics rather than random assignment [128]. Therefore, strong load-balancing regularization is unnecessary; we include it with a small weighting factor to prevent degenerate routing without interfering with domain-driven specialization.

5.2.5 Weighted Prediction Fusion

The final anomaly prediction is computed as a weighted sum of the selected experts' outputs (see Fig. 30 (f)). Given the top- K selected experts $\mathcal{K}$ with gating scores w_i , the weights are first re-normalized over the selected set:

$$\tilde{w}_i = \frac{w_i}{\sum_{j \in \mathcal{K}} w_j}, \quad \text{for } i \in \mathcal{K} \quad (14)$$

The final anomaly score is then obtained as

$$a_{\text{final}}(x) = \sum_{i \in \mathcal{K}} \tilde{w}_i \cdot a_i(x) \quad (15)$$

This re-normalization ensures that the selected high-confidence experts contribute proportionally while preserving a convex combination ($\sum_{i \in \mathcal{K}} \tilde{w}_i = 1$), producing a single, unambiguous anomaly score that is robust to cross-sensor and cross-dataset variation.

Finally, although inspired by sparse and soft-gated MoE architectures [128, 129], our approach uses domain-specialized FFNN-AD models as experts, routing occurs at the sample level rather than the token or trajectory level, and expert choice is implicitly driven by domain and sensor characteristics learned through routing, thus avoiding the need for extensive load-balancing regularization.

5.2.6 Training-time Domain Adaptation and Generalization

We formulate the proposed framework as a multi-task learning problem that jointly optimizes the objectives of anomaly detection and domain classification. DA is implemented using a GRL [176], which is placed between the backbone and the domain classifier. During training, the domain classifier is trained to correctly identify the dataset for each sample, while GRL inverts the gradients before they reach the backbone, thereby encouraging the backbone to learn feature representations that remain discriminative and informative for anomaly detection but insensitive to dataset-specific differences. As a result, the extracted features become more domain-invariant and generalize better to unseen datasets. In addition, we investigate domain generalization by training on multiple source domains to capture invariant factors shared across datasets [177].

5.2.7 Training

Training is performed jointly, reflecting the unified structure of the proposed architecture. Unlike the original GE-AD formulation 4.3, we have updated the meta-model with a feedforward network, enabling joint optimization in which FFNN-AD experts operate on both the outputs of base anomaly detectors and the backbone’s latent features. Each FFNN-AD expert is trained jointly with the downstream modules using labeled anomaly data from all datasets and is updated via MoE. This joint optimization improves cross-domain generalization and allows experts to adapt dynamically during training.

The HyperSL backbone is initialized from a pre-trained foundation model [175]. During training, the backbone is first frozen for 2–3 warm-up epochs to stabilize the gating and downstream modules, after which it is fine-tuned jointly with the gating (routing) network and the domain classifier. Due to inconsistent spectral dimensions across data sources, it is not possible to load them into the same batch for training the output model. To mitigate data bias in the model, we employed three separate data samplers (minority-class, domain, and spectral-dimension) within the loader. Given an input hyperspectral sample, the backbone extracts latent features that the gating network uses to assign weights to experts and that the domain classifier uses for domain-adversarial learning. During this stage, the model is optimized using a weighted combination of classification, domain, and routing-related losses, described below. The GRL ensures that the backbone learns representations that are discriminative for anomaly detection while remaining invariant to dataset-specific characteristics.

We used the AdamW optimizer [162] with an initial learning rate of 5×10^{-4} and employed Cosine Annealing with Warm Restarts [178] as the learning rate scheduler. We set the initial restart period $T_0 = 5$, $T_{mult} = 2$, and the minimum learning rate $\eta_{min} = 1 \times 10^{-8}$ for the scheduler. Following standard mixed-precision training practices [179, 180], all neural components were trained using bfloat16 precision, which preserves FP32-like dynamic range while reducing memory and computational cost. We can train to maximize F1-macro (threshold=0.50) and minimize the loss. Training was conducted on two NVIDIA A40 GPUs.

During inference, the threshold for F1-macro is determined using the validation set. At this stage, only the backbone, gating network, and FFNN-AD experts remain active; the domain classifier and GRL are omitted, resulting in a final single anomaly score for each input sample.

Loss Function

The overall objective combines anomaly classification, domain supervision, and routing regularization terms with stage-dependent activation.

Domain-Weighted Anomaly Classification. The primary supervision is a domain-weighted binary cross-entropy loss with logits:

$$L_{cls} = \sum_{d \in \mathcal{D}} w_d \cdot \mathbb{E}_{i \in d} [\ell_{BCE}(y_i, \hat{y}_i)] \quad (16)$$

where w_d denotes the domain-specific weight and ℓ_{BCE} is a class-weighted binary cross-entropy loss addressing class imbalance. The loss ℓ_{BCE} is defined as:

$$\ell_{BCE}(y, \hat{y}) = -[\alpha y \log \sigma(\hat{y}) + (1 - y) \log(1 - \sigma(\hat{y}))] \quad (17)$$

where $\sigma(\cdot)$ is the sigmoid function and α is a positive class weight that up-weights anomaly samples to address class imbalance.

In our implementation, anomaly samples are assigned a higher weight (via `pos_weight`) to compensate for their rarity. This formulation decouples domain imbalance (handled by w_d) from class imbalance (handled by α). The overall contribution of L_{cls} is gradually increased during training to optimize early exploration and stabilize the experts in the later stages.

Domain-Adversarial Loss. A domain classification loss (L_{grl}) is applied to the domain prediction head after a warm-up period using a Gradient Reversal Layer (GRL) [176]. The loss

is annealed as:

$$L_{grl} = \frac{1}{t+1} \cdot \ell_{CE}(d_i, \hat{d}_i) \quad (18)$$

where t is the training epoch and scaled by λ_{dom} , encouraging domain-invariant feature learning.

Load Balancing Loss. To prevent expert collapse during early training, we apply a load balancing loss (L_{lb}):

$$L_{lb} = E \cdot \sum_{j=1}^E p_j \cdot f_j \quad (19)$$

where E is the number of experts, p_j is the average routing probability, and f_j is the fraction of samples routed to expert j . This term is active only during the initial training phase. An expert collapse is where most samples are routed to a small subset of experts [179].

Domain Diversity Loss. To promote domain-specific expert specialization, we introduce a margin-based diversity loss with masked distance computation:

$$L_{div} = \frac{1}{|\mathcal{P}|} \sum_{(d,k)} \max(0, m - \|(\tilde{p}_d - \tilde{p}_k) \odot \mathbf{M}_{d,k}\|^2) \quad (20)$$

where $\tilde{p}_d$ and $\tilde{p}_k$ are normalized routing distributions, $\mathbf{M}_{d,k}$ is a top- k mask selecting dominant experts, and m is an adaptive margin.

This loss consists of two components: (i) separation between current batch-wise domain distributions, and (ii) separation between batch distributions and historical domain means. The historical means are estimated via an exponential moving average (EMA) and normalized to form valid distributions. Domain weights are applied when comparing against historical means.

Entropy Regularization. An entropy regularization term is applied to the router:

$$L_{\mathcal{H}} = -\mathbb{E} \left[\sum_j p_j \log p_j \right] \quad (21)$$

This term is subtracted from the loss to encourage confident routing and is only applied during early training.

Expert Regularization. When expert-specific outputs are enabled, an ℓ_2 regularization term is applied to expert logits to stabilize training:

$$L_{exp} = \mathbb{E} [\|z_{exp}\|^2] \quad (22)$$

Final Objective. The total loss is:

$$L = \lambda_{cls}(t)L_{cls} + \lambda_{dom}L_{grl} + \lambda_{lb}L_{lb} + \lambda_{div}L_{div} - \lambda_{\mathcal{H}}L_{\mathcal{H}} + \lambda_{exp}L_{exp} \quad (23)$$

Each component is activated conditionally. Load balancing and entropy regularization are applied during early training, while domain diversity and adversarial learning are introduced after a warm-up period. The classification weight $\lambda_{cls}(t)$ increases over time, enabling stable optimization followed by stronger task supervision.

5.3 Test-time Domain Adaptation

We address test-time DA (TTA) by introducing a Drift-aware hierarchical Test-time Adaptation (DTA) strategy designed for HS-AD. The goal is to adapt the model to distributional shifts that may occur at inference time while avoiding unnecessary or harmful updates when the test distribution closely matches the training distribution.

Given incoming test samples, we first estimate distributional drift relative to source-domain statistics computed during training. This drift estimation determines whether test-time adaptation should be enabled and, if so, which components of the model should be adapted. By conditioning adaptation on explicit drift measurements, the proposed strategy prevents over-adaptation and preserves anomaly-discriminative representations under stationary conditions.

The DTA framework operates hierarchically: when the detected drift is small, no adaptation is performed and source-domain statistics are preserved; as the magnitude of drift increases, progressively stronger adaptation mechanisms are selectively activated [121, 122]. This design enables stable adaptation in high-dimensional hyperspectral settings, where indiscriminate normalization updates may otherwise degrade performance [125].

Batch Normalization Drift Estimation

To quantify distributional shifts at the population level, we measure deviations between target-domain Batch Normalization (BN) statistics and the frozen source-domain BN statistics obtained after training. This approach is motivated by the moment-matching interpretation of BN [120] and its use in normalization-based test-time adaptation, but does not rely on entropy minimization to infer the presence or direction of shift.

Let a BN layer maintain frozen source statistics (μ_s, σ_s^2) . Given a target batch $x_t \in \mathbb{R}^{N \times C \times H \times W}$, we compute channel-wise target statistics:

$$\mu_t^c = \mathbb{E}_{n,h,w}[x_t^{n,c,h,w}], \quad \sigma_t^{2,c} = \text{Var}_{n,h,w}[x_t^{n,c,h,w}]$$

We define the Batch Normalization drift score D_{BN} as a normalized channel-wise deviation:

$$D_{\text{BN}} = \frac{1}{C} \sum_{c=1}^C \frac{|\mu_t^c - \mu_s^c|}{\sqrt{(\sigma_s^c)^2 + \epsilon}}$$

where ϵ is a small constant for numerical stability. This score provides a scale-invariant, layer-wise indicator of covariate shift without requiring access to stored source data, unlike discrepancy-based measures such as MMD [181].

Layer Normalization Drift Estimation

While BN captures population-level distributional shifts, Transformer-based architectures rely primarily on Layer Normalization (LN), which operates at the sample level and is invariant to batch statistics. As a result, BN-based drift measures alone are insufficient to detect representation-level shifts affecting LN-based models.

To estimate such representation drift, we use the energy score [182], computed from the anomaly classifier logits:

$$E_t = -\log \sum_{k=1}^K \exp(z_k) \quad (24)$$

where z_k denotes the logit corresponding to class k produced by the anomaly classifier head. By storing the mean energy score on a validation set during training, denoted as $\mathbb{E}[E_s]$, we define the Layer Normalization drift score D_{LN} as:

$$D_{\text{LN}} = |\mathbb{E}[E_t] - \mathbb{E}[E_s]|$$

This metric captures shifts in the model’s confidence distribution and provides a lightweight indicator of representation-level drift relevant to LN-based architectures.

5.3.1 Drift-Aware Adaptation Policy

Using the BN- and LN-based drift scores, we define a drift-aware adaptation policy that selectively activates test-time adaptation mechanisms. Given predefined thresholds τ_1 , τ_2 , and τ_3 , the policy operates as follows:

- (a) No domain shift** ($D_{\text{BN}} < \tau_1$): The target distribution closely matches the source distribution. All normalization statistics are restored to their source values, and no adaptation is performed.
- (b) Mild shift** ($\tau_1 \leq D_{\text{BN}} < \tau_2$): Adaptation is restricted to updating BN running statistics, correcting low-order distributional mismatches while preserving learned representations [122].

(c) Severe shift ($D_{\text{BN}} \geq \tau_2$): Stronger adaptation is enabled by updating BN affine parameters to improve prediction confidence [121].

(d) Representation drift ($D_{\text{LN}} \geq \tau_3$): When representation-level drift is detected, normalization affine parameters (including Layer Normalization parameters in Transformer-based components) are updated to reduce predictive uncertainty. If both severe BN drift and LN drift are present, adaptation is applied jointly while avoiding redundant updates.

Overall, the proposed DTA framework enables adaptive yet conservative test-time adaptation by explicitly conditioning updates on measured distributional drift. By leveraging the complementary properties of BN and LN, the method supports both convolutional and Transformer-based components while minimizing unnecessary updates and maintaining stability in HS-AD.

5.4 Results

We evaluated the performance of the proposed model on seven hyperspectral datasets (domains) acquired by different sensors (AVIRIS, HYDICE, Pika-L, and ROSIS-03) and across diverse scenes. We split each dataset individually in a stratified manner based on class labels, allocating 32% for training, 8% for validation, and 60% for testing, thereby preserving the original class imbalance in each domain. Using these splits, the training portions of all datasets were combined to train the complete neural network end-to-end, including five experts, enabling joint learning across domains while maintaining strict separation of the held-out test sets. The test data were not used during training, hyperparameter tuning, or model selection, and were employed exclusively for the final performance evaluation reported in this section.

Table 20 reports ROC-AUC for ranking ability, PR-AUC for robustness to class imbalance, F1-macro for detection quality, and TPR@FPR for performance under low false-alarm constraints. The threshold for F1-macro is selected using the validation set. These metrics provide a comprehensive evaluation under severe class imbalance. Overall performance

is reported as the mean and standard deviation across datasets to ensure equal contribution from each dataset.

Table 20: Quantitative performance of the proposed model across all datasets (mean $\pm$ standard deviation over 5 runs with different random seeds).

Dataset	ROC-AUC $\uparrow$	PR-AUC $\uparrow$	Best F1-macro $\uparrow$	TPR@1% FPR $\uparrow$	TPR@5% FPR $\uparrow$
HYDICE -Urban	0.991 ± 0.010	0.824 ± 0.061	0.925 ± 0.025	0.872 ± 0.096	0.949 ± 0.073
Salinas	0.947 ± 0.027	0.801 ± 0.040	0.913 ± 0.014	0.844 ± 0.031	0.844 ± 0.031
San-Diego	0.997 ± 0.001	0.783 ± 0.026	0.890 ± 0.009	0.950 ± 0.015	0.987 ± 0.005
ABU-Airport	0.956 ± 0.037	0.467 ± 0.053	0.752 ± 0.017	0.573 ± 0.052	0.857 ± 0.093
ABU-Beach	0.973 ± 0.010	0.420 ± 0.113	0.736 ± 0.057	0.574 ± 0.118	0.786 ± 0.096
ABU-Urban	0.992 ± 0.001	0.726 ± 0.022	0.868 ± 0.007	0.822 ± 0.019	0.968 ± 0.010
Arizona	0.748 ± 0.151	0.337 ± 0.092	0.666 ± 0.090	0.351 ± 0.103	0.473 ± 0.153
Average	0.943 ± 0.082	0.623 ± 0.191	0.821 ± 0.094	0.712 ± 0.200	0.838 ± 0.164

Table 20 summarizes the quantitative performance of the proposed model across seven hyperspectral datasets, with results reported as mean $\pm$ standard deviation over five runs with different random seeds. Figure 33 shows qualitative examples of the model’s best performance. On average, the model achieves a ROC-AUC of 0.943 ± 0.082 , indicating strong ranking ability across domains. The PR-AUC of 0.623 ± 0.191 reflects the increased difficulty under severe class imbalance, where precision-recall characteristics are more sensitive to rare anomalies.

The best F1-macro score of 0.821 ± 0.094 demonstrates moderate detection quality at an optimal threshold selected on the validation set. Additionally, the model achieves TPR values of 0.639 ± 0.320 at 1% FPR and 0.798 ± 0.260 at 5% FPR, indicating the model’s ability to detect anomalies under low false-alarm constraints, which is critical for practical hyperspectral anomaly detection systems. These 1% and 5% false-positive rates are commonly used operating points in anomaly detection, representing strict and moderately relaxed false-alarm regimes. They allow consistent comparison of models under practically meaningful conditions.

Performance varies across datasets. The model achieves strong results on

HYDICE-Urban, San-Diego, and ABU-Urban, with ROC-AUC values exceeding 0.99 and consistently high TPR at low FPR. In contrast, performance degrades on more challenging datasets, such as Arizona and ABU-Beach, as reflected by lower PR-AUC and F1 scores along with higher variance. This suggests that domain shift and scene complexity significantly impact detection performance.

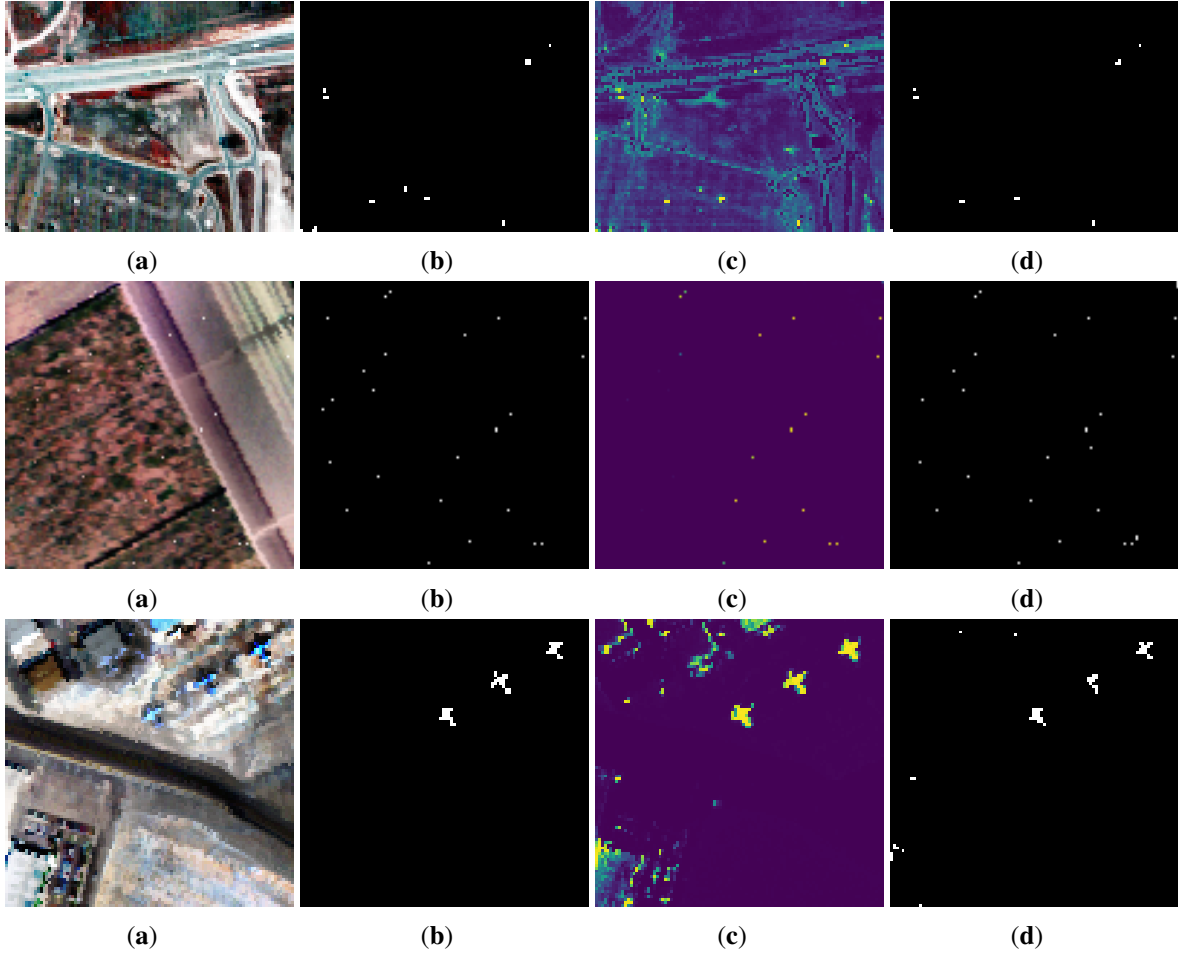


Figure 33: Qualitative results across datasets. Rows (top to bottom): HYDICE-Urban (F1=0.958), Salinas (F1=0.933), San-Diego image 02 (F1=0.913), ABU-Airport image 02 (F1=0.819), ABU-Beach image 01 (F1=0.976), ABU-Urban image 02 (F1=0.942), Arizona image 04 (F1=0.923). Columns (left to right): RGB image, ground-truth mask, anomaly score map, and binary detection result.

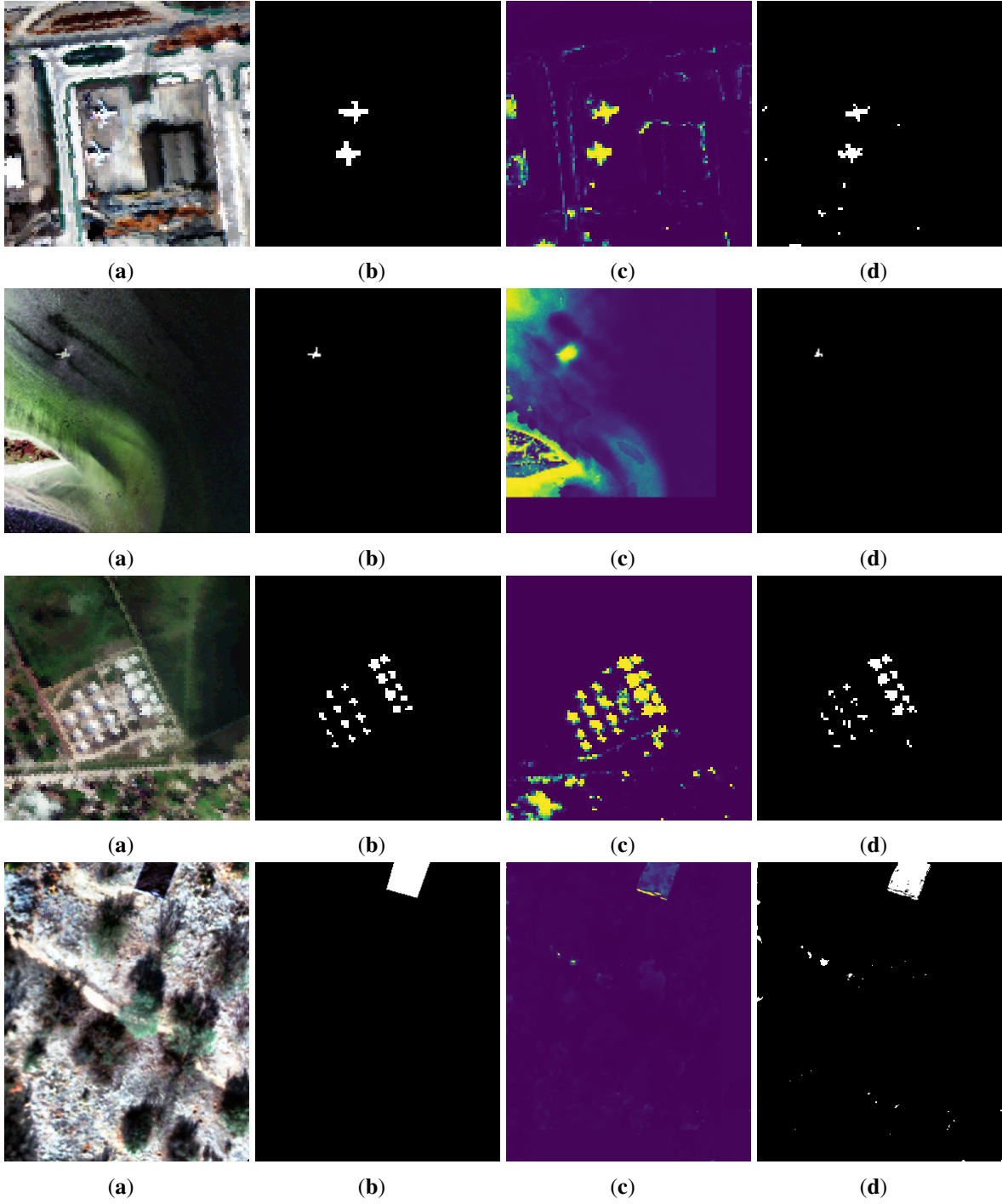


Figure 33: Qualitative results across datasets (continued).

Overall, the results demonstrate that the proposed model generalizes well across multiple sensors and domains, while highlighting the challenges posed by highly diverse and imbalanced hyperspectral datasets. However, the relatively large standard deviations across

certain datasets (e.g., ABU-Beach and Arizona) indicate that severe domain shifts remain challenging. Nevertheless, the overall performance across diverse datasets suggests that the mixture-of-experts design contributes to improved adaptability to domain-specific characteristics while maintaining generalization.

5.4.1 *Leave-One-Out Evaluation*

Table 21 reports the performance of the proposed model under a leave-one-dataset-out evaluation protocol. In each experiment, the model is trained on all but one dataset and evaluated exclusively on the held-out dataset, thereby simulating deployment under an unseen-domain scenario. The same set of imbalance-aware metrics is reported to enable direct comparison with the fully supervised results in Tab. 20. As F1 is threshold-dependent, we randomly selected 20% of the data to compute the best threshold during evaluation. The performance exhibits substantially higher variability than in the in-domain experiments, with F1-macro scores ranging from 0.650 to 0.888, indicating that this setting is significantly more challenging. Compared to the fully supervised results in Tab. 20, where the average F1-macro reaches 0.821, the leave-one-out setup leads to noticeable degradation on several datasets, highlighting the sensitivity of HS-AD models to domain shift when the target dataset is entirely excluded from training.

The largest performance degradation occurs on the Arizona dataset, which is collected using the Pika-L sensor. Because no other Pika-L dataset is available during training, the model lacks sensor-consistent representations, leading to a sharp drop in F1-macro (0.650 ± 0.014). This result underscores the importance of sensor diversity for cross-domain generalization.

A similar, though less severe, degradation is observed for ABU-Beach. This dataset includes imagery from multiple sensors, increasing intra-dataset heterogeneity and making it difficult for the model to learn a coherent anomaly representation when the dataset is fully excluded during training.

In contrast, performance on HYDICE-Urban remains comparatively stable despite the

Table 21: Performance of our new model, where we excluded one dataset completely during training. The Dataset column shows the excluded dataset.

Dataset	ROC-AUC $\uparrow$	PR-AUC $\uparrow$	Best F1-macro $\uparrow$	TPR@1% FPR $\uparrow$	TPR@5% FPR $\uparrow$
HYDICE -Urban	0.966 ± 0.015	0.418 ± 0.149	0.743 ± 0.083	0.718 ± 0.181	0.897 ± 0.036
Salinas	0.976 ± 0.007	0.656 ± 0.110	0.855 ± 0.031	0.822 ± 0.031	0.889 ± 0.031
San-Diego	0.996 ± 0.000	0.776 ± 0.022	0.888 ± 0.015	0.911 ± 0.023	0.987 ± 0.005
ABU-Airport	0.940 ± 0.008	0.375 ± 0.044	0.724 ± 0.010	0.473 ± 0.052	0.800 ± 0.007
ABU-Beach	0.972 ± 0.005	0.430 ± 0.034	0.716 ± 0.020	0.599 ± 0.092	0.806 ± 0.057
ABU-Urban	0.984 ± 0.006	0.653 ± 0.115	0.834 ± 0.045	0.729 ± 0.106	0.926 ± 0.034
Arizona	0.538 ± 0.018	0.226 ± 0.031	0.650 ± 0.014	0.246 ± 0.019	0.276 ± 0.017

absence of other HYDICE-Urbandatasets in training. This can be attributed to scene-level similarity with ABU-Urban, which allows knowledge transfer across datasets even when sensor characteristics differ.

For datasets collected using AVIRIS (e.g., San-Diego, ABU-Airport, ABU-Urban, and Salinas), the model achieves consistently higher performance. The availability of multiple AVIRIS datasets enables collaborative learning among experts, mitigating domain shift and improving the robustness of anomaly detection in the leave-one-out setting.

These results further demonstrate that the mixture-of-experts framework benefits from exposure to diverse domains during training, while performance degrades when domain-specific characteristics are entirely absent.

5.4.2 Results from Base Methods

Table 22 presents a quantitative comparison between the proposed method and base HS-AD and unmixing-AD methods. Because most of these methods are unsupervised and produce continuous anomaly scores without a canonical (i.e., method-intrinsic and dataset-independent) decision threshold, threshold-dependent metrics such as the F1-score

would require additional assumptions and heuristic tuning. This complements the evaluation in Sec. 5.4, where threshold-dependent metrics are used, by providing a consistent comparison with unsupervised base methods.

In contrast, ROC-AUC enables a threshold-independent comparison and evaluates the ranking quality of anomaly scores across all possible operating points. Thus, to ensure a fair, reproducible, and assumption-free comparison with prior work, we adopt ROC-AUC as the primary evaluation metric here.

Table 22: ROC-AUC performance comparison between the proposed method and base HS-AD and unmixing-AD methods

Dataset	San -Diego	ABU -Airport	ABU -Urban	HYDICE -Urban	Salinas	ABU -Beach	Arizona	Average
Our method	0.997 ± 0.001	0.956 ± 0.037	0.992 ± 0.001	0.991 ± 0.010	0.947 ± 0.027	0.973 ± 0.010	0.748 ± 0.151	0.943 ± 0.082
Abundance	0.902	0.710	0.807	0.866	0.957	0.772	0.575	0.748 ± 0.187
AED	0.987	0.976	0.934	0.963	0.911	0.982	0.336	0.819 ± 0.272
KIFD	0.990	0.965	0.962	0.997	0.994	0.944	0.561	0.874 ± 0.187
LSUNR -SORAD	0.968	0.972	0.988	0.997	1.000	0.987	0.372	0.844 ± 0.271

Table 22 shows that the proposed method consistently outperforms or matches the input HS-AD and unmixing-based methods across most evaluated datasets, particularly on complex scenes such as San-Diego and ABU-Urban. This improvement demonstrates the effectiveness of the MoE framework in leveraging complementary information from the base methods to achieve enhanced anomaly ranking performance.

5.4.3 Results from Individual GE-AD Methods

To present the results from individual GE-AD methods, we conducted a training and evaluation round and calculated all the metrics. We also trained and evaluated our proposed

method on this same dataset. Table 23 shows the performance of GE-AD models trained and evaluated on the same dataset. These models achieve F1-macro scores ranging from 0.785 to 0.894 using the same training-validation-test split, which serves as an upper bound when no domain shift is present. In contrast, the proposed GMoE-AD model (Sec. 5.4) maintains competitive performance while generalizing across datasets.

Figure 34 further illustrates cross-domain behavior. The single-domain GE-AD models (rows 1–7) achieve strong performance primarily along the diagonal, indicating good in-domain performance but poor generalization across datasets. In contrast, the proposed model (last row) remains stable across domains and maintains consistently higher F1-macro scores across all datasets, demonstrating improved cross-domain robustness. Compared to single-domain GE-AD models, the proposed approach achieves comparable F1-macro performance while significantly improving cross-domain generalization, as demonstrated by leave-one-out and multi-domain results. Notably, while single-domain models exhibit significant performance degradation on unseen datasets (e.g., values below 0.3 in cross-domain settings), the proposed model maintains stable performance across all domains.

Table 23: Performance of GE-AD models trained and evaluated on the same dataset (for completeness, we report standard classification metrics for the single-domain setting, where a fixed threshold is well-defined).

Dataset	Precision	Recall	F1-macro	ROC-AUC	Accuracy	Cohen's Kappa
San -Diego	0.711	0.916	0.779	0.991	0.988	0.558
ABU -Airport	0.772	0.896	0.822	0.993	0.990	0.644
ABU -Urban	0.787	0.954	0.851	0.996	0.988	0.702
HYDICE -Urban	0.727	0.883	0.785	0.983	0.997	0.570
Salinas	0.861	0.933	0.894	1.000	0.999	0.787
ABU -Beach	0.854	0.891	0.871	0.997	0.998	0.743
Arizona	0.837	0.836	0.837	0.942	0.980	0.673

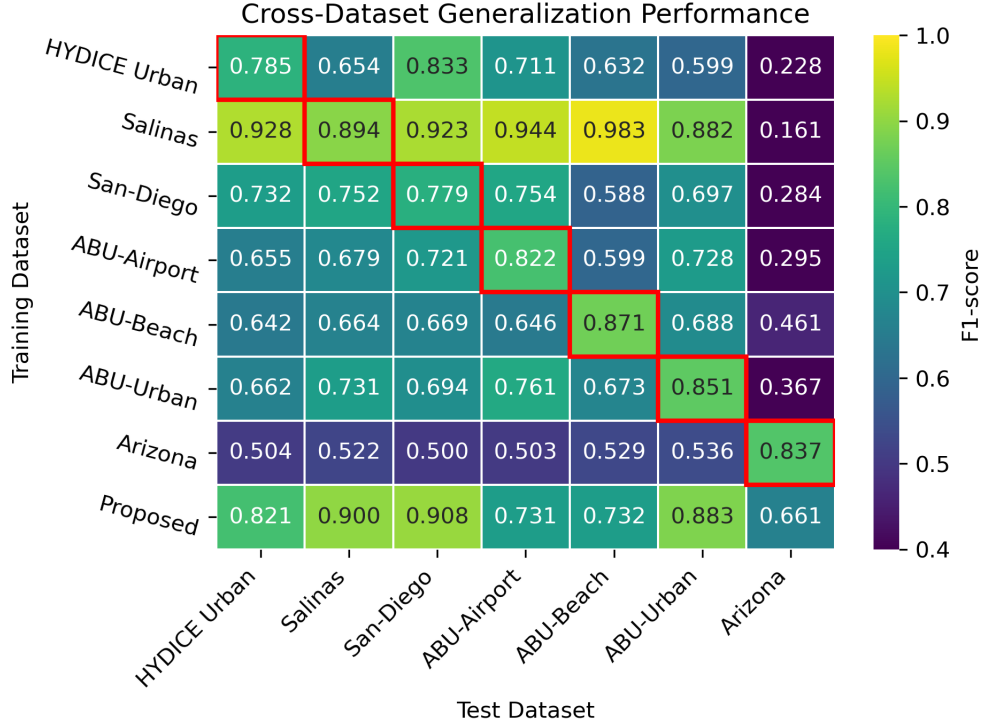


Figure 34: Comparison of cross-domain F1-macro performance. Rows correspond to models trained on a single dataset (rows 1–7) and the proposed mixture-of-experts model (last row), while columns denote the evaluation dataset. Single-domain GE-AD models exhibit strong performance primarily along the diagonal (in-domain) but degrade significantly on unseen datasets (off-diagonal). In contrast, the proposed model maintains consistently higher and more stable performance across all domains, highlighting its ability to generalize under domain shift.

5.4.4 Ablation Test

We formulate the problem as a multi-task learning setup by incorporating domain classification as an auxiliary task. In the ablation study, we remove the domain classifier head, effectively reducing the model to a standalone anomaly detector, and evaluate the impact. Table 24 shows that removing the domain classifier degrades F1-macro performance relative to the full model (with GRL and domain classifier, 0.821 ± 0.094 in Tab. 20), indicating that the auxiliary task improves anomaly detection. This improvement suggests that domain classification regularizes the shared representation, encouraging more structured and stable feature learning and enhancing generalization.

Table 24: Performance of our new model without the domain classifier head and with the domain classifier head but without the Gradient Reversal Layer, averaged among all datasets.

Training	Precision	Recall	F1-macro	ROC AUC
No domain classifier head	0.729	0.782	0.753	0.918
No GRL	0.742	0.783	0.761	0.924

When using the domain classifier head, we apply a GR Layer with a domain-adversarial loss (L_{grl}). Without the GRL, this reduces to a standard multi-class cross-entropy loss. As shown in Table 24, removing the GRL further degrades performance relative to the full model, highlighting the importance of adversarial domain alignment for learning domain-invariant representations. Overall, these results confirm that both the auxiliary domain classification task and adversarial training contribute to improved cross-domain generalization.

5.5 Discussion

Table 20 shows consistent F1-macro (0.821 ± 0.094), PR-AUC (0.623 ± 0.191), and ROC-AUC (0.943 ± 0.082), validating balanced anomaly retrieval and strong class separability, particularly under severe class imbalance. Together, these results indicate that the proposed model achieves reliable and robust anomaly detection across domains.

Furthermore, comparisons with base HS-AD and unmixing-AD methods (Table 22) show that the proposed MoE framework consistently matches or outperforms the individual methods. This demonstrates its ability to effectively combine complementary anomaly cues from multiple experts, thereby improving ranking performance across diverse datasets.

We formulate the problem as a multi-task learning framework by incorporating domain classification as an auxiliary objective. The ablation study confirms its positive contribution to anomaly detection performance (see Table 24). This result indicates that the auxiliary task acts as a regularizer on the shared representation, promoting more structured feature learning and improved generalization.

5.5.1 *Impact of MoE on Anomaly Detection Metrics*

The proposed framework does not always match the upper-bound performance of single-domain models trained and evaluated on the same dataset; this gap is expected, given the substantially more challenging evaluation setting. In contrast to prior HS-AD methods (see Table 23) that assume known test-time domains, our approach must infer the appropriate expert through MoE routing, introducing additional sources of error. Importantly, despite this constraint, the model maintains competitive performance across heterogeneous datasets without retraining, highlighting its robustness and practical applicability in realistic deployment scenarios where domain labels are unavailable.

5.5.2 *Experiment Leave-One-Out*

The leave-one-out results reveal pronounced variability in performance, highlighting the vulnerability of HS-AD models to domain shift when the target dataset is completely unseen during training (see Tab. 21). Severe performance degradation on the Arizona dataset demonstrates the critical role of sensor diversity, as the absence of Pika-L data prevents learning sensor-consistent representations. In contrast, more stable performance on datasets with similar scenes or shared sensor characteristics indicates that both sensor-level and scene-level similarity facilitate cross-domain knowledge transfer. Overall, these findings emphasize that robust generalization in HS-AD depends strongly on domain diversity and inter-dataset complementarity.

5.5.3 *Train-time adaptation*

We observe in Tab. 24 that, compared with the adversarial setup, removing GRL results in a decrease in anomaly detection performance (F1-macro score drops from 0.821 to 0.761), indicating that adversarial DA during training is beneficial.

5.6 Summary

We proposed a unified MoE framework for HS-AD that extends ensemble-based detection to a cross-domain setting and integrates TL and DA to improve robustness when the target domain is unknown at test time. The approach combines traditional unsupervised anomaly detectors within a stacking ensemble and a soft-gated MoE router, initialized via TL and refined through feature distribution alignment, offering a balance between interpretability and flexibility. Evaluated on seven real-world hyperspectral datasets from four sensors in a cross-domain setting, the proposed model achieves an F1-macro score of 0.821 ± 0.094 , PR-AUC of 0.623 ± 0.191 , and ROC-AUC of 0.943 ± 0.082 (mean $\pm$ standard deviation across datasets; each value averaged over multiple runs). While a performance gap remains compared to single-domain models, this reflects the added challenge of cross-domain generalization and cumulative errors from expert routing and detectors when the test domain is unknown. Overall, the results demonstrate that combining TL and DA within an ensemble MoE framework improves resilience to distribution shifts and enables generalized anomaly detection in operational settings. We replace the Random Forest component of the GE-AD stacking ensemble with feedforward networks and enable joint end-to-end training within the proposed framework. This work highlights the potential of domain-aware ensemble architectures for scalable, deployable hyperspectral anomaly detection and motivates further research on adaptive expert selection, improving expert specialization, unsupervised domain discovery, and continual adaptation in dynamic remote sensing systems.

Chapter 6

Hyperspectral Image Semantic Segmentation for Aided Target Recognition

Previous chapters focused on hyperspectral anomaly detection, where the goal is to identify pixels or regions that deviate from the dominant background distribution without requiring semantic labels. While anomaly detection is effective for highlighting unusual or potentially relevant regions, it does not provide semantic interpretation of the observed scene.

In many Aided Target Recognition (AiTR) scenarios, anomaly detection may serve as an initial filtering stage, followed by downstream tasks that provide semantic understanding, such as classification or segmentation. However, semantic understanding may also be required independently of anomaly detection, particularly when the objective is to characterize land cover, vegetation types, or crop distributions across a scene.

This chapter investigates hyperspectral semantic segmentation as a downstream exploitation task that complements, but does not depend on, anomaly detection. The objective is to demonstrate that meaningful semantic information can be extracted from hyperspectral imagery using task-adapted deep learning models, even under practical constraints such as limited labeled data, restricted computational resources, and the absence of large-scale hyperspectral pre-trained models.

6.1 Motivation for Target-Focused Hyperspectral Segmentation

Crop mapping identifies the types of crops grown in a specific area and their spatial distribution, which facilitates important agricultural activities such as assessing plant health and estimating crop production. Remote sensing has increasingly supplanted traditional field surveys for crop mapping. In particular, hyperspectral imagery acquired via unmanned aerial vehicles (UAVs) provides significantly richer spectral information than conventional RGB imagery, which can potentially improve crop discrimination.

However, applying deep learning to hyperspectral imagery presents several challenges.

One straightforward approach is to use deep classifiers that process hyperspectral pixels individually, but such methods often suffer from overfitting due to the large number of model parameters and limited labeled data. In addition, pixel-wise classification is computationally inefficient at inference time.

Semantic segmentation addresses these limitations by predicting labels for all pixels simultaneously, enabling faster inference and improved spatial consistency. Nevertheless, semantic segmentation requires detailed pixel-level annotations, which are scarce for hyperspectral agricultural imagery. Furthermore, most existing segmentation architectures are designed for RGB inputs and are pre-trained on large-scale RGB datasets, making them difficult to apply directly to hyperspectral data with hundreds of spectral channels.

Training hyperspectral segmentation models from scratch is often infeasible due to dataset scarcity, while naively modifying RGB pre-trained models may destroy useful learned representations. These challenges motivate the need for task-adapted deep learning strategies that leverage existing RGB pre-trained models while minimally modifying their architectures to accommodate hyperspectral inputs.

6.2 Task-Adapted Deep Learning for Hyperspectral Crop Mapping

To address the scarcity of large-scale hyperspectral training datasets, this work adopts a task adaptation strategy based on transfer learning and fine-tuning. Transfer learning enables a model trained on a large RGB dataset to be adapted to a related task and data modality, while retaining generalizable feature representations learned during pre-training.

We apply transfer learning to TransUNet, which incorporates a Vision Transformer (ViT) encoder pre-trained on RGB images. Many deep neural networks trained on natural images learn generic low-level features, such as edge detectors and texture patterns, that can be reused across domains. However, hyperspectral imagery requires modifications to the model’s input layers to accommodate a large number of spectral channels.

We investigate whether a dynamically adapted pre-trained model can effectively process

hyperspectral inputs while preserving most of its pre-trained knowledge. To achieve this, architectural modifications are kept minimal and localized. The input layer is adapted to accept hyperspectral imagery, and only the immediately subsequent layers are adjusted accordingly.

Introducing hundreds of input channels without proportionally increasing the model’s internal feature dimensions can create a representational bottleneck. Conversely, significantly expanding hidden layer dimensions increases computational cost and discards pre-trained weights. To address this trade-off, we incorporate a spectral channel attention mechanism that enables the network to selectively emphasize informative spectral features while maintaining a compact and efficient architecture.

6.3 Methodology

Given an HS agricultural image $x \in R^{H \times W \times C}$ with the spatial resolution of $H \times W$ and C number of channels. We aim to predict a pixel-wise label map of size $H \times W$ with class values ranging from 0 to $N - 1$. Here, N is the number of classes. The most common approach is to directly train a CNN (e.g., U-Net) to convert images into high-level abstract feature representations, which are then upsampled for feature reconstruction along the expansive path to the full spatial resolution for classification. Here, we followed the design of TransUNet [183]. Our approach differs from theirs by thoroughly examining the attention mechanisms along the contraction path. As discussed in the previous section, adapting RGB-based architectures to hyperspectral inputs introduces representational bottlenecks due to the large number of spectral channels. To address this challenge, we incorporate spectral channel attention in the contraction path. Spectral channel attention is crucial for effectively utilizing the model’s limited number of feature channels. This attention mechanism prioritizes vital features while suppressing less relevant ones, improving model performance. Figure 35 outlines the design of our proposed segmentation network.

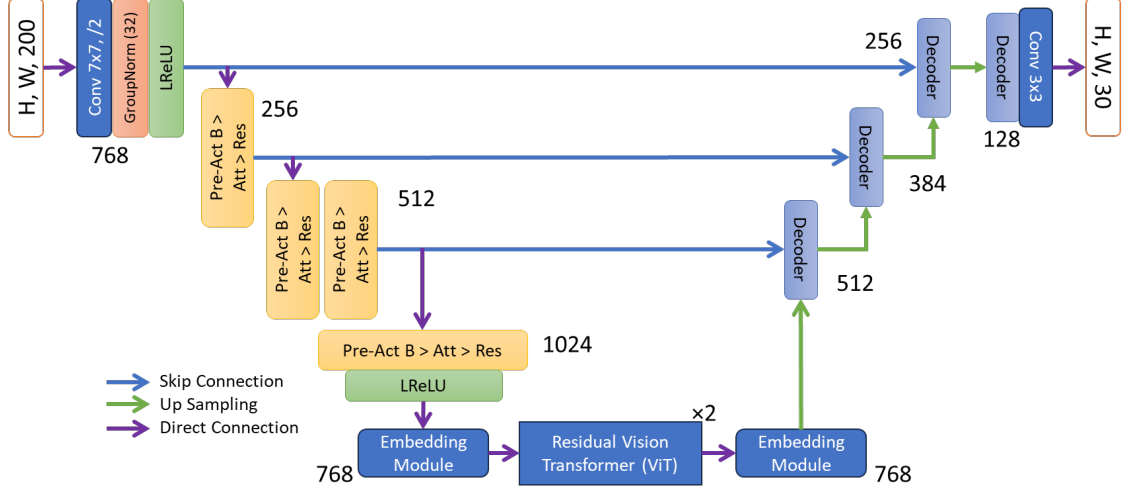


Figure 35: Outline of our proposed segmentation network.

6.3.1 Using Residual Blocks from ResNetv2 in Contraction Path

We start our contracting path with the first layer from the ResNet architecture, which uses a single 7x7 convolution with a stride of 2 (Figure 35), optimally reducing computation for the subsequent layers. It combines the advantages of three 3x3 convolutions while maintaining the same receptive field and applying stride to downsample the image. This dual function makes it more computationally efficient.

Additionally, we have updated the input channels of this first layer to accommodate images with more than three channels. We used Eq. 25 to increase the output channels from 64 to 768 adaptively. The original 64 channels were designed to extract features from just three input channels, so it is logical to raise the number of output channels in response to the increase in input channels from three to two hundred.

$$channel_{out} = 3 * 2^{\lceil \log_2(channel_{in}) \rceil} \quad (25)$$

The later units of TransUNet are Pre-Activation Bottleneck encoders from ResNet models pre-trained on ImageNet-21k and fine-tuned on ImageNet datasets [184]. The original ResNet-50

had a depth of four with [3,4,6,3] blocks, each reducing the image’s resolution by a factor of two. In the TransUNet variant, the pre-trained ResNet had a depth of three with [3,4,9] units per block, and only the first unit of each block had a 3x3 convolution layer with a stride of 2.

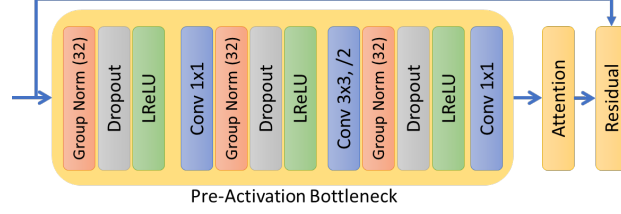


Figure 36: ResNetV2 block with Pre-Activation Bottleneck, Attention mechanism, and Residual block.

ResNet-V2 [185] concentrates on making the second non-linearity as identity mapping. Thus, the gradient calculated in the output layer can easily reach the initial layers without changing the network. We modified the original Pre-Activation Bottleneck encoders with this design, as shown in Figure 36. Our contracting path for feature abstraction has [1,2,1] units per block. Similar to the previous design, we used a 3x3 convolution layer with a stride 2 to downsample in the first unit of each block. We employed the Leaky ReLU (LReLU) as the activation function. To mitigate overfitting, we implemented an adaptive dropout technique. Equation 26 shows the way we computed the dropout probability ($drop_{prob}$) for a given number of input feature channels (c_{in}).

$$drop_{prob} = \max(0.1, \min(c_{in}/2560, 0.25)) \quad (26)$$

We used the Max Pooling instead of convolution with strides in the skip connection of the residual pathway to match the resolution of the residuals and a 1x1 convolution layer to match the number of channels. Max Pooling resolves gradient propagation issues in DNNs and introduces non-linearity to the network [186].

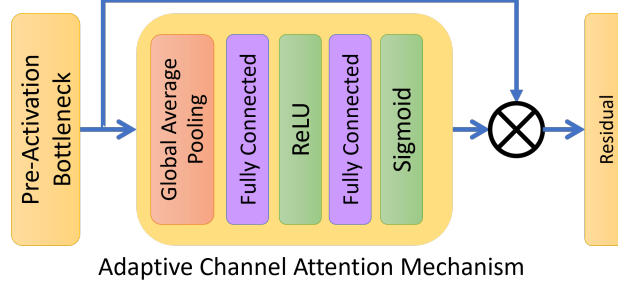


Figure 37: Adaptive Channel Attention Mechanism

6.3.2 Using Attention in Residual Blocks of ResNetv2

Hu et al. [142] first proposed the attention module, which they termed the squeeze-and-excitation (SE) block, in their newly proposed SENet. CNNs use convolutional filters to extract hierarchical feature representations from input images. In standard architectures, all feature channels are treated equally when generating output feature maps. Squeeze-and-excitation blocks address this limitation by adaptively re-weighting feature channels based on their importance. We incorporate channel attention into the ResNetv2 encoder using an adaptive reduction ratio to better capture informative spectral features. Figure 37 shows our channel attention mechanism. The model uses low-level features extracted at the initial levels to generate high-level features for object identification.

In a typical CNN architecture, the network weights each channel equally when creating the output feature maps. Squeeze-and-excitation blocks add parameters to each channel of a convolutional block, allowing the network to adaptively adjust the weighting of each feature map and recalibrate it. We added channel attention to our ResNetv2 encoder, similar to SENet. However, we updated its fixed reduction ratio of 16 in the bottleneck with an adaptive reduction ratio [143] as mentioned in Eq. 27. Here, C_{in} is the number of channels coming into the attention block. It preserves more information with fewer features and discards more when there is an increased number of input features.

$$reduction-ratio = 2 * \log_2 C_{in} \quad (27)$$

6.3.3 Transformer as Encoder Bottleneck

Our Vision Transformer (ViT) as encoder design is identical to the TransUNet [183]. The first embedding module performs tokenization [187], converting input features coming from the contraction path into the shape of $\frac{H \times W}{P^2}$, where patch size is $P \times P$. Then, it maps the vectorized patches into a latent d_{enc} -dimensional embedding space using a trainable linear projection, where learnable position embeddings are added to the patch embeddings to retain positional information. Each Transformer layer consists of Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks with layer normalization operation. These ViTs are pre-trained on ImageNet-21k and fine-tuned on ImageNet datasets. The second embedding module converts these tokenized patches back to the shape of input features.

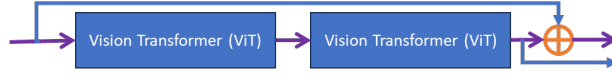


Figure 38: Design of our residual ViT

When we evaluated our model, we found that gradients are not flowing backward properly within the ViT, and 1-layer and 12-layer ViT performance is similar, which Chen et al. [183] also found in their evaluation. Thus, we took the inspiration from HSI-TransUNet [42], designed our own residual ViT, and added a skip connection to create an alternating residual ViT as seen in Figure 38. We used a 2-layer residual ViT in our network and found no significant improvement in adding more layers.

6.3.4 Expansion Path

We used a cascaded upsampler (CUP) similar to TransUNet [43]. It comprises a $2 \times$ upsampling step and two convolution blocks. Each convolution block has a 3×3 convolution layer, a Batch Normalization layer, a Dropout layer, and a ReLU layer. To mitigate overfitting,

we employed the same adaptive dropout technique here too. Before the segmentation head, we also reversed the Eq. 25 to compute the number of input channels. Equation 28 shows our updated solution for input channels. The original 16 channels were designed to provide features to the segmentation head to predict four classes. As the number of classes increased to thirty, our last CUP now has 128 channels. We subsequently updated the prior layers to adapt to this change.

$$channel_{in} = 2^{\lceil \log_2(channel_{out} * 3) \rceil} \quad (28)$$

The upsampling step in the original U-Net paper involved the upsampling feature map followed by a 2×2 convolution (“up-convolution”) that halved the number of feature channels [188]. However, this deconvolution can create checkerboard artifacts [189]. Odena et al. [189] found that nearest-neighbor interpolation provided suitable results. On the other hand, we could utilize operations with learnable weights for upsampling in the expansion path, such as sub-pixel convolutions [190]. Sub-pixel convolution takes in r^2 channels and outputs r channels with double the resolution. It would substantially increase the number of feature channels in hidden layers, creating significant contraction and expansion path changes. Thus, for the sake of simplicity, we continued using the nearest upsampling.

6.4 Experimental Setup

6.4.1 Loss Function

As shown in Eq. 29, we used a combination of Weighted (w-ce) and Categorical Cross-Entropy (cat-ce) loss. We used the “balanced” heuristic [191] to compute the class weights for the weighted cross-entropy and 0.7 as the β value. Weighted Cross-Entropy ensures that less represented classes will get higher weights when computing the loss. However, we kept the β value high to keep the $loss_{w-ce}$ low and reduce false positives from these under-represented classes.

$$loss_{comb-ce} = \beta * loss_{cat-ce} + (1 - \beta) * loss_{w-ce} \quad (29)$$

We then combined this cross-entropy loss with Log-Cosh Dice Loss function [192] for our model, as shown in Eq. 30. Our dice loss is defined as $loss_{dice} = \log(\cosh(1 - Dice_{coeff}))$, and we used 0.4 as the α value. This α value balances performance for segmentation and classification tasks.

$$loss_{total} = \alpha * loss_{dice} + (1 - \alpha) * loss_{comb-ce} \quad (30)$$

6.4.2 Training Details

Task adaptation (transfer learning and fine-tuning) in deep semantic segmentation refers to modifying a pre-trained model to perform a different yet related task without retraining from scratch [138]. Our first training step tried cross-domain transfer, label set modification, and adaptation. Transfer learning leverages knowledge from a pre-trained model on a related large dataset. We replace the last classification layer to match the number of target classes. We kept our pre-trained ResNetV2 and ViT model frozen and used them as fixed feature extractors; only the up-sampling expansive path for feature reconstruction and the final convolutional layer for classification are trained using the limited NIR-RG image dataset. We chose it due to the small size of our new dataset and the desire to leverage knowledge from a larger dataset. During this step, we used a learning rate of 0.0001. Later, we modified the input layer to handle different input modalities. We trained only the modified contracting path for feature abstraction with a learning rate 0.00001. Finally, we unfroze all layers of the model, used a very low learning rate of 0.000 001, and fine-tuned it. Fine-tuning trains the pre-trained model on a new dataset to adapt it more closely to the new task. This iterative approach allowed more adaptation to the new task, resulting in higher accuracy than general transfer learning. We used AdamW [162] as the optimizer in our training. The batch size for our training was 32. It was large enough to suppress noise but not too large to hinder learning from a small dataset. We set the max epoch to 500 with

early stopping, and all runs ended before 200 epochs.

6.5 Experimental Results on UAV Hyperspectral Imagery

Our evaluation investigated our proposed HRViTUNet (HSI-ResNetV2-ViT-UNet) model’s performance on HSI images from UAV-HSI-Crop Dataset [42]. We wrote our code in Python (PyTorch) and ran non-deterministic training four times with the same configuration on a 48GB NVIDIA RTX 6000 Ada Generation GPU with 20 CPU cores and 160GB RAM, ran evaluations on the test dataset, computed the performance metrics per test image, and averaged over all the runs.

For the qualitative evaluation, the light-blue area in Figure 39(b) represents the ground-truth target crop. Figure 39(c) presents results from our runs where our model made some errors by misclassifying the crop as background. More examples from our model are shown in Section 6.5.1. For our comparative evaluation, we trained the TransUNet [43] model using RGB and HS images with a learning rate of 0.001. We updated the data loader and kept all other settings as the default for the TransUNet ¹. As shown in Figure 39(e), the TransUNet trained with HS images performs qualitatively better than when trained with RGB images. For evaluating the HSI-TransUNet [42], we modified the code base of TransUNet ¹ to match their paper. However, we could not make it work perfectly and had difficulty reproduce their performance. Here, in the tables, we report what we achieved.



Figure 39: The visualization shows an example of semantic segmentation from our model using an image from UAV-HSI-CropDataset. (a) RGB, (b) ground-truth, (c) Our HRViTUNet fine-tuned on this agricultural HSI dataset, (d) TransUNet fine-tuned on RGB dataset, (e) TransUNet fine-tuned on HSI dataset.

¹TransUNet

We used the Dice and Jaccard coefficients to evaluate the performance of our model’s segmentation tasks. Table 25 shows these values.

Table 25: Performance comparison of HRViTUNet using segmentation metrics. Bold marks the better result than other models for that metric.

Model	Dice (mean, median)	Jaccard (mean, median)	Params (M)	Tflops (Tera)
HRViTUNet (Our)	0.749 $\pm$ 0.119, 0.749	0.737 $\pm$ 0.123, 0.732	57.70	2.19
TransUNet (RGB)	0.779 $\pm$ 0.097, 0.794	0.771 $\pm$ 0.099, 0.785	105.16	111.58
TransUNet (HSI)	0.766 $\pm$ 0.097, 0.760	0.757 $\pm$ 0.099, 0.753	105.78	111.65
HSI-TransUNet	0.643 $\pm$ 0.133, 0.648	0.631 $\pm$ 0.137, 0.633	99.33	1.18

To evaluate the performance of our model’s classification task, we used Precision, Recall, F1-score, Accuracy, and Cohen-Kappa scores. Table 26 presents these values.

Table 26: Performance comparison of HRViTUNet using classification metrics. Bold marks the better result than other models for that metric.

Model	Precision	Recall	F1-score	Accuracy	Cohen-Kappa
HRViTUNet (Our)	0.334 $\pm$ 0.166	0.552 $\pm$ 0.201	0.277 $\pm$ 0.159	0.766 $\pm$ 0.121	0.532 $\pm$ 0.222
TransUNet (RGB)	0.315 $\pm$ 0.171	0.409 $\pm$ 0.238	0.226 $\pm$ 0.168	0.606 $\pm$ 0.248	0.345 $\pm$ 0.247
TransUNet (HSI)	0.315 $\pm$ 0.154	0.470 $\pm$ 0.231	0.242 $\pm$ 0.155	0.668 $\pm$ 0.232	0.434 $\pm$ 0.270
HSI-TransUNet	0.258 $\pm$ 0.141	0.554 $\pm$ 0.216	0.205 $\pm$ 0.140	0.746 $\pm$ 0.185	0.521 $\pm$ 0.275

Our semantic segmentation model achieved 76.6% accuracy and 74.9% dice score. The confusion matrix shown in Figure 40 demonstrates the model’s effectiveness in identifying most objects, primarily because of Weighted Cross-Entropy loss. It is important to note that no examples of bok choy (class 7) or okra (class 29) exist in the test dataset. Our model’s precision is highlighted by its zero false-positive for bok choy (class 7). There are some other low-performing classes for the model, along with 100% misclassification for Kohlrabi (class 13) and sesame (class 18). Compared to the original HSI-TransUNet, we have fewer classes missing. Weighted Cross-Entropy, while providing overall lower quantitative scores than other loss functions, is a good choice if the goal is to identify the most targets, including underrepresented and hard-to-identify classes. We aimed to balance classification and segmentation accuracy and

find a decent semantic segmentation model. Our loss function and task adaptation technique successfully achieved that goal.

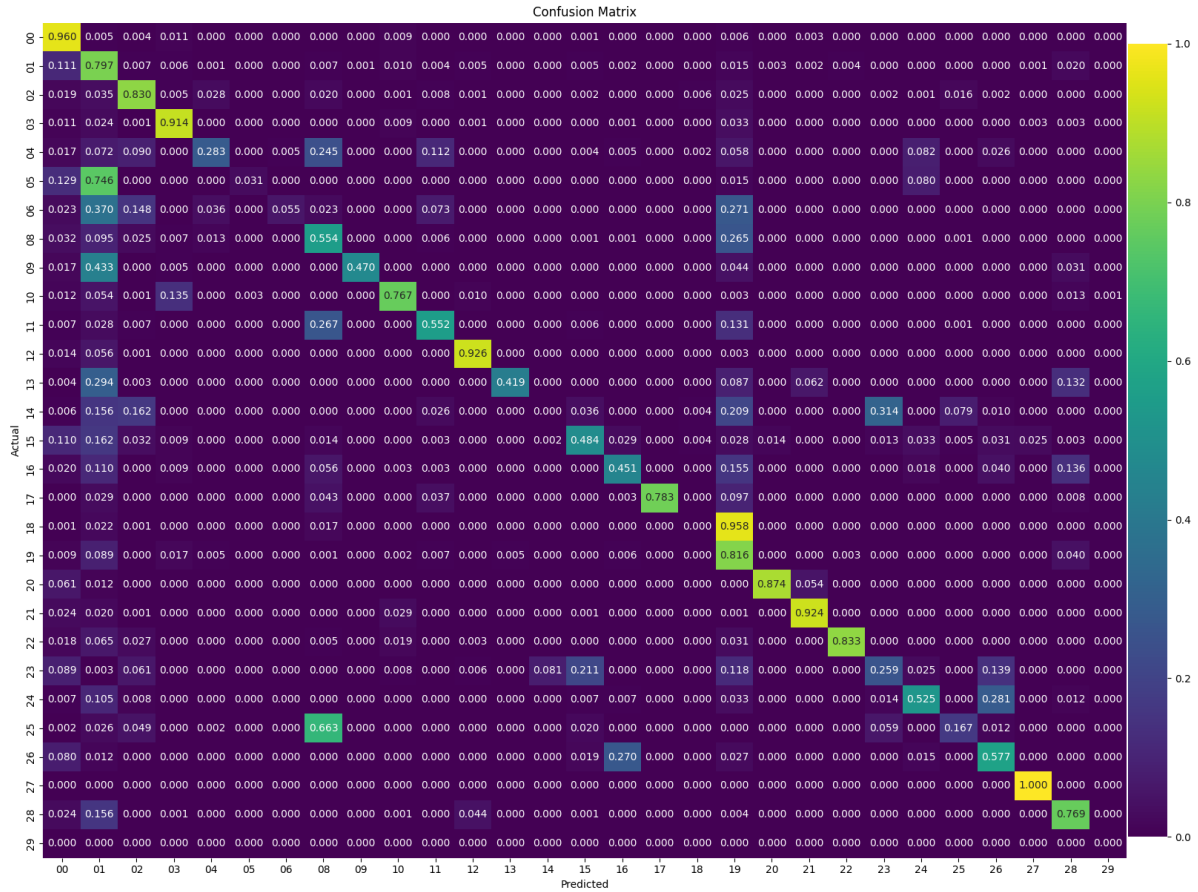


Figure 40: Confusion-Matrix from one of the runs shows low missing classes.

6.5.1 Qualitative Evaluation

Visualizing results to understand a computer vision model's performance is often helpful. Figure 41 shows four more prediction examples. Our model makes more mistakes if the scene is cluttered, whereas simple scenes with clear boundaries are mostly correctly segmented.

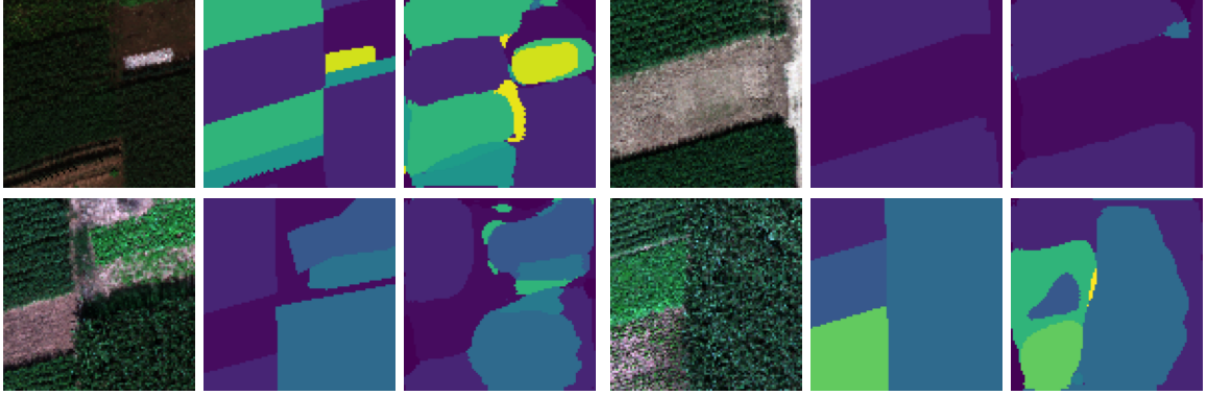


Figure 41: The visualization of our semantic segmentation model’s predictions using images from UAV-HSI-CropDataset. In each section: (a) RGB, (b) ground-truth, (c) Our HRViTUNet’s prediction.

6.6 Computational Considerations and Practical Implications

6.6.1 Model Complexity and Memory Footprint

Hyperspectral semantic segmentation models are often constrained by memory usage and computational cost due to the large number of spectral channels and high spatial resolution of UAV imagery. These constraints are particularly relevant in agricultural settings, where access to high-end computational infrastructure may be limited.

Our proposed HRViTUNet model contains 57.70 million parameters (total size of 4755.06 MB), which is significantly smaller than the standard TransUNet variants evaluated in this study. Despite incorporating both convolutional and transformer-based components, the overall model size and computational requirements remain moderate. Params’ size is 229.45 MB ². During training, we began our training with 26.36 M Params Out of 57.70 M Params and the forward and backward pass required approximately 4.3 GB (4346.45 MB) of GPU memory, while inference required substantially fewer resources.

The total number of floating-point operations (FLOPs) during training was 2.19 TFLOPs, and inference required approximately 546.57 GFLOPs ³. These values indicate that the proposed

²Computed using: PyTorch Model Size Estimator, Torch Info

³Computed using: DeepSpeed

model achieves a favorable balance between representational capacity and computational efficiency, making it suitable for practical deployment scenarios.

6.6.2 Architectural Efficiency under Hyperspectral Constraints

One of the primary challenges in hyperspectral deep learning is accommodating a large number of input channels without excessively increasing model complexity. Naively expanding convolutional layers to match the spectral dimensionality can lead to a substantial increase in parameters and memory usage, often negating the benefits of transfer learning.

To address this issue, we deliberately limited the depth of both the ResNetV2 encoder and the Vision Transformer bottleneck. Rather than increasing the width of hidden layers, we incorporated a spectral channel attention mechanism that enables the network to selectively emphasize informative spectral features. This approach enables the model to leverage hyperspectral information while maintaining a compact architecture.

By maintaining minimal structural changes to the pre-trained backbone, the proposed design avoids introducing unnecessary computational overhead and reduces the risk of overfitting on small hyperspectral datasets.

6.6.3 Transfer Learning and Weight Reuse

Preserving pre-trained weights is critical when adapting RGB-based models to hyperspectral imagery. Large architectural modifications can disrupt learned feature hierarchies and diminish the benefits of transfer learning. In this work, architectural changes were intentionally confined to the input layer and its immediately subsequent layers.

When modifications to convolutional input dimensions were unavoidable, pre-trained weights were reused wherever possible, and newly introduced parameters were initialized with the average of weights from existing pre-trained channels. This strategy minimizes distributional shifts at early layers and stabilizes training during fine-tuning.

This weight reuse mechanism enables effective cross-domain transfer while avoiding the

need for large-scale hyperspectral pre-training, which is often impractical due to data scarcity.

6.6.4 Implications for Real-World Deployment

The computational characteristics of the proposed model make it well-suited to real-world agricultural applications. UAV-based hyperspectral systems often operate with limited onboard or edge computing resources, necessitating efficient inference pipelines.

By employing a reduced-depth architecture and avoiding excessive channel expansion, the proposed approach enables faster inference and lower memory consumption than larger segmentation models. This enables practical deployment in time-sensitive applications such as crop monitoring, disease detection, yield assessment, and military surveillance.

More broadly, the design principles explored in this chapter—minimal architectural modification, selective spectral attention, and task-adapted transfer learning—are applicable beyond crop mapping. They provide a general framework for adapting RGB pre-trained models to hyperspectral downstream tasks under realistic computational constraints.

6.7 Summary and Integration with Detection Pipelines

This chapter demonstrated that hyperspectral semantic segmentation can be effectively performed as a downstream exploitation task using task-adapted deep learning models. By minimally modifying an RGB pre-trained TransUNet architecture and incorporating spectral channel attention, the proposed approach successfully transfers knowledge from natural image domains to hyperspectral agricultural imagery.

The experimental results show that competitive segmentation and classification performance can be achieved under realistic constraints, including limited labeled data and reduced computational resources. Importantly, these results demonstrate that effective hyperspectral semantic segmentation can be achieved independently of anomaly detection and without large-scale hyperspectral pre-training or excessively complex architectures.

In the broader context of this dissertation, hyperspectral semantic segmentation

complements anomaly detection by providing semantic interpretation of detected regions or by operating as a standalone scene understanding module. Future work may integrate dimensionality reduction techniques, such as hyperspectral unmixing, to further improve generalization across sensors and datasets. Together, these components form a foundation for deployable hyperspectral analysis pipelines that support both detection-driven and task-specific AiTR applications.

Chapter 7

Discussion, Conclusions, and Future Research

This dissertation investigated how robust hyperspectral anomaly detection and downstream scene understanding can be achieved under realistic operational constraints for aided and automatic target recognition. HS sensing offers rich spectral discrimination that is valuable for Intelligence, Surveillance, and Reconnaissance (ISR), agriculture, and environmental monitoring. However, the same properties that make HS data powerful—high dimensionality, sensitivity to acquisition conditions, limited labeled samples, and severe class imbalance—also complicate deployment. Models often degrade when applied across different scenes, sensors, or environmental conditions. Performance degradation across scenes, sensors, and environmental conditions remains a central obstacle. The overarching question guiding this work has therefore been: how can HS-AD systems be designed to remain accurate, stable, and interpretable under domain variability and practical UAV constraints?

A central theme of this dissertation is that robustness in HS-AD does not arise from a single architectural innovation, but from disciplined integration of signal stewardship, principled ensemble design, physical spectral representation, and domain-aware learning. Rather than treating preprocessing, modeling, and evaluation as independent stages, this work frames them as interdependent components of a unified detection pipeline.

At the signal level, anomaly detection accuracy is fundamentally constrained by the physical fidelity of the input hypercube. Sensor noise, atmospheric absorption, and corrupted spectral bands introduce distortions that propagate through subsequent modeling stages. Structured preprocessing—through SNR-guided band removal, conservative channel selection, and normalization—serves not merely as data cleaning but as an inductive bias that promotes physically plausible and statistically stable representations. Quantile normalization, in particular, improved ensemble coherence by aligning heterogeneous distributions of detector scores. This result shows that statistical comparability across models can be as influential as architectural

sophistication. These findings reinforce the principle that improving machine detection accuracy is inseparable from disciplined management of the spectral signal itself.

In addition to machine-oriented preprocessing, this dissertation addressed human interpretability through a structured enhancement-selection framework. Rather than applying generic image enhancement heuristics, enhancement was formulated as a multi-task learning problem. A shared visual backbone extracts degradation-sensitive representations, while task-specific heads predict both degradation type and the appropriate enhancement strategy. The results demonstrate that discriminative features for heterogeneous degradations can be learned even with a frozen backbone, enabling reliable enhancement selection with reduced training complexity and computational cost. This formulation elevates image enhancement from manual adjustment to data-driven decision-making and supports AiTR by improving human visibility and scene understanding prior to automated analysis.

Building upon this foundation, the dissertation introduced the Greedy Ensemble Anomaly Detection (GE-AD) framework to address the instability of individual HS anomaly detectors. Classical detectors rely on distinct background modeling assumptions and often exhibit scene-dependent variability. GE-AD formalizes ensemble construction as a greedy base-model selection problem, guided by cross-validated performance under severe class imbalance. By systematically selecting complementary detectors and adaptively weighting them through supervised or unsupervised stacking, GE-AD reduces reliance on any single modeling assumption. Across diverse datasets, this strategy showed that aggregating heterogeneous background models yields more stable and reliable detection than individual methods. The observed variability of selected base detectors across datasets further confirms that no single anomaly detector is universally optimal; adaptive ensemble composition is essential for scenario-specific robustness.

The integration of physically motivated spectral representations extends this robustness beyond ensemble diversity. Spectral unmixing-derived abundance maps and spectral similarity measures introduce material awareness into detection. Importantly, the enhanced anomalous

abundance formulation shifts the focus from predefined target spectra to spectral rarity within the scene, enabling detection of small or previously unseen objects without requiring prior target knowledge. This transition from target-dependent modeling to rarity-driven representation increases generality and supports deployment in previously unknown operational environments. By coupling ensemble learning with physically interpretable spectral features, the framework bridges statistical modeling and physical interpretability. This integration reinforces detection performance while reducing false alarms.

Recognizing that real-world deployment often involves domain shifts across sensors and scenes, the dissertation further extended GE-AD into a unified MoE stacking ensemble architecture that incorporates transfer learning, adversarial domain adaptation, and mixture-of-experts (MoE) routing. A pre-trained transformer-based feature extractor provides deep representations, while dataset-specialized FFNN-AD experts are adaptively routed across heterogeneous domains. The incorporation of domain classification as an auxiliary task and the use of adversarial alignment encourage domain-invariant feature learning that mitigates distribution mismatch. Although cross-domain performance does not fully match single-domain training scenarios—an expected consequence of unknown test-time domains—the unified model maintains competitive performance across multiple datasets and sensors without retraining. These results demonstrate that domain-aware ensemble architectures can meaningfully reduce performance degradation under realistic deployment constraints.

Further analysis of MoE routing behavior clarified that cross-domain robustness depends not only on routing accuracy but also on inter-expert similarity. Misrouting between domains with similar anomaly-detection behavior does not necessarily cause substantial degradation, whereas confusion between dissimilar experts amplifies performance loss. This observation highlights both the promise and limitations of expert-based adaptation. Robust generalization is facilitated when training data capture sufficient sensor and scene diversity; conversely, leave-one-out experiments reveal vulnerability when critical sensor characteristics are absent during training. These findings emphasize that domain diversity and complementary datasets are

foundational for transferable HS representations.

The evaluation methodology for extreme class imbalance constitutes another key contribution of this dissertation. The comparative analysis showed that ROC-AUC alone can be insufficient under severe class imbalance, as it may obscure false-positive behavior. F1-macro and precision–recall–oriented evaluation provide complementary insight into threshold-dependent performance. Repeated cross-validation and statistical testing ensured the stability and reproducibility of ensemble construction. Aligning the greedy search objective with operational metrics emerged as an essential consideration, reinforcing that evaluation design directly shapes model behavior.

Beyond anomaly detection as a cueing mechanism, this dissertation extended HS analysis to downstream semantic segmentation for agricultural scene understanding. By minimally adapting a pre-trained RGB segmentation architecture using channel-attention mechanisms and task-adaptive transfer learning, effective HS semantic segmentation was achieved under limited labeled data and computational constraints. This result shows that HS scene interpretation can be integrated into AiTR pipelines without extensive architectural redesign or large-scale HS pre-training. In the broader system perspective, anomaly detection and semantic segmentation serve complementary roles: anomaly detection identifies spectrally unusual regions, while segmentation provides semantic context and fine-grained interpretation.

Several limitations define the scope of the present work. First, ensemble performance depends on the reliability of its base detectors; severe noise or extreme domain mismatch can degrade results. Second, domain adaptation relies on sensor diversity available during training; sensor characteristics that are entirely unseen remain challenging. Third, MoE experts are independently pre-trained rather than jointly optimized end-to-end, limiting the full potential of collaborative learning. Finally, preprocessing decisions for certain datasets required manual intervention, indicating the need for more automated signal-quality assessment mechanisms. These constraints delineate the operational boundary conditions under which the reported results hold.

Future research directions naturally emerge from these findings. Joint end-to-end training of expert networks within the MoE framework may reduce routing-induced performance gaps and enable deeper cross-domain feature sharing. Exploration of alternative objective functions, such as metrics more robust to class imbalance, may further align training with operational priorities. Automated degradation detection and adaptive preprocessing could reduce manual intervention and improve deployment scalability. Additionally, expanding domain adaptation toward unsupervised domain discovery and continual learning would support long-term operation in evolving sensing environments.

Taken together, this dissertation demonstrates that robust HS target detection under real operational variability is achievable by unifying physically interpretable spectral modeling, greedy ensemble learning, domain-adaptive representation alignment, and task-adaptive transfer learning, while preserving computational efficiency suitable for UAV deployment. HS sensing is inherently information-rich yet sensitive to environmental and acquisition variability. Achieving reliability, therefore, requires not only powerful models but coordinated adaptation across signal processing, ensemble design, domain alignment, and evaluation methodology. By integrating these components into a coherent framework, this work advances scalable and deployable HS analysis systems that support both detection-driven and semantic AiTR applications under practical constraints.

References

- [1] M. Sharif, G. Han, W. Liu, X. Huang, Cultivating multidisciplinary research and education on gpu infrastructure for mid-south institutions at the university of memphis: Practice and challenge (2025). arXiv:2504.14786.
URL <https://arxiv.org/abs/2504.14786>
- [2] J. A. Ratches, Review of current aided/automatic target acquisition technology for military target acquisition tasks, *Optical Engineering* 50 (7) (2011) 072001–072001.
- [3] M. Harbaugh, Unmanned aerial systems (uas) for intelligence, surveillance, and reconnaissance (isr), State-of-the-Art-Report (SOAR), Defense Systems Information Analysis Center (DSIAC).
URL <https://dsiac.org/wp-content/uploads/2018/05/UNMANNED-AERIAL-SYSTEMSUAS-FOR-INTELLIGENCE-SURVEILLANCE-AND-RECONNAISSANCE-ISR.pdf>
- [4] M. Patterson, A. Brescia, Operation of small sensor payloads on tactical sized unmanned air vehicles, *The Aeronautical Journal* 114 (1157) (2010) 427–436.
- [5] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, A. Abbas, Modern trends in hyperspectral image analysis: A review, *IEEE Access* 6 (2018) 14118–14129.
- [6] M. Paoletti, J. Haut, J. Plaza, A. Plaza, Deep learning classifiers for hyperspectral imaging: A review, *ISPRS Journal of Photogrammetry and Remote Sensing* 158 (2019) 279–317.
- [7] J. Liu, X. Wang, Plant diseases and pests detection based on deep learning: a review, *Plant Methods* 17 (2021) 1–18.
- [8] P. S. Teh, N. Zhang, A. B. J. Teoh, K. Chen, A survey on touch dynamics authentication in mobile devices, *Computers & Security* 59 (2016) 210–235.
- [9] A. Lasisi, R. Ghazali, T. Herawan, Application of real-valued negative selection algorithm to improve medical diagnosis, in: *Applied Computing in Medicine and Health*, Elsevier, 2016, pp. 231–243.
- [10] R. Nayak, U. C. Pati, S. K. Das, A comprehensive review on deep learning-based methods for video anomaly detection, *Image and Vision Computing* 106 (2021) 104078.
- [11] D. G. Manolakis, R. B. Lockwood, T. W. Cooley, *Hyperspectral imaging remote sensing: physics, sensors, and algorithms*, Cambridge University Press, 2016.
- [12] G. Geography, Multispectral vs hyperspectral imagery explained, <https://gisgeography.com/multispectral-vs-hyperspectral-imagery-explained/>, accessed: 2024-11-23 (2014).
URL <https://gisgeography.com/multispectral-vs-hyperspectral-imagery-explained/>

- [13] H. Su, Z. Wu, H. Zhang, Q. Du, Hyperspectral anomaly detection: A survey, *IEEE Geoscience and Remote Sensing Magazine* 10 (1) (2021) 64–90.
- [14] I. Racetin, A. Krtalić, Systematic review of anomaly detection in hyperspectral remote sensing applications, *Applied Sciences* 11 (11) (2021) 4878.
- [15] N. Keshava, J. F. Mustard, Spectral unmixing, *IEEE signal processing magazine* 19 (1) (2002) 44–57. doi:10.1109/79.974727.
- [16] C.-I. Chang, *Hyperspectral data processing: algorithm design and analysis*, John Wiley & Sons, 2013.
- [17] J. A. Richards, X. Jia, *Remote sensing digital image analysis: an introduction*, Springer, 2006.
- [18] X. Hu, C. Xie, Z. Fan, Q. Duan, D. Zhang, L. Jiang, X. Wei, D. Hong, G. Li, X. Zeng, et al., Hyperspectral anomaly detection using deep learning: A review, *Remote Sensing* 14 (9) (2022) 1973. doi:10.3390/rs14091973.
- [19] C.-I. Chang, S.-S. Chiang, Anomaly detection and classification for hyperspectral imagery, *IEEE Transactions on geoscience and remote sensing* 40 (6) (2002) 1314–1325.
- [20] R. Zhao, B. Du, L. Zhang, A robust nonlinear hyperspectral anomaly detection approach, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (4) (2014) 1227–1234.
- [21] R. Zhao, Z. Shi, Z. Zou, Z. Zhang, Ensemble-based cascaded constrained energy minimization for hyperspectral target detection, *Remote Sensing* 11 (11) (2019) 1310.
- [22] J. Opitz, S. Burst, Macro F1 and Macro F1, arXiv preprint arXiv:1911.03347.
- [23] X. Kang, X. Zhang, S. Li, K. Li, J. Li, J. A. Benediktsson, Hyperspectral anomaly detection with attribute and edge-preserving filters, *IEEE Transactions on Geoscience and Remote Sensing* 55 (10) (2017) 5600–5611.
- [24] B. Hosseiny, M. Mahdianpari, M. Hemati, A. Radman, F. Mohammadimanesh, J. Chanussot, Beyond supervised learning in remote sensing: A systematic review of deep learning approaches, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2023) 1035–1052.
- [25] S. Liu, Z. Li, G. Wang, X. Qiu, T. Liu, J. Cao, D. Zhang, Spectral–spatial feature fusion for hyperspectral anomaly detection, *Sensors* 24 (5) (2024) 1652.
- [26] X. Zhang, K. Gao, J. Wang, P. Wang, Z. Hu, Z. Yang, X. Zhao, W. Li, Target detection adapting to spectral variability in multi-temporal hyperspectral images using implicit contrastive learning, *Remote Sensing* 16 (4) (2024) 718.
- [27] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, Q. He, Deep subdomain adaptation network for image classification, *IEEE transactions on neural networks and learning systems* 32 (2021) 1713–1722. doi:10.1109/tnnls.2020.2988928.

- [28] H. Liu, R. Jiang, B. Zhou, An ensemble approach for unsupervised anomaly and cyber attack detection, *Journal of Computer Security* 33 (2025) 215–238. doi:10.1177/0926227x251330216.
- [29] S. Nazat, W. Alayed, L. Li, M. Abdallah, Ensemble learning framework for anomaly detection in autonomous driving systems, *Sensors* 25 (16) (2025) 5105.
- [30] M. S. Younis, M. Hossain, A. L. Robinson, L. Wang, C. Preza, Hyperspectral unmixing-based anomaly detection, in: *Computational Imaging VII*, Vol. 12523, SPIE, 2023, p. 1252302. doi:10.1117/12.2664706.
- [31] M. Hossain, M. Younis, A. Robinson, L. Wang, C. Preza, Greedy ensemble hyperspectral anomaly detection, *Journal of Imaging* 10 (6) (2024) 131. doi:10.3390/jimaging10060131.
- [32] M. Hossain, A. Robinson, L. Wang, C. Preza, Investigation of unsupervised and supervised hyperspectral anomaly detection, in: *Applications of Machine Learning 2024*, Vol. 13138, SPIE, 2024, p. 1313817. doi:10.1117/12.3029916.
- [33] M. Hossain, A. Robinson, L. Wang, C. Preza, Improving semantic segmentation through task adaptation for uav hyperspectral agricultural imagery, in: *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping X*, Vol. 13475, SPIE, 2025, p. 1347507. doi:10.1117/12.3053426.
- [34] M. Hossain, A. Robinson, C. Preza, L. Wang, Gmoe-ad: Generalized hyperspectral anomaly detection via mixture-of-experts and domain-invariant learning, *Sensors* 26 (2026) 5661. doi:10.3390/s26175661.
- [35] H. Kwon, N. M. Nasrabadi, Kernel rx-algorithm: A nonlinear anomaly detector for hyperspectral imagery, *IEEE Transactions on Geoscience and Remote Sensing* 43 (2) (2005) 388–397.
- [36] N. Acito, M. Diani, G. Corsini, Gaussian mixture model based approach to anomaly detection in multi/hyperspectral images, in: *Image and Signal Processing for Remote Sensing XI*, Vol. 5982, SPIE, 2005, pp. 209–217.
- [37] A. Schaum, Joint subspace detection of hyperspectral targets, in: *2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No. 04TH8720)*, Vol. 3, IEEE, 2004, p. 1824.
- [38] S. Li, K. Zhang, P. Duan, X. Kang, Hyperspectral anomaly detection with kernel isolation forest, *IEEE Transactions on Geoscience and Remote Sensing* 58 (1) (2019) 319–329.
- [39] K. Tan, Z. Hou, F. Wu, Q. Du, Y. Chen, Anomaly detection for hyperspectral imagery based on the regularized subspace method and collaborative representation, *Remote sensing* 11 (11) (2019) 1318.
- [40] J. Sill, G. Takács, L. Mackey, D. Lin, Feature-weighted linear stacking, *arXiv preprint arXiv:0911.0460*.

- [41] C.-I. Chang, An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis, *IEEE Transactions on Information Theory* 46 (5) (2000) 1927–1932. doi:10.1109/18.857802.
- [42] B. Niu, Q. Feng, B. Chen, C. Ou, Y. Liu, J. Yang, Hsi-transunet: A transformer based semantic segmentation model for crop mapping from uav hyperspectral imagery, *Computers and Electronics in Agriculture* 201 (2022) 107297.
- [43] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, TransUNet: Transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306* 1.
- [44] T. P. Watson, K. McKenzie, A. Robinson, K. Renshaw, R. Driggers, E. L. Jacobs, J. Conroy, Evaluation of aerial real-time rx anomaly detection, in: *Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXIX*, Vol. 12519, SPIE, 2023, pp. 254–260.
- [45] M. K. Griffin, H.-h. K. Burke, Compensation of hyperspectral data for atmospheric effects, *Lincoln laboratory journal* 14 (1) (2003) 29–54.
- [46] G. P. Anderson, G. W. Felde, M. L. Hoke, A. J. Ratkowski, T. W. Cooley, J. H. Chetwynd Jr, J. Gardner, S. M. Adler-Golden, M. W. Matthew, A. Berk, et al., Modtran4-based atmospheric correction algorithm: Flaash (fast line-of-sight atmospheric analysis of spectral hypercubes), in: *Algorithms and technologies for multispectral, hyperspectral, and ultraspectral imagery VIII*, Vol. 4725, SPIE, 2002, pp. 65–71.
- [47] T. Wang, T. Guan, F. Qiu, L. Liu, X. Zhang, H. Zeng, Q. Zhang, Evaluation of scale effects on uav-based hyperspectral imaging for remote sensing of vegetation, *Remote Sensing* 17 (6) (2025) 1080.
- [48] A. Maity, A. Pattanaik, S. Sagnika, S. Pani, A comparative study on approaches to speckle noise reduction in images, in: *2015 International Conference on Computational Intelligence and Networks*, IEEE, 2015, pp. 148–155.
- [49] R. C. Gonzalez, R. E. Woods, *Digital image processing*, Pearson education india, 2009.
- [50] K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, Image denoising by sparse 3-d transform-domain collaborative filtering, *IEEE Transactions on image processing* 16 (8) (2007) 2080–2095.
- [51] Y. Mäkinen, L. Azzari, A. Foi, Collaborative filtering of correlated noise: Exact transform-domain variance for improved shrinkage and patch matching, *IEEE Transactions on Image Processing* 29 (2020) 8339–8354.
- [52] A. A. Green, M. Berman, P. Switzer, M. D. Craig, A transformation for ordering multispectral data in terms of image quality with implications for noise removal, *IEEE Transactions on geoscience and remote sensing* 26 (1) (1988) 65–74.

- [53] C. C. Aggarwal, Outlier ensembles: position paper, ACM SIGKDD Explorations Newsletter 14 (2) (2013) 49–58.
- [54] A. Zimek, R. J. Campello, J. Sander, Ensembles for unsupervised outlier detection: challenges and research questions a position paper, Acm Sigkdd Explorations Newsletter 15 (1) (2014) 11–22.
- [55] T. Välikangas, T. Suomi, L. L. Elo, A systematic evaluation of normalization methods in quantitative label-free proteomics, Briefings in bioinformatics 19 (1) (2018) 1–11.
- [56] Y. Zhao, L. Wong, W. W. B. Goh, How to do quantile normalization correctly for gene expression data analyses, Scientific reports 10 (1) (2020) 15534.
- [57] R. Farmer, Data Transformation.
URL <https://dataalltheway.com/posts/001-data-transformation>
- [58] S. C. Hicks, R. A. Irizarry, When to use quantile normalization?, BioRxiv (2014) 012203.
- [59] A. Ghadiri, Implementation of an automated image processing system for observing the activities of honey bees, Ph.D. thesis, Appalachian State University Boone, NC, USA (2013).
- [60] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, O. Chae, A dynamic histogram equalization for image contrast enhancement, IEEE transactions on consumer electronics 53 (2) (2007) 593–600.
- [61] D. T. Susetianingtias, S. Madenda, D. A. Rahayu, et al., Retinal microaneurysm detection using maximally stable external region algorithm, International Journal on Advanced Science, Engineering and Information Technology 6 (5) (2016) 644–648.
- [62] I. The MathWorks, Contrast limited adaptive histogram equalization, <https://www.mathworks.com/help/visionhdl/ug/contrast-adaptive-histogram-equalization.html>, accessed: 2024-12-23 (2024).
URL <https://www.mathworks.com/help/visionhdl/ug/contrast-adaptive-histogram-equalization.html>
- [63] K. Zuiderveld, Contrast limited adaptive histogram equalization, Academic Press Professional, Inc., USA, 1994, Ch. 45, p. 474–485.
- [64] H. R. Sheikh, M. F. Sabir, A. C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Transactions on image processing 15 (11) (2006) 3440–3451.
- [65] S. Athar, Z. Wang, A comprehensive performance evaluation of image quality assessment algorithms, Ieee Access 7 (2019) 140030–140070.
- [66] I. S. Reed, X. Yu, Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution, IEEE Transactions on acoustics, speech, and signal processing 38 (10) (1990) 1760–1770.

- [67] P. C. Mahalanobis, On the generalized distance in statistics, *Sankhyā: The Indian Journal of Statistics, Series A* (2008-) 80 (1) (2018) S1–S7.
- [68] A. P. Schaum, A. D. Stocker, Joint hyperspectral subspace detection derived from a bayesian likelihood ratio test, in: *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery VIII*, Vol. 4725, SPIE, 2002, pp. 225–233.
- [69] A. Schaum, Hyperspectral anomaly detection beyond rx, in: *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIII*, Vol. 6565, SPIE, 2007, pp. 13–25.
- [70] M. J. Carlotto, A cluster-based approach for detecting man-made objects and changes in imagery, *IEEE Transactions on Geoscience and Remote Sensing* 43 (2) (2005) 374–387.
- [71] P. C. Hytla, R. C. Hardie, M. T. Eismann, J. Meola, Anomaly detection in hyperspectral imagery: comparison of methods using diurnal and seasonal data, *Journal of Applied Remote Sensing* 3 (1) (2009) 033546.
- [72] B. Du, R. Zhao, L. Zhang, L. Zhang, A spectral-spatial based local summation anomaly detection method for hyperspectral images, *Signal Processing* 124 (2016) 115–131.
- [73] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: *2008 eighth ieee international conference on data mining, IEEE, 2008*, pp. 413–422.
- [74] D. Hong, N. Yokoya, J. Chanussot, X. X. Zhu, An augmented linear mixing model to address spectral variability for hyperspectral unmixing, *IEEE Transactions on Image Processing* 28 (4) (2018) 1923–1938.
- [75] C.-I. Chang, Q. Du, Estimation of number of spectrally distinct signal sources in hyperspectral imagery, *IEEE Transactions on Geoscience and Remote Sensing* 42 (3) (2004) 608–619. doi:10.1109/TGRS.2003.819189.
- [76] M. E. Winter, N-findr: an algorithm for fast autonomous spectral end-member determination in hyperspectral data, in: M. R. Descour, S. S. Shen (Eds.), *Imaging Spectrometry V*, Vol. 3753, International Society for Optics and Photonics, SPIE, 1999, pp. 266 – 275. doi:10.1117/12.366289.
URL <https://doi.org/10.1117/12.366289>
- [77] Y. Qu, R. Guo, W. Wang, H. Qi, B. Ayhan, C. Kwan, S. Vance, Anomaly detection in hyperspectral images through spectral unmixing and low rank decomposition, in: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2016, pp. 1855–1858.
- [78] S. Ravel, C. Fossati, S. Bourennane, Spectral unmixing of hyperspectral images in the presence of small targets, *Remote Sensing* 7 (1).
- [79] W. Shang, M. Jouni, Z. Wu, Y. Xu, M. Dalla Mura, Z. Wei, Hyperspectral anomaly detection based on regularized background abundance tensor decomposition, *Remote Sensing* 15 (6) (2023) 1679. doi:10.3390/rs15061679.

- [80] J. Zhang, X. Zhang, L. Jiao, Dual-view hyperspectral anomaly detection via spatial consistency and spectral unmixing, *Remote Sensing* 15 (13) (2023) 3330.
- [81] C. Chenot, M. Yaghoobi, M. E. Davies, Y. Altmann, Anomaly detection with high resolution hyperspectral observations, in: 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2018, pp. 963–967. doi:10.1109/GlobalSIP.2018.8646705.
- [82] K. Jiang, W. Xie, J. Lei, T. Jiang, Y. Li, Lren: Low-rank embedded network for sample-free hyperspectral anomaly detection, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, 2021, pp. 4139–4146. doi:10.1609/aaai.v35i5.16536.
- [83] X. Shang, Y. Duan, X. Wang, B. Fu, X. Sun, Anomaly-background separation and particle swarm optimization based band selection for hyperspectral anomaly detection, *IET Image Processing* 18 (8) (2024) 2053–2063.
- [84] Y. Lu, X. Zheng, H. Xin, H. Tang, R. Wang, F. Nie, Ensemble and random collaborative representation-based anomaly detector for hyperspectral imagery, *Signal Processing* 204 (2023) 108835.
- [85] N. Merrill, C. C. Olson, Unsupervised ensemble-kernel principal component analysis for hyperspectral anomaly detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 112–113.
- [86] X. Yang, X. Huang, M. Zhu, S. Xu, Y. Liu, Ensemble and random rx with multiple features anomaly detector for hyperspectral image, *IEEE Geoscience and Remote Sensing Letters* 19 (2022) 1–5.
- [87] S. Wang, W. Feng, Y. Quan, W. Bao, G. Dauphin, L. Gao, X. Zhong, M. Xing, Subfeature ensemble-based hyperspectral anomaly detection algorithm, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022) 5943–5952.
- [88] S. Fatemifar, M. Awais, A. Akbari, J. Kittler, A stacking ensemble for anomaly based client-specific face spoofing detection, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 1371–1375.
- [89] J. Nalepa, M. Myller, L. Tulczyjew, M. Kawulok, Deep ensembles for hyperspectral image data classification and unmixing, *Remote Sensing* 13 (20) (2021) 4133.
- [90] X. Zhang, H. Zhao, Hyperspectral-cube-based mobile face recognition: A comprehensive review, *Information Fusion* 74 (2021) 132–150.
- [91] G. Seni, J. Elder, Ensemble methods in data mining: improving accuracy through combining predictions, Morgan & Claypool Publishers, 2010.
- [92] C. C. Aggarwal, S. Sathe, Theoretical foundations and algorithms for outlier ensembles, *Acm sigkdd explorations newsletter* 17 (1) (2015) 24–47.

- [93] C. Zhang, Y. Ma, Ensemble machine learning: methods and applications, Springer, 2012.
- [94] L. Breiman, Stacked regressions, Machine learning 24 (1996) 49–64.
- [95] B. Pavlyshenko, Using stacking approaches for machine learning models, in: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), IEEE, 2018, pp. 255–258.
- [96] P. de Tchébychef, Average values, Journal of Pure and Applied Mathematics 12 (1867) 177–184.
- [97] X. Kang, Airport-Beach-Urban (ABU) Datasets, <http://xudongkang.weebly.com/data-sets.html>, , accessed: 2023-04-23 (2017).
URL <http://xudongkang.weebly.com/data-sets.html>
- [98] Y.-Q. Zhao, J. Yang, Hyperspectral image denoising via sparse representation and low-rank constraint, IEEE Transactions on Geoscience and Remote Sensing 53 (1) (2014) 296–308.
- [99] L. Zhu, G. Wen, Hyperspectral anomaly detection via background estimation and adaptive weighted sparse representation, Remote Sensing 10 (2) (2018) 272.
- [100] M. Grana, M. A. Veganzons, B. Ayerdi, Hyperspectral remote sensing scenes, Grupo de Inteligencia Computacional (GIC), Accessed: 2024-01-13.
URL https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes
- [101] W. Li, G. Wu, Q. Du, Transferred deep learning for anomaly detection in hyperspectral imagery, Git Hub, , accessed: 2024-01-13 (2019).
URL <https://github.com/yousefan/CNND/blob/master/HYDICE%5FUrban.mat>
- [102] X. Liu, M. Tanaka, M. Okutomi, Single-image noise level estimation for blind denoising, IEEE transactions on image processing 22 (12) (2013) 5226–5237.
- [103] N. J. P. Laboratory, Aviris - airborne visible / infrared imaging spectrometer, <https://aviris.jpl.nasa.gov/>, , accessed: 2024-01-23 (2006).
URL <https://aviris.jpl.nasa.gov/>
- [104] R. Basedow, E. Zalewski, Characteristics of the hydice sensor, in: Summaries of the Fifth Annual JPL Airborne Earth Science Workshop. Volume 1: AVIRIS Workshop, Vol. 1, 1995, p. 9.
- [105] D. B. Delbosque, Hypercube, spectral imagery analysis software, <https://www.erd.usace.army.mil/media/fact-sheets/fact-sheet-article-view/article/610433/hypercube/>, , accessed: 2024-04-22 (July 2015).
URL <https://www.erd.usace.army.mil/media/fact-sheets/fact-sheet-article-view/article/610433/hypercube/>

- [106] Resonon, Pika l - hyperspectral sensors - lightweight, compact vnir - benchtop systems, <https://resonon.com/Pika-L>, , accessed: 2023-09-13 (2022).
URL <https://resonon.com/Pika-L>
- [107] C. C. Aggarwal, Outlier ensembles, in: *Outlier Analysis*, Springer, 2016, pp. 185–218.
- [108] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve., *Radiology* 143 (1) (1982) 29–36.
- [109] C.-I. Chang, Comprehensive analysis of receiver operating characteristic (roc) curves for hyperspectral anomaly detection, *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022) 1–24. doi:10.1109/TGRS.2022.3211786.
- [110] r. rashmaravindrasharma, Auc-roc curve, <https://www.geeksforgeeks.org/auc-roc-curve/>, , accessed: 2024-01-13 (January 2022).
URL <https://www.geeksforgeeks.org/auc-roc-curve/>
- [111] N. Chinchor, Muc-4 evaluation metrics, in: *Proceedings of the 4th Conference on Message Understanding, MUC4 '92*, Association for Computational Linguistics, USA, 1992, p. 22–29. doi:10.3115/1072064.1072067.
URL <https://doi.org/10.3115/1072064.1072067>
- [112] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, *PloS one* 10 (3) (2015) e0118432.
- [113] L. Meegahapola, H. Hassoune, D. Gatica-Perez, M3bat: Unsupervised domain adaptation for multimodal mobile sensing with multi-branch adversarial training, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8 (2) (2024) 46:7. doi:10.1145/3659591.
- [114] A. Pillai, S. K. Nepal, W. Wang, M. Nemesure, M. Heinz, G. Price, D. Lekkas, A. C. Collins, T. Griffin, B. Buck, S. M. Preum, T. Cohen, N. C. Jacobson, D. Ben-Zeev, A. Campbell, Investigating generalizability of speech-based suicidal ideation detection using mobile phones, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7 (4). doi:10.1145/3631452.
URL <https://doi.org/10.1145/3631452>
- [115] Q. Wang, P. Bu, T. P. Breckon, Unifying unsupervised domain adaptation and zero-shot visual recognition, in: *2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2019, pp. 1–8.
- [116] M. Mancini, S. R. Buló, B. Caputo, E. Ricci, Best sources forward: domain generalization through source-specific nets, in: *2018 25th IEEE international conference on image processing (ICIP)*, IEEE, 2018, pp. 1353–1357.
- [117] J. Ryu, G. Kwon, M.-H. Yang, J. Lim, Generalized convolutional forest networks for domain generalization and visual recognition, in: *International conference on learning representations*, 2019, p. 0.

- [118] U. Muhammad, J. Laaksonen, D. Romaissa Beddiar, M. Oussalah, Domain generalization via ensemble stacking for face presentation attack detection, *International Journal of Computer Vision* 132 (12) (2024) 5759–5782.
- [119] J. Liang, R. He, T. Tan, A comprehensive survey on test-time adaptation under distribution shifts, *International Journal of Computer Vision* 133 (1) (2025) 31–64.
- [120] S. Ioffe, Batch normalization: Accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167*.
- [121] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, T. Darrell, Tent: Fully test-time adaptation by entropy minimization, *arXiv preprint arXiv:2006.10726*.
- [122] M. J. Mirza, J. Micorek, H. Possegger, H. Bischof, The norm must go on: Dynamic unsupervised domain adaptation by normalization, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14765–14775.
- [123] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, *arXiv preprint arXiv:1607.06450*.
- [124] Q. Wang, O. Fink, L. Van Gool, D. Dai, Continual test-time domain adaptation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7201–7211.
- [125] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, M. Tan, Efficient test-time model adaptation without forgetting, in: *International conference on machine learning*, PMLR, 2022, pp. 16888–16905.
- [126] J. Liang, D. Hu, J. Feng, Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, in: *International conference on machine learning*, PMLR, 2020, pp. 6028–6039.
- [127] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive mixtures of local experts, *Neural computation* 3 (1) (1991) 79–87.
- [128] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, *arXiv preprint arXiv:1701.06538*.
- [129] B. Zoph, Designing effective sparse expert models, in: *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, IEEE, 2022, pp. 1044–1044.
- [130] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, W. Fedus, ST-MoE: Designing stable and transferable sparse expert models, *arXiv preprint arXiv:2202.08906*.
- [131] R. Feng, N. Xi, D. Chu, R. Wang, Z. Deng, A. Wang, L. Lu, J. Wang, Y. Huang, Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving, *arXiv preprint arXiv:2504.19580*.

- [132] Z. Luo, W. Yang, Y. Yuan, R. Gou, X. Li, Semantic segmentation of agricultural images: A survey, *Information Processing in Agriculture* 11 (2) (2024) 172–186. doi:10.1016/j.inpa.2023.02.001.
URL <https://www.sciencedirect.com/science/article/pii/S2214317323000112>
- [133] S. Tian, Y. Dong, R. Feng, D. Liang, L. Wang, Mapping mountain glaciers using an improved u-net model with cse, *International Journal of Digital Earth* 15 (1) (2022) 463–477.
- [134] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, *arXiv preprint arXiv:1804.03999* 1.
- [135] Z. Fan, K. Liu, J. Hou, F. Yan, Q. Zang, Jaunet: A u-shape network with jump attention for semantic segmentation of road scenes, *Applied Sciences* 13 (3) (2023) 1493.
- [136] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, *Nature methods* 18 (2) (2021) 203–211.
- [137] G. Liu, L. Bai, M. Zhao, H. Zang, G. Zheng, Segmentation of wheat farmland with improved u-net on drone images, *Journal of applied remote sensing* 16 (3) (2022) 034511–034511.
- [138] D. Karimi, S. K. Warfield, A. Gholipour, Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations, *Artificial intelligence in medicine* 116 (2021) 102078.
- [139] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al., Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?, *IEEE transactions on medical imaging* 37 (11) (2018) 2514–2525. doi:10.1109/TMI.2018.2837502.
- [140] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, *Medical image analysis* 53 (2019) 197–207.
- [141] M. Amiri, R. Brooks, H. Rivaz, Fine tuning u-net for ultrasound image segmentation: which layers?, in: *MICCAI Workshop on Domain Adaptation and Representation Transfer*, Springer, 2019, pp. 235–242.
- [142] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [143] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11534–11542.

- [144] Q. Du, H. Yang, Similarity-based unsupervised band selection for hyperspectral image analysis, *IEEE geoscience and remote sensing letters* 5 (4) (2008) 564–568.
- [145] R. B. Lyngdoh, A. S. Sahadevan, T. Ahmad, P. S. Rathore, M. Mishra, P. K. Gupta, A. Misra, Avhyas: A free and open source qgis plugin for advanced hyperspectral image analysis, in: *2021 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, IEEE, 2021, pp. 71–76.
- [146] M. Younis, Investigation of Hyperspectral Data Unmixing and Its Impact on the Task of Anomaly Detection, The University of Memphis, 2023.
- [147] D. Manolakis, D. Marden, G. A. Shaw, et al., Hyperspectral image processing for automatic target detection applications, *Lincoln laboratory journal* 14 (1) (2003) 79–116.
- [148] A. M. Baldridge, S. J. Hook, C. Grove, G. Rivera, The aster spectral library version 2.0, *Remote sensing of environment* 113 (4) (2009) 711–715.
- [149] S. K. Meerdink, S. J. Hook, D. A. Roberts, E. A. Abbott, The ecostress spectral library version 1.0, *Remote Sensing of Environment* 230 (2019) 111196.
- [150] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, *ACM computing surveys (CSUR)* 41 (3) (2009) 1–58.
- [151] R. C. Gonzalez, Digital image processing, Pearson education india, 2009.
- [152] A. K. Jain, Fundamentals of digital image processing, Prentice-Hall, Inc., 1989.
- [153] G. K. Wallace, The jpeg still picture compression standard, *Communications of the ACM* 34 (4) (1991) 30–44.
- [154] C. Tomasi, R. Manduchi, Bilateral filtering for gray and color images, in: *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, IEEE, 1998, pp. 839–846.
- [155] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE transactions on image processing* 13 (4) (2004) 600–612.
- [156] A. A. Michelson, Studies in optics, Courier Corporation, 1995.
- [157] J. R. Rice, The algorithm selection problem, in: *Advances in computers*, Vol. 15, Elsevier, 1976, pp. 65–118.
- [158] K. A. Smith-Miles, Cross-disciplinary perspectives on meta-learning for algorithm selection, *ACM Computing Surveys (CSUR)* 41 (1) (2009) 1–25.
- [159] A. G. Barto, Reinforcement learning: An introduction. by richard’s sutton, *SIAM Rev* 6 (2) (2021) 423.

- [160] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, IEEE, 2009, pp. 248–255.
- [161] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision (IJCV)* 115 (3) (2015) 211–252. doi:10.1007/s11263-015-0816-y.
- [162] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 1.
- [163] D. R. Cox, The regression analysis of binary sequences, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 20 (2) (1958) 215–232.
- [164] J. R. Quinlan, C4. 5: programs for machine learning, Elsevier, 2014.
- [165] L. Breiman, Random forests, *Machine learning* 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.
- [166] T. Fushiki, Estimation of prediction error by using k-fold cross-validation, *Statistics and Computing* 21 (2011) 137–146.
- [167] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, S. Ridella, et al., The ‘k’ in k-fold cross validation., in: ESANN, 2012, pp. 441–446.
- [168] T.-T. Wong, P.-Y. Yeh, Reliable accuracy estimates from k-fold cross validation, *IEEE Transactions on Knowledge and Data Engineering* 32 (8) (2019) 1586–1594.
- [169] F. Wilcoxon, Individual comparisons by ranking methods, *Biometrics Bulletin* 1 (6) (1945) 80–83.
URL <http://www.jstor.org/stable/3001968>
- [170] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the american statistical association* 32 (200) (1937) 675–701.
- [171] E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* (1988) 837–845.
- [172] W. J. Krzanowski, D. J. Hand, ROC curves for continuous data, Chapman and Hall/CRC, 2009.
- [173] S. (<https://stats.stackexchange.com/users/22311/sycorax>), Statistical significance (p-value) for comparing two classifiers with respect to (mean) roc auc, sensitivity and specificity, Cross Validated, , uRL:<https://stats.stackexchange.com/q/358598> (version: 2018-08-01).
arXiv:<https://stats.stackexchange.com/q/358598>.
URL <https://stats.stackexchange.com/q/358598>

- [174] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, *BMC genomics* 21 (2020) 1–13.
- [175] W. Kong, B. Liu, X. Bi, C. Yu, X. Li, Y. Chen, Hypersl: A spectral foundation model for hyperspectral image interpretation, *IEEE Transactions on Geoscience and Remote Sensing*.
- [176] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: *International conference on machine learning*, PMLR, 2015, pp. 1180–1189.
- [177] A. G. Khoei, Y. Yu, R. Feldt, Domain generalization through meta-learning: a survey, *Artificial Intelligence Review* 57 (10) (2024) 285.
- [178] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, *arXiv preprint arXiv:1608.03983*.
- [179] W. Fedus, B. Zoph, N. Shazeer, Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, *Journal of Machine Learning Research* 23 (120) (2022) 1–39.
- [180] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al., Mixed precision training, *arXiv preprint arXiv:1710.03740*.
- [181] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, *The journal of machine learning research* 13 (1) (2012) 723–773.
- [182] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, *Advances in neural information processing systems* 33 (2020) 21464–21475.
- [183] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, et al., Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers, *Medical Image Analysis* 97 (2024) 103280.
- [184] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
- [185] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV* 14, Springer, 2016, pp. 630–645.
- [186] S. Sun, J. Pang, J. Shi, S. Yi, W. Ouyang, Fishnet: A versatile backbone for image, region, and pixel level prediction, *Advances in neural information processing systems* 31.
- [187] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: *International Conference on Learning Representations*, 2021, p. 0.
URL <https://openreview.net/forum?id=YicbFdNTTy>

- [188] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.
- [189] A. Odena, V. Dumoulin, C. Olah, Deconvolution and checkerboard artifacts, *Distill* 1 (10) (2016) e3.
- [190] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.
- [191] G. King, L. Zeng, Logistic regression in rare events data, *Political analysis* 9 (2) (2001) 137–163.
- [192] S. Jadon, A survey of loss functions for semantic segmentation, in: 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB), IEEE, 2020, pp. 1–7.