ITC 1/55
Information Technology and Control
Vol. 55 / No. 1/ 2026
pp. 125-142
DOI 10.5755/j01.itc.55.1.43040

Complicated Scene Classification Using a Deep Active Learning Paradigm for Wetland Remote Sensing Analysis

Received 2025/10/11 | Accepted after revision 2026/01/05

HOW TO CITE: Huang, F., Zhao, Q. (2026). Complicated Scene Classification Using a Deep Active Learning Paradigm for Wetland Remote Sensing Analysis. *Information Technology and Control*, 55(1), 125-142. https://doi.org/10.5755/j01.itc.55.1.43040

Complicated Scene Classification Using a Deep Active Learning Paradigm for Wetland Remote Sensing Analysis

Fenghua Huang*

College of Artificial Intelligence, Yango University, Fuzhou 350015, China; fhhuang@ygu.edu.cn (F.H.)

Fujian Key Laboratory of spatial information perception and intelligent processing (Yango University), Fuzhou 350015, China; e-mail: fhhuang@ygu.edu.cn (F.H.)

Fujian University Engineering Research Center of Spatial Data Mining and Application (Yango University), Fuzhou 350015, China; e-mail: fhhuang@ygu.edu.cn (F.H.)

Qianyu Zhao

Academy of Digital China (Fujian), Fuzhou University, Fuzhou 350108, China; e-mail: 245527004@fzu.edu.cn (Q.Z.)

Corresponding author: fhhuang@ygu.edu.cn

Investigating the intricate semantics of diverse wetland landscapes is vital for the development of Intelli gent computing systems within remote sensing applications. This study introduces a cutting-edge approach that utilizes memristor-based architectures to integrate multi-channel perceptual visual features for classifying wetland remote sensing images, characterized by complex spatial and ecological structures. Our method leverages a deep hierarchical model designed to emulate human gaze dynamics through a memristor-enabled processing unit, employing the BING objectness metric to accurately detect key ecological features and details across multiple scales within wetland scenes. To enhance the human-like visual attention mechanism, we propose a Memristor-Enhanced Robust Deep Active Learning (MRDAL) strategy, which systematically generates gaze shifting paths (GSPs) and extracts their deep representations using memristor-based networks. A distinctive aspect of MRDAL is its resilience against label noise, achieved through a sparse penalty mechanism embedded within the memristor architecture, effectively filtering out irrelevant GSP features. We subsequently apply a manifold-regularized feature selector (MRFS) integrated with memristive com-

ponents to extract high-quality deep GSP features, which are then utilized to train a linear Support Vector Machine (SVM) for the classification of wetland scenes. Empirical evaluations reveal the method's superior performance over conventional models, demonstrating its exceptional capability in discerning complex patterns within a comprehensive dataset of large-scale wetland remote sensing images. This advancement highlights the potential of memristor-based intelligent computing technologies for ecological monitoring and environmental analysis.

KEYWORDS: Active learning, Wetland scene, Deep feature, Feature selection, Manifold distribution.

1. Introduction

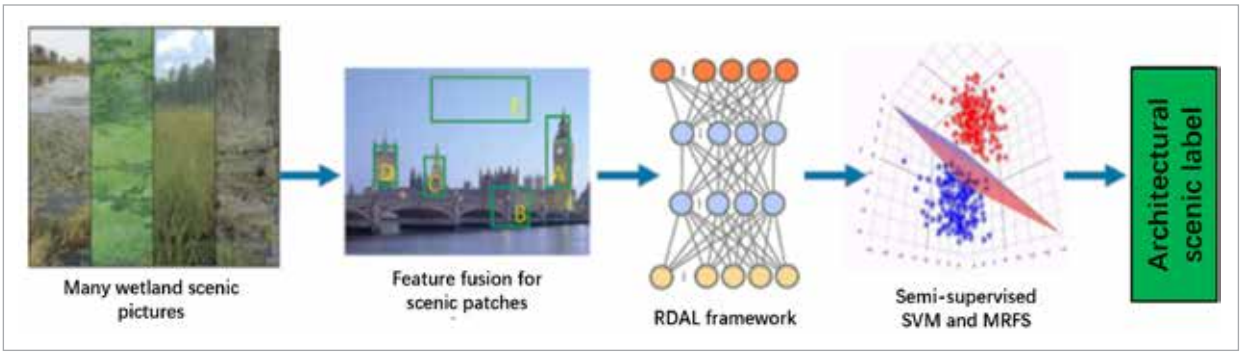
In the advanced field of AI technology, the ability to assign multiple labels to each scene is crucial for wetland remote sensing analysis. For instance, the refinement of environmental monitoring systems can benefit from identifying specific scene attributes such as water bodies, vegetation types, and land use patterns. Similarly, modern conservation efforts rely on detecting certain scene elements, such as water levels and vegetation health, to enhance real-time surveillance of ecosystems and wildlife movements. Notably, wetland ecosystems are dynamic and prone to changes, such as seasonal flooding or droughts. Thus, swift and precise categorization of various scene types is vital for deploying multi-sensor satellite monitoring systems at crucial wetland areas. Implementing this approach greatly enhances the ability to monitor unusual environmental activities, such as water level shifts or habitat degradation. In the realm of visual categorization and scene labeling for wetland analysis, several advanced algorithms have been developed to intricately capture the nuances of remote sensing imagery across various resolutions. Prominent methods include: 1) applying Multiple Instance Learning (MIL) and CNN-based strategies for region localization under weak super vision [46, 4]; 2) implementing semantically-rich graph models to parse scenes in detail [25, 23]; and 3) designing complex hierarchical structures for precise annotation of remote sensing photos [44, 13, 14]. Despite these technological strides, current approaches sometimes struggle to fully represent the complexity of wetland scenes, encountering several significant challenges:

Firstly, the detection of prominent objects or features in high-resolution scenes requires a method influenced by biological vision systems, designed to mimic the human ability to identify visually or semantically important areas. This includes precisely monitoring gaze shift paths (GSP) that reflect human attention shifts across pertinent parts of the image, tackling the problem of label noise in extensive datasets, and incorporating semantic labels at the patch level to accurately represent scene contents; Secondly, key areas in scenes are often characterized by diverse low-level descriptors, each high-lighting different scene attributes. An integrated approach that effectively combines these descriptors is crucial, necessitating a methodical strategy to evenly distribute the impact of each feature type. This task is made more complex by the necessity to fine-tune the weights of feature channels to accommodate the variety in scenic imagery collections.

To address the above challenges, we introduce an innovative framework for wetland scene classification that leverages a deep and active exploration of human gaze dynamics, incorporating the capabilities of intelligent computing technology based on new types of memristors (shown in Figure 1). Memristor-based architectures are a new type of electronic architecture that utilizes memristors as core components, aiming to break through the bottleneck of traditional von Neumann architecture and improve the efficiency of neural network model learning through hardware parallel acceleration. It is a commonly used method and generally does not need specific introduction of its hardware structure in the work. In this work, Memristor-based architectures play an important role in improving the efficiency of the complicated scene classification using a Deep Active Learning Paradigm for wetland remote sensing analysis.

In the framework above, each wetland image patch is characterized by a unique blend of diverse low-level features. Utilizing the BING objectness metric, we identify multiple object-specific patches across a wide range of remote sensing images, even when

Figure 1
The overview of our designed scenery categorization by perceptual feature selection.



label data may be inaccurate. Our model integrates memristor-based computing with human visual perception alignment through the Robust Deep Active Learning (RDAL) model. This model excels in predicting human gaze shift paths (GSPs) and extracting deep representations of these dynamics, effectively managing label noise and data redundancy, leveraging memristor's capacity for efficient data storage and computation. RDAL uses a semi-supervised learning approach, employing a subset of available semantic labels for training. We then implement a memristor-enhanced manifold-regularized feature selector (MRFS) to identify and extract highly discriminative deep GSP features, which are subsequently used to train a linear SVM tailored for the complex task of wetland scene classification. The use of memristor technology enhances the efficiency and speed of the SVM, enabling rapid processing of high-dimensional feature spaces. The effectiveness of our method and its superiority over existing models have been confirmed through comprehensive empirical tests on six publicly available datasets and a specialized dataset developed for wetland monitoring. This validation demonstrates the substantial benefits of integrating gaze-informed modeling with memristor-based computing for efficient and accurate wetland scene classification.

This research accomplishes two major advancements. First, we introduce the Robust Deep Active Learning (RDAL) framework, enhanced by memristor technology, which effectively captures and processes human gaze patterns while extracting visual features influenced by these dynamics. RDAL excels in simulating human attention through Gaze Shift Paths (GSPs) while handling label noise via a sparse penalty strategy. The integration of memristors enables the model to rapidly focus on relevant scene regions, disregarding irrelevant features with greater computational efficiency. Second, we compile a million-scale wetland remote sensing image dataset, featuring high-resolution satellite imagery of diverse wetland ecosystems globally. The dataset is organized into 12 distinct categories: marshes, swamps, bogs, fens, coastal wetlands, estuaries, mangroves, river flood-plains, seasonal wetlands, inland freshwater lakes, surrounding vegetation, and water bodies. The images are sourced from various earth observation satellites, including Landsat, Sentinel, and commercial platforms. Extensive empirical evaluations validate the superiority of our memristor-enhanced method, demonstrating significant improvements over traditional models in accurately classifying and analyzing wetland scenes, emphasizing the role of intelligent computing technology in advancing environmental monitoring.

2. Previous Work Review

In the domain of computer vision, the rise of deep learning models has radically reshaped the field of scene classification. Key to this evolution are the advancements in hierarchical Convolutional Neural Networks (CNNs) and sophisticated frameworks designed to handle large-scale image datasets, particularly ImageNet. Research highlighted in [8] and [15] has illustrated the remarkable performance of these models in scene recognition, especially when applied to subsets of ImageNet. Initially developed to address a wide range of visual tasks, the ability of ImageNet-based CNNs to extract deep features has driven

significant advancements across various computer vision applications, including video processing and anomaly detection. In recent years, improvements in CNNs inspired by ImageNet have mainly focused on expanding the volume of training data and optimizing architectural frameworks. Methods such as selective search [28] have been pivotal in generating comprehensive, category-neutral patch samples by merging search strategies with semantic labels. At the same time, Region-based CNNs (R-CNNs) [9] have concentrated on acquiring high-quality patch samples for detailed image interpretation. The development of large, scene-specific datasets for training [45] and the application of pre-trained hierarchical CNNs to identify and visualize localized scene components [31] mark significant breakthroughs. Furthermore, the integration of multi-task and multi-scale approaches into scene categorization, which involves manifold-based regularization [21] and the fusion of low-rank feature extraction with Markov models for deeper semantic exploration [19], helps preserve feature distribution consistency. Advancements towards unsupervised learning models for extracting deep features from scenic images [37] suggest further progress in using unlabeled data for training purposes. Methods combining discriminative feature learning with soft labeling techniques, incorporating multi-layer sparse autoencoders for enhanced visual representations [35], have significantly augmented the arsenal for scene classification.

Similarly, the analysis of aerial imagery has advanced through the application of cutting-edge computational models based on sophisticated machine learning techniques. The introduction of a multi-modal learning framework for annotating high-resolution aerial images [11] marks a significant advancement in this area. Research on multi-attention mechanisms [3] highlights the adaptability of these methods across different resolutions of aerial photography. Nonetheless, applying these models to low-resolution images often faces the challenge of accurately identifying small, crucial objects that may appear blurred. To address this, a shift toward region-level modeling is essential for the accurate identification and localization of key objects within low-resolution aerial views. Developments in facial recognition technologies [38], addressing the challenges of incomplete multi-view clustering [39], and employing multi-layer deep learning strategies for object detection at various scales [29] reflect ongoing innovations in aerial image analysis. A focus on high-accuracy vehicle localization [34], crafting geographic object detection models for high-resolution imagery [7], and integrating feature engineering with soft labeling techniques [36] highlight the continuous evolution and diversification of approaches designed to enhance aerial imagery analysis.

Recent advancements in deep learning have brought significant improvements to scene classification tasks by leveraging perceptual features extracted through Convolutional Neural Networks (CNNs) and other neural architectures. Kuznetsova et al. [16] provide an extensive survey on how CNN-based models have evolved to handle challenges like large intraclass variation and semantic ambiguity, which are prominent in scene classification tasks. Neupane and Aryal [26] explore scene classification in remote sensing using various deep learning architectures, including Vision Trans-formers (ViTs) and Generative Adversarial Networks (GANs), with their meta-analysis confirming the dominance of CNN-based approaches across popular datasets. Xu et al. [33] and Jiao et al. [20] further enhance CNN performance by introducing fast perception networks and two-stage feature fusion, respectively, addressing the need for efficiency in processing high-resolution aerial images. Hu et al. [12] delve into spectral clustering for better feature learning, and Thorpe and van Gennip [27] examine the boundaries of deep residual networks (ResNets), which remain popular due to their ability to train very deep models without gradient vanishing issues.

In addition to supervised learning approaches, unsupervised and semi-supervised models are gaining traction. Cheriyadat [6] applies unsupervised feature learning to aerial images, improving scene classification when labeled data is scarce, while Zhang et al. [40] use saliency-guided unsupervised learning to focus on perceptually significant regions within an image. Residual learning techniques, as explored by He et al. [10], and spectral clustering methods introduced by Hu et al. [12], both of-fer new directions for enhancing feature extraction. Scene classification continues to benefit from datasets like AID [32], which Xia et al. use to benchmark deep models. In specialized domains like food scene classification and acoustic scene analysis, hierarchical deep networks [22] and

deep scattering spectra [1] introduce domain-specific improvements, showcasing the versatility of deep learning across various scene classification tasks.

3. Methods

3.1. Detecting Semantically Meaningful Patches

In our study, we explore a crucial element of human cognition and psychology, the inherent tendency of individuals to focus on regions within an architectural scene that possess significant semantic or visual importance, as evidenced by recent studies [30, 2]. These insights reveal that attention within a scene is not uniformly distributed but concentrated on specific areas perceived as most pertinent. To tailor our scenery classification strategy to these natural human gaze behaviors, we introduce a sophisticated method that combines the identification of object-aware patches with an innovative Robust Deep Active Learning (RDAL) technique. Our goal is to accurately detect and analyze those segments of a scene that naturally engage human observers.

Empirical studies confirm that the human visual system is instinctively drawn to semantically rich or visually compelling elements, such as vehicles or architectural details, which are crucial in de-fining the perception of a scene. For effective pinpointing of these essential areas, we utilize the BING objectness measure [5], renowned for its efficiency in isolating distinct, high-quality object-specific patches within diverse scenic environments. The BING algorithm is particularly noted for its exceptional precision in identifying relevant patches with minimal computational demand, its ability to improve the Gaze Shift Paths (GSP) extraction by securing exemplary object-level patches, and its flexibility to generalize across a wide range of object categories beyond those seen during training. These attributes make our scene classification framework exceptionally versatile and effective across various datasets.

3.2. Implementing Robust Deep Active Learning (RDAL)

Utilizing the BING algorithm [5], we gather a wide array of BING object patches, from hundreds to tens of thousands, across various scenes. It is essential to note that human attention typically focuses on a small set of scene elements, indicative of a more selective perception process. In this way, we have to select those BING patches that are visually/semantically attractive to human observers, based on which we can deeply learn their deeply features for scene categorization. Aiming at this, we introduce an innovative Robust Deep Active Learning (RDAL) approach to select a few visually/semantically salient BING patches and further learn the deep features. This strategy is meticulously crafted to identify an optimal set of scenic patches, denoted by L, for constructing Gaze Shift Paths (GSP) and extracting their deep representations. The RDAL framework thoughtfully incorporates crucial elements: the spatial layout of scenes, the semantic significance of certain object patches, and the challenges of correcting potentially inaccurate semantic labels, ensuring a thorough and accurate representation of scenes that aligns with natural human observation patterns.

3.2.1. Analyzing the Spatial Structure of Architectural Scenery

Our ability to classify architectural scenery effectively is deeply linked to our understanding of the spatial structure within a scene, particularly how foreground and background elements are positioned. This understanding requires a method that evaluates the significance of each scenic patch based on its spatial relationships with surrounding patches. This representation process employs an optimization method that recognizes the spatial interconnections among patches, allowing for an enhanced depiction of scenic arrangements that is vital for effective architectural scenery classification. The optimization model is defined as Formula (1).

$$\begin{aligned} &\underset{\boldsymbol{E}}{arg\,min} \sum_{i=1}^{N} \left\| z_i - \sum_{j=1}^{N} \boldsymbol{F}_{ij} z_j \right\| \\ &s.t. \sum_{j=1}^{N} \boldsymbol{F}_{ij} = 1, \boldsymbol{F}_{ij} = 0 \; if \; z_i \notin N(z_j). \end{aligned} \tag{1}$$

In this model, we represent the collection of deep features extracted from N architectural scenic patches identified by the BING algorithm [5] as $z_1,\ldots,z_N \in R^{N\times A}$,where A denotes the depth of feature representation for each patch. The matrix $\boldsymbol{F}_{ij}$ quantifies how much the i-th patch contributes to the reconstruction

of the j-th patch. Furthermore, $N(z_i)$ includes patches that are spatially close to the i-th patch, emphasizing the spatial coherence of the scene.

3.2.2. Assessing the Semantic Value of Architectural Scenic Patches

Recognizing the semantic significance of the selected scenic patches is crucial in creating Gaze Shift Paths (GSPs). Using the specified reconstruction error in Formula (1), we can outline the reconstructed scenic patches as $g_1,...,g_N$. The process of identifying the most semantically significant L patches is carried out for minimizing the objective function as Formula (2):

$$\theta(h_1,\cdots,h_N)=\sum_{i=1}^{N}\sum_{j=1}^{L}\left\|h_i-h_{q_i}\right\|^2+\gamma\sum_{j=1}^{N}\left\|g_i-\sum_{j=1}^{N}\boldsymbol{F}_{ij}g_j\right\|^2. \quad (2)$$

In this model, γ acts as the regularization parameter, and the set includes the L scenic patches selected through our RDAL approach. The goal is to minimize the reconstruction error, thus maintaining the spatial and semantic coherence of the selected patches. This method ensures that the reconstructed patches accurately represent the visual and semantic elements of the original scene. By optimizing Formula (2), we select a subset of patches that effectively mimic human perception of the scene.

For analytical purposes, we introduce the matrices $\boldsymbol{A}=[z_i]$ and $\boldsymbol{H}=[h_i]$, and define Δ, a matrix that is diagonal, as an indicator of scenic patch selection. Specifically, Δ_{ii} is set to 1 for indices i within the set $q_1,...,q_L$, indicating the chosen patches, and 0 for all other indices. This setup refines the objective function as Formula (2) in the following manner: $\Delta_{ii}=1$ for $i\in\{q_1,...,q_L)$ and 0 otherwise, facilitating an improved formulation of the objective function as Formula (3).

$$\eta(\boldsymbol{Q})=tr\left(\left(\boldsymbol{H}-\boldsymbol{A}\right)^T\Delta(\boldsymbol{H}-\boldsymbol{A})\right)+\tau tr\left(\boldsymbol{H}^T\boldsymbol{L}\boldsymbol{H}\right). \quad (3)$$

In this case, $\boldsymbol{L}=(\boldsymbol{I}-\boldsymbol{F})^T(\boldsymbol{I}-\boldsymbol{F})$ serves as a key component of our optimization strategy. To efficiently optimize Formula (3), we calculate the gradient of $\eta(\boldsymbol{H})$ and set it to zero, which leads us to Formula (4):

$$\Delta(\boldsymbol{H}-\boldsymbol{A})+\tau\boldsymbol{L}\boldsymbol{H}=0. \quad (4)$$

In this setting, the reconstructed scenic patches are determined through Formula (5):

$$\boldsymbol{H}=(\tau\boldsymbol{L}+\Delta)^{-1}\Delta\boldsymbol{A}. \quad (5)$$

Utilizing the reassembled scenic patches, we refine the reconstruction error through the process described below as Formula (6):

$$\eta(z_{q_1},\cdots,z_{q_k})=\|\boldsymbol{Z}-\boldsymbol{G}\|_F^2=\left\|\boldsymbol{Z}-(\tau\boldsymbol{K}+\Delta)^{-1}\Delta\boldsymbol{Z}\right\|_F^2=\left\|(\tau\boldsymbol{K}+\Delta)^{-1}\tau\boldsymbol{K}\boldsymbol{Z}\right\|_F^2. \quad (6)$$

Here, the Frobenius norm is represented as $\|\cdot\|_F^2$, which is a type of matrix norm.

3.2.3. The RDAL Framework Explained

Our approach employs a multilayered deep learning strategy to identify and dissect the visual traits inherent to various scenes. As illustrated in Figure 2, the RDAL framework operates using a deep model consisting of R layers, which systematically breaks down the semantic label matrix into a series of factor matrices from $\boldsymbol{U}_1$ to $\boldsymbol{U}_R$, along with $\boldsymbol{V}$. This methodical decomposition aids in the extraction of scene-specific deep features and equips the model with the capability to precisely characterize new scenes, starting with the initial equation $\boldsymbol{U}_1=\boldsymbol{W}_1\boldsymbol{X}$ at the foundational layer.

The core of our RDAL technique is rooted in the application of successive linear combinations to unveil the latent attributes unique to each scene, steering clear of overly intricate models. This streamlined approach enhances the interpretation of scene dynamics, as demonstrated in the structured design of our deep learning framework as Formula (7).

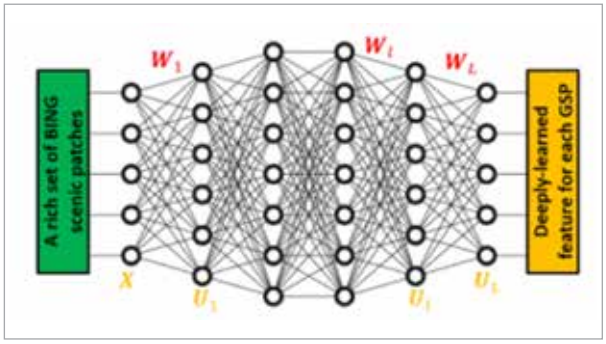
$$\boldsymbol{G}\leftarrow\boldsymbol{P}\boldsymbol{Q}_R,\boldsymbol{Q}_R=\boldsymbol{U}_R\boldsymbol{P}_{R-1},\cdots\boldsymbol{Q}_1=\boldsymbol{U}_1\boldsymbol{Y}. \quad (7)$$

In our framework, each matrix $\boldsymbol{U}_1$ serves as the transformation matrix for the i-th layer, while $\boldsymbol{P}$represents the matrix of semantic labels, which might not be explicitly observable. The matrix $\boldsymbol{Q}_i$ encapsulates the scene's representation at the i-th deep layer. Additionally, $\boldsymbol{Y}$ comprises y_i, which collects the B-dimensional features of each scenic patch. The deepest and most comprehensive representation, achieved at the topmost layer, is denoted by $\boldsymbol{Q}=\boldsymbol{Q}_L$. During the training process, as outlined in Formula (7), our deep

learning approach strives to uncover the latent factor $\boldsymbol{P}$ along with the progression of transformation matrices from $\boldsymbol{U}_R$ to $\boldsymbol{U}_1$, enabling a nuanced and layered exploration of the scene's characteristics.

Figure 2

Details of the deeply encoded and semantically rich Gaze Shift Path (GSP) model.



In summary, the comprehensive active learning process, guided by our deep model, can be articulated through the following mathematical framework as Formula (8):

$$\min_{\boldsymbol{P},\Delta\boldsymbol{U}_1,\cdots,\boldsymbol{U}_R} \frac{1}{2}\|\boldsymbol{F}-\boldsymbol{PQ}\|_F^2+\frac{\alpha}{2}\|\boldsymbol{P}\|_F^2+\frac{\alpha}{2}\sum_{i=1}^{R}\|\boldsymbol{U}_i\|_F^2+\frac{\beta}{2}\|\boldsymbol{U}\|_{2,1}. \quad (8)$$

Within this framework, the matrix $\boldsymbol{F}\in R^{R\times N}$ captures semantic labels, with $\boldsymbol{F}_{ij}=1$ indicating that the i-th scenic image is associated with the j-th label, and $\boldsymbol{F}_{ij}=0$ denoting the absence of such a link. Here, R represents the total number of distinct semantic labels, α serves as the regularization parameter to mitigate overfitting, and β promotes sparsity within the columns of $\boldsymbol{U}_i$. Considering the potential for visual features to be overlapping, redundant, or contaminated by noise, the integration of a sparse model using the l_{21}-norm is essential. This approach effectively filters out low-quality, noisy features. The solution of Formula (8) is detailed further in Section 3.3.

3.3. Step-by-Step Solution of Formula (8)

There are three Objective Components in Formula (8):

- The first term, $\frac{1}{2}\|\boldsymbol{F}-\boldsymbol{PQI}\|_F^2$, is the reconstruction error of the factorization. The goal is to minimize the difference between the original matrix $\boldsymbol{F}$ and the product of the factors $\boldsymbol{P}$, $\boldsymbol{Q}$ and $\boldsymbol{I}$.
- The terms $\frac{\alpha}{2}\|\boldsymbol{P}\|_F^2$ and $\frac{\alpha}{2}\sum_{i=1}^{R}\|\boldsymbol{U}_i\|_F^2$ apply L_2 regularization (Ridge) on the matrices to control overfitting and ensure numerical stability.
- The last term, $\frac{\beta}{2}\|\boldsymbol{U}\|_{2,1}$, promotes group sparsity through the $L_{2,1}$ norm. This encourages certain rows of the matrix to become zero, which can help in feature selection or structured sparsity.

The following approach can be used to solve Formula (8):

3.3.1. Alternating Minimization

Since Formula (8) involves multiple variables $\boldsymbol{P}$, $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_R$, a common approach is to use alternating minimization. This involves fixing all but one variable, solving for the remaining one, and iterating. Steps: 1) Fix $\boldsymbol{P}$ and $\boldsymbol{U}_i$, and solve for $\boldsymbol{Q}$ by minimizing the objective function with respect to $\boldsymbol{Q}$. 2) Fix $\boldsymbol{Q}$ and$\boldsymbol{U}$i, and solve for $\boldsymbol{P}$. 3) Fix $\boldsymbol{P}$ and $\boldsymbol{Q}$,and solve for $\boldsymbol{U}_i$ for eachi. 4) The Frobenius norm-based terms allow direct solutions through gradient descent or closed-form updates, while the $l_{2,1}$ norm term typically requires more specialized algorithms (e.g., proximal gradient descent).

3.3.2. Proximal Gradient Descent

The $l_{2,1}$ norm requires using proximal gradient methods due to its non-smooth nature. The update for $\boldsymbol{U}_i$ would use a soft-thresholding operation during gradient descent to handle the sparsity inducing regularizer.

3.3.3. Gradient Descent

For the smooth parts of the objective function, such as $\frac{1}{2}\|\boldsymbol{F}-\boldsymbol{PQI}\|_F^2$, standard gradient descent or stochastic gradient descent (SGD) can be used to update the variables.

General Steps: (1) Initialize the variables $\boldsymbol{P}$, $\boldsymbol{Q}$, $\boldsymbol{U}_i$ randomly or through some heuristic. (2) **Iterate:** 1) Update $\boldsymbol{Q}$ by solving the corresponding minimization subproblem (using gradient descent or closed-form solutions). 2) Update $\boldsymbol{P}$ similarly. 3) Update $\boldsymbol{U}_i$ using proximal gradient descent to handle the $l_{2,1}$ norm. (3) **Convergence:** Continue the iterations until the objective value converges.

Importantly, diverging from prior analyses that focus on the spatial organization of scenery and semantic parsing at the patch level, our RDAL method operates within a semi-supervised learning frame-

work. This implies that the model's training relies on only a subset of available semantic labels as delineated in Equation (8), offering a substantial benefit for analyzing large image collections where exhaustive semantic labeling is unfeasible due to the sheer volume of manual annotation required.

Afterward, we tackle the challenges of manifold feature selection and classification in the context of high-dimensional deep Gaze Shift Path (GSP) features. The main challenges addressed include: 1) reducing the dimensionality of features when large numbers of image patches are selected, which makes classifier training difficult, and 2) overcoming the scarcity of labeled samples by using both labeled and unlabeled data for learning. A semi-supervised feature selection [41] strategy is employed, combining labeled and unlabeled scenic images to train a classifier. The feature matrix represents multiple $\boldsymbol{C}$ path representations, and a graph modeling approach using manifold regularization ensures the smoothness of the decision function across samples.

The proposed model incorporates a linear SVM, where the regularization term integrates graph Laplacians, and an error margin η_i balances the regularization effects. Additionally, a semi-supervised feature selection (FS) approach is integrated, where a selection matrix $\boldsymbol{E}$ modifies the sample representation based on chosen features. The optimization problem is framed as a min-max problem, designed to minimize error while selecting an optimal feature subset and ensuring spatial coherence in the sample distribution.

4. Results

This section investigates the efficacy of our Robust Deep Active Learning (RDAL) approach for classifying wetland scenes across four diverse experimental configurations. We begin by outlining the experimental setup and introducing six benchmark datasets that serve as the foundation for our assessments. After establishing the framework, we perform a comparative study to evaluate the performance of our RDAL model against various traditional and contemporary deep learning-based scene recognition algorithms. Moreover, we explore key parameters and analyze their impact on the RDAL model's effectiveness. In the final part of this analysis, we showcase how the deep Gaze Shift Path (GSP) features, derived through our RDAL technique, significantly improve the classifi-cation of architectural scenes.

4.1. Data Sets and Setting

Our model is rigorously tested across six diverse scenic image datasets, combining traditional benchmarks with contemporary collections to highlight its adaptability. Illustrative images from these datasets, displayed in Figure 3, exemplify the broad spectrum of scenes evaluated. Crucially, two fundamental datasets, Scene-15 [17] and MIT Indoor Scene-67[24], serve as key benchmarks for assessing our model's performance. The evaluation encompasses:

- **Scene-15:** This dataset includes 15 distinct scene categories, initially compiled by Feifei [18], featuring 200 to 400 images per category, with a typical resolution of pixels. Images are primarily sourced from the COREL database, supplemented by individual photographers and Google images.
- **Scene-67:** Focused on indoor scenes, this dataset spans 67 varied categories of interior spaces, sourced from Picasa, Altavista for diverse indoor settings, various photography sharing sites, and the extensive LabelMe database. It provides a comprehensive insight into a variety of private and public indoor environments. Additionally, our extensive evaluation incorporates four modern scenic image collections that offer varied perspectives on scenic imagery:
  - ZJU Aerial Imagery [42] offers a distinct aerial perspective on landscapes and urban areas, capturing the complexity of terrain and infrastructure from above.
  - ILSVRC-2010 [8], part of the broader ImageNet project, focuses on a diverse array of scenes and objects, aiding in the development and testing of sophisticated image recognition algorithms.
  - SUN [43] provides a detailed examination of a wide range of natural and man-made environments, emphasizing the diversity and depth of scene comprehension.
  - Places [45] features a vast collection of images intended to train and evaluate algorithms in scene recognition, covering a comprehensive array of indoor and outdoor settings.

Figure 3
Sample Images from Six Scene Datasets.

Additionally, we introduce a specially curated dataset named Massive-Scale Wetland Imagery (MSWI). This dataset is a customized dataset including approximately 976,000 images across 12 wetland categories, as outlined in Table 1, designed to support research and educational applications in ecological monitoring and remote sensing. The dataset provides extensive coverage of wetland ecosystems, offering valuable insights into various wetland types and their spatial characteristics. In MSWI, the various remote sensing images of wetland scenes are primarily sourced from public datasets. Prior to the experiments, all the aforementioned sample images underwent preprocessing steps such as radiometric calibration, geometric correction, topographic correction, and atmospheric correction. Currently, this dataset is not publicly available for download, and interested readers can contact the authors of this paper. Statistics of this data set MSWI is detailed in Table 1, demonstrating its diverse and comprehensive categorization.

Table 1
Details of our MSWI image set.

Wetland category	Training image#	Validation image#
Marshes	83434	23165
Swamps	78546	25876
Bogs	79943	43234
Fens	87657	27765
Coastal wetlands	69485	34657
Estuaries	73442	26576
Mangroves	78794	27440
River floodplains	77453	28854
Seasonal wetlands	82184	29849
Inland freshwater lakes	76340	27345
Surrounding vegetation	78345	28560
Water bodies	81230	29350

This dataset provides a broad exploration of wetland ecosystems, supporting a range of applications from environmental monitoring to research in wetland conservation and management. To set the groundwork for our baseline model evaluations, we have carefully designed our experimental setup to provide a fair and thorough comparison of all involved algorithms. Our key configurations include:

1. **Object Patch Extraction:** Using the BING algorithm, we extract 1000 scene patches from each dataset to ensure a thorough evaluation of object detection performance across a diverse range of environments.
2. **Spatial Neighbor Configuration:** For every object patch, five neighboring regions are considered to mirror the human tendency to focus on a small number of areas within a scene, capturing important spatial relationships.
3. **Low-Level Feature Extraction:** Three distinct sets of low-level features are employed to effectively capture the visual essence of each patch: 16-dimensional color moments, 64-dimensional Histogram of Oriented Gradients (HOG), and a 160-dimensional edge and color histogram. These features are selected for their capacity to represent essential visual attributes.
4. **Internal Gaze Shift Path (GSP) Regions:** Each GSP is configured with five internal regions. This design reflects the natural human inclination to focus on a few key areas within a scene, ensuring that our model aligns with realistic attention patterns.
5. **Dimensionality of Deep Features at Patch Level:** The dimensionality of the deep features at the patch level is standardized to 212. This ensures that the features are both manageable and consistent, facilitating effective and efficient classification.

These methodological choices are intended to replicate human perceptual processes, thereby enhancing the accuracy and relevance of our scene categorization model across various visual contexts.

4.2. Comparison with Other Recognition Models

4.2.1. Scene Categorization Task

In our analysis, we compared our advanced scene categorization method with a variety of established shallow and deep learning-based classifiers to assess its effectiveness. For shallow classifiers, we evaluated the following:

1. **Fixed-Length Walk Kernel (FWK) and Tree Kernel (FTK):** These models are designed to capture image structural patterns, with FTK extending FWK by integrating hierarchical data structures.
2. **Multi-Resolution Histogram (MRH):** This method analyzes textures across multiple scales for detailed texture-based classification.
3. **Spatial Pyramid Matching with Kernel Methods (SPM):** Variants such as Locality-constrained Linear Coding with SPM, Sparse Coding with SPM, and Object Bank with SPM were included, each enhancing SPM with distinct feature encoding strategies for improved scene representation.
4. **Super Vector Coding (SVC) and Supervised Image Coding (SSC):** These techniques refine image representation through advanced vector quantization and supervised learning strategies.

Each model was optimized to ensure a fair comparison. For example, the parameters for FWK and FTK were fine-tuned for maximum performance, while MRH utilized RBF smoothing at different grayscale levels to boost texture analysis accuracy.

For the deep learning evaluations, we examined leading models such as ImageNet CNN, Region-based CNN, Meta Object CNN, Deep Mining CNN, and Spatial Pyramid Pooling CNN. The M-CNN model underwent specific adjustments, including selecting an optimal number of region proposals per image and standardizing feature representation using the FC7 layer of a well-known CNN. Additionally, superpixels per scene were generated and enhanced using Linear Discriminant Analysis or by identifying visually significant patches. We used average classification accuracy as the primary performance metric, representing the ratio of correctly classified scene images to the total.

Our Robust Deep Active Learning (RDAL) framework notably improved the analysis by integrating low-level features to identify key superpixels that are both semantically and visually important, known as Gaze Shift Paths. This innovative approach enabled the construction of a graph-based superpixel framework, critical to our scene classification kernel machine. The clear advantage of our method, particularly the use of BING-derived patches instead of traditional superpixels, was demonstrated, highlighting our

Table 2

Categorization Accuracies for Different Models on Various Benchmarks.

Benchmark	XFWK	XFTK	XMRH	XPM	XLLC-SP	XSC-SP	XOB-SP	XSV	XSSC
Dataset-15	73.1%	74.2%	65.8%	76.0%	78.1%	80.2%	75.4%	81.0%	86.3%
Dataset-67	42.6%	41.2%	35.4%	46.5%	49.2%	50.1%	49.3%	47.2%	53.1%
Aerial-ZJU	68.7%	70.1%	64.2%	74.1%	77.6%	79.3%	76.9%	79.2%	84.1%
ImageNet-2010	33.8%	31.4%	29.3%	33.2%	38.0%	36.8%	38.1%	37.7%	39.2%
Set-SUN397	17.8%	16.2%	15.1%	23.4%	40.1%	41.3%	39.1%	36.4%	41.9%
Urban-Places	23.4%	22.9%	21.2%	28.1%	32.4%	33.1%	32.3%	32.5%	33.1%
MS-WI	48.8%	49.5%	52.0%	47.3%	52.0%	55.7%	48.1%	52.6%	54.1%
Dataset	**XIN-CNN**	**XR-CNN**	**XM-CNN**	**XDM-CNN**	**XSPP-CNN**	**XSP-S**	**XSP-GBVS**	**XSP-LDA**	**XMesnil**
Dataset-15	84.4%	87.2%	88.9%	89.5%	92.3%	90.6%	87.2%	86.8%	87.4%
Dataset-67	58.2%	69.0%	72.5%	68.9%	67.3%	77.2%	72.1%	71.8%	73.0%
Aerial-ZJU	77.3%	79.8%	80.2%	81.0%	78.9%	81.5%	79.8%	81.7%	80.6%
ImageNet-2010	36.4%	39.1%	40.8%	41.1%	42.0%	41.6%	40.9%	40.8%	40.9%
Set-SUN397	49.8%	48.0%	51.2%	49.5%	52.1%	52.4%	51.3%	51.2%	51.5%
Urban-Places	41.2%	44.1%	44.3%	45.2%	49.4%	50.1%	48.8%	48.7%	49.3%
MS-WI	53.5%	51.2%	53.5%	55.4%	57.8%	54.3%	59.0%	61.2%	62.9%
Dataset	**Xiao**	**Cong**	**Fast R-CNN**	**Faster R-CNN**	**Ours**				
Dataset-15	84.9%	88.4%	91.0%	92.5%	92.6%				
Dataset-67	72.1%	72.3%	71.8%	74.1%	75.8%				
Aerial-ZJU	81.8%	81.4%	79.8%	81.4%	84.0%				
ImageNet-2010	41.8%	42.6%	41.8%	41.5%	44.8%				
Set-SUN397	51.8%	52.4%	54.1%	53.4%	56.3%				
Urban-Places	49.8%	48.2%	48.9%	50.2%	52.1%				
MS-WI	60.2%	61.5%	63.1%	64.4%	73.9%				

Table 3
Derivation Results for Various Models on Different Datasets.

Benchmark	XFWK	XFTK	XMRH	XSP	XLLC-SP	XSC-SP	XOB-SP	XSV	XSSC
Dataset-15	0.014	0.011	0.011	0.016	0.017	0.019	0.012	0.014	0.013
Dataset-67	0015	0.013	0016	0.015	0.015	0.014	0.014	0.015	0015
Aerial-ZJU	0.015	0.016	0.017	0.016	0.017	0.016	0.015	0.014	0.015
ImageNet-2010	0.015	0.014	0.014	0.014	0.015	0.014	0.013	0.014	0.015
Set-SUN397	0.013	0.015	0.015	0.014	0.015	0.016	0.017	0.014	0.016
Urban-Places	0.014	0.015	0.016	0.015	0.017	0.015	0.017	0.016	0.018
MS-WI	0.014	0.010	0.015	0.012	0.009	0.013	0.013	0.014	0.013
Dataset	**XIN-CNN**	**XR-CNN**	**XM-CNN**	**XDM-CNN**	**XSPP-CNN**	**XSP-S**	**XSP-GBVS**	**XSP-LDA**	**XMesnil**
Dataset-15	0.017	0.014	0.015	0.015	0.016	0.014	0.015	0.014	0.016
Dataset-67	0.014	0.016	0.014	0.014	0.015	0.014	0.016	0.014	0.013
Aerial-ZJU	0.014	0.015	0.016	0.015	0.014	0.015	0.014	0.017	0.015
ImageNet-2010	0.016	0.014	0.015	0.014	0.016	0.019	0.014	0.016	0013
Set-SUN397	0.014	0.015	0.016	0.013	0.015	0.013	0.015	0.015	0.016
Urban-Places	0.013	0.015	0.013	0.014	0.014	0.015	0.014	0.013	0.014
MS-WI	0015	0.012	0.016	0.015	0.007	0014	0.012	0.017	0.016
Dataset	**Xiao**	**Cong**	**Fast R-CNN**	**Faster R-CNN**	**Ours**				
Dataset-15	0.013	0.015	0.014	0.015	0.010				
Dataset-67	0.018	0.013	0.014	0.014	0.008				
Aerial-ZJU	0.015	0.014	0.015	0.013	0.009				
ImageNet-2010	0.014	0.014	0.015	0.012	0.010				
Set-SUN397	0.013	0.014	0.015	0.014	0.010				
Urban-Places	0.014	0.013	0.015	0.013	0.009				
MS-WI	0.014	0.012	0.015	0.013	0.008				

model's enhanced descriptiveness and robustness in scene categorization, as evidenced by comparisons with state-of-the-art scene categorization methods.

Following a thorough evaluation of the results in Tables 2-3, we conducted a detailed statistical analysis to compare our approach against a broad range of scene recognition models, including both modern deep learning technologies and traditional visual recognition techniques. This comparative study involved running each test 20 times to ensure the reliability of the results, with standard deviation metrics calculated to assess consistency. The analysis clearly demonstrates that our proposed method outperforms other techniques in terms of classification accuracy and stability. Notably, when tested on our exclusive MSWI dataset, our Robust Deep Active Learning (RDAL) method exhibited superior performance, surpassing the closest competing model by over 8%. This substantial margin high- lights the effectiveness of our approach, especially in specialized or niche datasets where precision and detailed detection are paramount.

In summary, the following observations were made:

1. Our RDAL-based scene classification model significantly outperforms traditional shallow recognition models for several reasons: 1) Models such as SPM, SC-SPM, LLC-SPM, SV, and SSC rely predominantly on SIFT descriptors, which struggle to integrate multi-channel visual descriptors. Moreover, with the exception of SSC, these models generally overlook data from additional channels, such as color moments. 2) The object detectors used in OB-SPM are trained on generic object categories, which may not provide sufficient detail for specific scene categories. 3) RHM, which focuses on capturing coarse structural information, tends to perform poorly. 4) Due to the inherent instability in graph-based descriptors, both walk and tree kernels fail to provide sufficiently descriptive features for scene modeling.
2. Our perception-guided scenery recognition algorithm competes effectively against various deep learning models, attributed to two main factors: 1) Our method selectively targets perception-aware salient regions within a scene, while other deep learning models often use entire images, randomly select patches, or identify discriminative patches only during training. 2) Our RDAL effectively simulates human gaze behavior, offering valuable insights into how different scenes are perceived, a feature generally lacking in competing models.
3. Despite relying on a limited set of semantically labeled scenic images, our approach remains competitive with other semantic scene recognition models, showcasing RDAL's robustness in bridging the semantic gap in scenery understanding. Additionally, the ability to extract semantics from a small subset of weakly labeled images proves crucial for enhancing scene recognition capabilities.

4.3. Tuning Parameters for Enhanced Performance

To elevate the performance of our scene recognition model, we undertook a comprehensive parameter tuning process aimed at optimizing the model's effectiveness in classifying scenes. This process involved meticulous adjustments and evaluations of several key parameters to identify the optimal settings. Specifically, we focused on three critical parameters:

1. L-Proximity of Object Patches: We experimented with varying the number of neighboring patches involved in reconstructing any given object patch to assess its influence on the model's scene recognition capabilities.
2. K-Number of Object Patches in GSP: We evaluated how changing the number of object patches included in a Gaze Shift Path (GSP) affects the model's performance.
3. Regularization Parameters α, β, γ: These parameters play a crucial role in the model's regularization framework. They help adjust the model's behavior to prevent overfitting and enhance its general applicability across different scenarios.
4. The tuning efforts were primarily conducted using the Scene-15 dataset [31] because of its manageable size, which avoids the computational burdens associated with larger datasets. This focused approach allowed us to determine the most effective parameter configurations that would significantly improve the model's accuracy and robustness in scene categorization. Through this detailed examination, our goal was to establish the ideal settings that would boost the model's precision and longevity in accurately classifying scenes. Fine-tuning the parameter L, which represents the number of neighboring patches used in reconstructing each scenic patch, is a crucial optimization task in our model. Maintaining the proximity of object patches is essential for the effectiveness of our feature fusion approach. We conducted a thorough assessment of L, exploring a range from one to fifteen, to determine its impact on scene recognition accuracy. The results, depicted in Figure 4, show a clear trend: as L increases, the accuracy of scene recognition initially improves peaking when L is between three to five, before declining with larger values.

This pattern suggests that an optimal range of three to five neighboring patches is ideal for accurate scenery reconstruction. Specifically, within the Scene-15 dataset, it was observed that scenic patches typically interact with three to five adjacent patches, confirming this range as effective for maintaining scene locality. Additionally, Figure5 highlights the negative consequences of expanding the number of neighboring patches beyond this range, as it can introduce excessive noise and irrelevant information, thereby reduc-

Figure 4
Categorization precision by adjusting L.

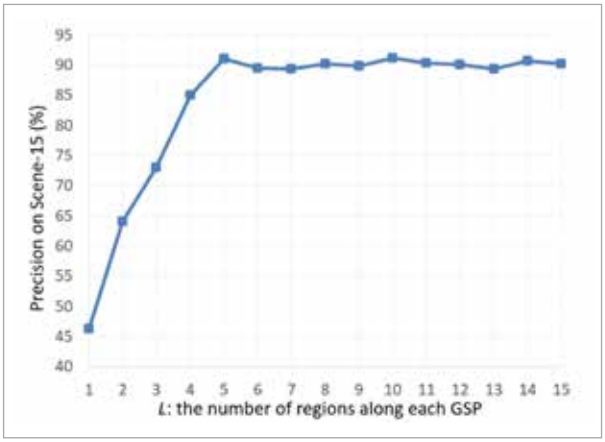
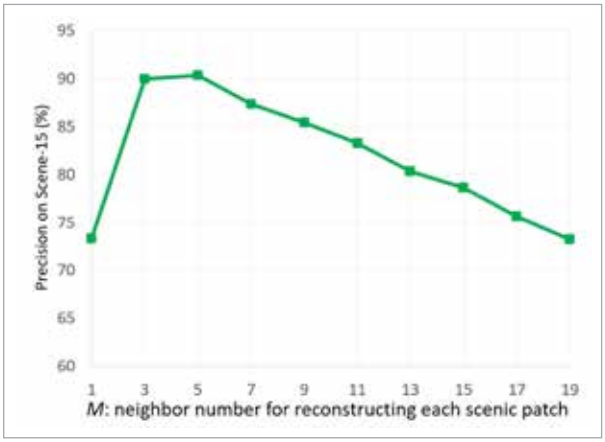


Figure 5
Categorization precision by adjusting M



ing the model's accuracy and efficiency. In our study, we assess the impact of the regularization parameters α, β, and γ on the scene categorization process. Initially setting each parameter to 0.1, we then adjust them independently to find their optimal values. We begin by modifying α from 0 to 0.95 and monitor its effect on the accuracy of scene classification. As illustrated in Table 4, there is a consistent increase in accuracy until α=0.25 , beyond which the performance starts to decline. This suggests that a moderate increase in α can enhance the model's resistance to overfitting; however, excessively high values may adversely affect the model's ability to efficiently manage feature sparsity and the semantic analysis of scene patches. Therefore, α=0.25 is established as the ideal value.

We continue this approach with β and γ, exploring how adjustments to these parameters affect scene classification, as detailed in Tables 5-6. Following a similar procedure as with α, the optimal settings for β and γ are determined to be 0.3 and 0.2, respectively. This precise tuning of parameters allows our model to effectively balance accuracy and generalization capabilities, optimizing it for the complex requirements of scene classification.

Table 4
Performance of Categorization with Varying α Values.

α	Accuracy	α	Accuracy
0	72.10%	0.04	70.30%
0.08	77.50%	0.12	66.00%
0.16	79.20%	0.20	64.10%
0.24	82.40%	0.28	61.20%
0.32	85.00%	0.36	57.40%
0.40	87.60%	0.44	53.70%
0.48	85.90%	0.52	50.00%
0.56	83.20%	0.60	46.60%
0.60	81.50%	0.68	43.20%
0.64	80.10%	0.72	41.10%
0.68	75.50%		

Table 5
Performance of Categorization with Varying β Values.

β	Accuracy	β	Accuracy
0	74.10%	0.04	76.30%
0.08	78.30%	0.12	73.50%
0.16	81.00%	0.20	72.00%
0.24	82.20%	0.28	70.10%
0.32	84.50%	0.36	69.90%
0.40	85.70%	0.44	71.00%
0.44	86.00%	0.52	68.80%
0.48	87.40%	0.56	67.00%
0.60	85.60%	0.64	64.30%
0.64	83.70%	0.72	61.50%
0.68	81.90%		

Table 5

Performance of Categorization with Varying γ Values.

γ	Accuracy	γ	Accuracy
0	75.90%	0.04	77.10%
0.08	79.40%	0.12	74.10%
0.16	85.30%	0.20	72.50%
0.24	86.60%	0.28	71.10%
0.32	89.00%	0.36	69.30%
0.36	86.30%	0.44	70.00%
0.48	83.90%	0.52	68.30%
0.52	82.00%	0.60	67.00%
0.56	80.10%	0.64	65.60%
0.60	78.30%		

5. Discussion

First, the computational efficiency of the proposed in this work needs to be further improved. Due to the large scale of remote sensing images and the need for multiple iterations in active learning to query the optimal representative image blocks, in each iteration of active learning, the entire large images or unlabeled datasets need to be inferred to calculate the uncertainty or information of the relevant image blocks. Therefore, the computational cost is relatively high, and compared with ordinary machine learning methods, the advantage in computational cost is not obvious.

Second, due to the complexity of surface object features, different wetland scenes may exhibit the same or similar spectrum characteristics in remote sensing images, the same or similar wetland scenes may exhibit different spectrum characteristics. Active learning often selects the most blurry and difficult to be identified areas in the remote sensing images, and some areas may even be difficult for the related experts to accurately label. Therefore, using the RDAL model proposed in this work for classify overly complex wetland scenes may result in errors. The performance of the complex scene classification method for wetland remote sensing analysis based on deep active learning paradigm proposed by our work still needs to be improved.

Third, the proposed method is specifically tailored for wetland remote sensing analysis, it cannot be indirectly used to other complex scenes classification because of the different active learning strategies in different scenes. The generalization of RDAL model can be improved by retraining based on more representative samples from the different scenes.

6. Conclusions

Accurately classifying wetland scenes is essential for various remote sensing and ecological applications. This research introduces an approach called Robust Deep Active Learning (RDAL), specifically tailored for wetland remote sensing analysis. RDAL creates an advanced image kernel that effectively captures areas of wetland images that are visually and semantically significant. The method begins by assembling a large-scale dataset of wetland scenes, covering different ecosystems such as marshes, swamps, and estuaries. Through RDAL, we identify key areas of these scenes that are relevant for analysis, establishing Gaze Shift Paths (GSPs) that represent important visual features. These deep GSP features are further processed and selected using the Manifold-Regularized Feature Selector (MRFS) algorithm. A linear classifier is then trained to categorize wetland scenes based on the refined feature set. The robustness and precision of this approach are validated through extensive evaluations, demonstrating its effectiveness in classifying complex wetland ecosystems. Of course, there are some shortcomings in the proposed method, the authors will try their best to further improve the computational efficiency, classification accuracy and generalization of RDAL model in the future work.

Author Contributions

Conceptualization, F.H.; methodology, F.H.; validation, F.H. and Q.Z.; formal analysis, F.H.; investigation, Q.Z.; data curation, Q.Z.; writing—original draft preparation, F.H.; writing—review and editing, F.H. and Q.Z.; supervision, F.H.; funding acquisition, F.H. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Fujian Province Science and Technology Plan Project (Guided Project, 2023N0021), the National Natural Science Foundation of China (NSFC, 41501451), China Postdoctoral Science Foundation (2015M571963).

References

1. Balestriero, R., Baratin, A., Denoyer, L., Gallinari, P. Deep Scattering Spectra With Deep Neural Networks for Acoustic Scene Classification Tasks. Chinese Journal of Electronics, 2019, 28(6), 1177-1183 https://doi.org/10.1049/cje.2019.07.006
2. Bruce, N. D. B., Tsotsos, J. K. Saliency, Attention, and Visual Search: An Information Theoretic Approach. Jo-urnal of Vision, 2009, 9(5), 1-24. https://doi.org/10.1167/9.3.5
3. Cai, W. W., Wei, Z. G., Li, H., Zhang, Q., Zhao, X., Wang, F. Remote Sensing Image Classification Based on a Cross-Attention Mechanism and Graph Convolution. IEEE Geoscience and Remote Sensing Letters, 2022, 19, 1-5. https://doi.org/10.1109/LGRS.2020.3026587
4. Cao, L. J., Luo, F., Chen, L., Sheng, Y. H., Zhang, J., Li, X., Wang, Q., Huang, T. Weakly Supervised Vehicle De-tection in Satellite Images via Multi-Instance Discriminative Learning. Pattern Recognition, 2017, 64, 417-424. https://doi.org/10.1016/j.patcog.2016.10.033
5. Cheng, M.-M., Lin, W.-L.,m Zhang, Z., Rosin, P. L., Torr, P. H. S. BING: Binarized Normed Gradients for Object-ness Estimation at 300fps. Computational Visual Media, 2019, 5(1), 3-20. https://doi.org/10.1007/s41095-018-0120-1
6. Cheriyadat, A. M. Unsupervised Feature Learning for Aerial Scene Classification. IEEE Transactions on Geos-cience and Remote Sensing, 2014, 52(1), 439-451. https://doi.org/10.1109/TGRS.2013.2241444
7. Costea, D., Leordeanu, M. Aerial Image Glocalization From Recognition and Matching of Roads and Inter-sections. arXiv: Computer Science/Computer Vision and Pattern Recognition, 2016, 1-6. https://doi.org/10.5244/C.30.118
8. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, June 20-25, 2009, 248-255. https://doi.org/10.1109/CVPR.2009.5206848
9. Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Se-mantic Segmentation. In Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), Columbus, OH, USA, June 23-28, 2014, 3214-3221. https://arxiv.org/pdf/1311.2524v1.
10. He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, June 27-30, 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90
11. Hong, D. F., Gao, L. R., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B. More Diverse Means Better: Multi-modal Deep Learning Meets Remote-Sensing Imagery Classification. IEEE Transactions on Geoscience and Re-mote Sensing, 2021, 59(5), 4340-4354. https://doi.org/10.1109/TGRS.2020.3016820
12. Hu, F., Xia, G. S., Wang, Z., Huang, X., Zhang, L., Sun, H. Unsupervised Feature Learning via Spectral Clustering of Multidimensional Patches for Remotely Sensed Scene Classification. IEEE Journal of Selected Topics in Ap-plied Earth Observations and Remote Sensing, 2015, 8(5), 2015-2030. https://doi.org/10.1109/JSTARS.2015.2444405
13. Kampffmeyer, M., Salberg, A. B., Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncer-tainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2016), Las Vegas, NV, USA, June 26 - July 01, 2016, 680-688. https://doi.org/10.1109/CVPRW.2016.90
14. Kemker, R., Salvaggio, C., Kanan, C. Algorithms for Semantic Segmentation of Multispectral Remote Sensing Imagery Using Deep Learning. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 145, 60-77. https://doi.org/10.1016/j.isprsjprs.2018.04.014
15. Krizhevsky, A., Sutskever, I., Hinton, G. E. ImageNet Classification With Deep Convolutional Neural Networks. Communications

of the ACM, 2017, 60(6), 84-90. https://doi.org/10.1145/3065386

16. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. Deep Learning for Scene Classification: A Survey. arXiv: Computer Science/Computer Vi-sion and Pattern Recognition, 2021, 1-24. https://arxiv.org/pdf/2101.10531.
17. Lazebnik, S., Schmid, C., Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York, NY, USA, June 17-22, 2006, 2169-2178. https://doi.org/10.1109/CVPR.2006.68
18. Li, F. F., Perona, P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Die-go, CA, USA, June 20-25, 2005, 524-531. https://doi.org/10.1109/CVPR.2005.16
19. Li, X. L., Mou, L. C., Lu, X. Q. Scene Parsing From an MAP Perspective. IEEE Transactions on Cybernetics, 2015, 45(9), 1876-1886. https://doi.org/10.1109/TCYB.2014.2361489
20. Liu, Y., Liu, Y., Ding, L. Scene Classification Based on Two-Stage Deep Feature Fusion. IEEE Geoscience and Remote Sensing Letters, 2018, 15(2), 183-186. https://doi.org/10.1109/LGRS.2017.2779469
21. Lu, X. Q., Li, X. L., Mou, L. C. Semi-Supervised Multi-Task Learning for Scene Recognition. IEEE Transactions on Cybernetics, 2015, 45(9), 1967-1976. https://doi.org/10.1109/TCYB.2014.2362959
22. Martinez, E. T., Leyva-Vallina, M., Sarker, M. K., Puig, D., Petkov, N., Radeva, P. Hierarchical Approach to Classify Food Scenes in Egocentric Photo-Streams. IEEE Journal of Biomedical and Health Informatics, 2020, 24(3), 866-877. https://doi.org/10.1109/JBHI.2019.2922390
23. Porway, J., Wang, Q. C., Zhu, S. C. A Hierarchical and Contextual Model for Aerial Image Parsing. International Journal of Computer Vision, 88(2), 254-283. https://doi.org/10.1007/s11263-009-0306-1
24. Quattoni, A., Torralba, A. Recognizing Indoor Scenes. In Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, June 20-25, 2009, 413-420. https://doi.org/10.1109/CVPR.2009.5206537
25. Shu, T., Xie, D., Rothrock, B., Todorovic, S., Zhu, S. C. Joint Inference of Groups, Events and Human Roles in Aerial Videos. In Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, June 07-12, 2015, 4576-4584. https://doi.org/10.1109/CVPR.2015.7298706
26. Thapa, A., Horanont, T., Neupane, B., Aryal, J. Deep Learning for Remote Sensing Image Scene Classification: A Review and Meta-Analysis. Remote Sensing, 2023, 15(19), 4804-4840. https://doi.org/10.3390/rs15194804
27. Thorpe, M., van Gennip, Y. Deep Limits of Residual Neural Networks. Research in the Mathematical Sciences, 2023, 10(6), 1-44. https://doi.org/10.1007/s40687-022-00370-y
28. Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., Smeulders, A. W. M. Selective Search for Object Recogni-tion. International Journal of Computer Vision, 2013, 104(2), 154-171. https://doi.org/10.1007/s11263-013-0620-5
29. Wang, C., Bai, X., Wang, S., Zhou, J., Ren, P. Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters, 2019, 16(2), 310-314. https://doi.org/10.1109/LGRS.2018.2872355
30. Wolfe, J. M., Horowitz, T. S. What Attributes Guide the Deployment of Visual Attention and How Do They Do It? Nature Reviews Neuroscience, 2004, 5, 495-501. https://doi.org/10.1038/nrn1411
31. Wu, R. B., Wang, B. Y., Wang, W. P., Yu, Y. Z. Harvesting Discriminative Meta Objects With Deep CNN Featu-res for Scene Classification. In Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, December 07-13, 2015, 1287-1295. https://doi.org/10.1109/ICCV.2015.152
32. Xia, G. S., Hu, F., Wang, Z. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classifica-tion. IEEE Transactions on Geosci-

ence and Remote Sensing, 2017, 55(7), 3965-3981. https://doi.org/10.1109/TGRS.2017.2685945

33. Xu, D., Jiao, L., Zhao, J., An, J. A Fast Deep Perception Network for Remote Sensing Scene Classification. Re-mote Sensing, 2020, 12(4), 729-741. https://doi.org/10.3390/rs12040729. https://doi.org/10.3390/rs12040729
34. Yang, M. Y., Liao, W. T., Li, X. B., Rosenhahn, B. Deep Learning for Vehicle Detection in Aerial Images. In Pro-ceedings of 2018 25th IEEE International Conference on Image Processing (ICIP 2018), Athens, Greece, October 07-10, 2018, 3079-3083. https://doi.org/10.1109/ICIP.2018.8451454
35. Yao, X. W., Han, J. W., Cheng, G., Qian, X. M., Guo, L. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(6), 3660-3671. https://doi.org/10.1109/TGRS.2016.2523563
36. Yu, Y., Yang, X., Li, J., Gao, X. B. Object Detection for Aerial Images With Feature Enhancement and Soft Label Assignment. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-16. https://doi.org/10.1109/TGRS.2022.3177255
37. Yuan, Y., Mou, L. C., Lu, X. Q. Scene Recognition by Manifold Regularized Deep Learning Architecture. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10), 2222-2233. https://doi.org/10.1109/TNNLS.2014.2359471
38. Zhang, C., Li, H. X., Chen, C. L., Qian, Y. H., Zhou, X. Z. Enhanced Group Sparse Regularized Nonconvex Reg-ression for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5), 2438-2452. https://doi.org/10.1109/TPAMI.2020.3033994
39. Zhang, C., Li, H. X., Lv, W., Huang, Z. Z., Gao, Y., Chen, C. L. Enhanced Tensor Low-Rank and Sparse Represen-tation Recovery for Incomplete Multi-View Clustering. In Proceedings of the AAAI Conference on Artificial In-telligence (AAAI 2023), Washington, DC, USA, February 7-14, 2023, 11174-11182. https://doi.org/10.1609/aaai.v37i9.26323
40. Zhang, F., Du, B., Zhang, L. Saliency-Guided Unsupervised Feature Learning for Scene Classification. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(4), 2175-2184. https://doi.org/10.1109/TGRS.2014.2357078
41. Zhang, J. R., Lin, X., Liu, Y. Manifold-Regularized Feature Selector for High-Resolution Aerial Photographs Categorization. IEEE Access, 2024, 12, 41354-41363. https://doi.org/10.1109/ACCESS.2024.3377241.
42. Zhang, L., Han, Y. H., Yang, Y., Song, M. L., Yan, S. C., Tian, Q. Discovering Discriminative Graphlets for Aerial Image Categories Recognition. IEEE Transactions on Image Processing, 2013, 22(12), 5071-5084. https://doi.org/10.1109/TIP.2013.2278465
43. Zhang, L. Y., Tong, M. H., Marks, T. K., Shan, H. H., Cottrell, G. W. SUN: A Bayesian Framework for Saliency Using Natural Statistics. Journal of Vision, 2008, 8(32), 1-20. https://doi.org/10.1167/8.7.32
44. Zheng, Z., Zhong, Y. F., Wang, J. J., Ma, A. L. Foreground-Aware Relation Network for Geospatial Object Seg-mentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, USA, June 13-19, 2020, 1-10. https://doi.org/10.1109/CVPR42600.2020.00415
45. Zhou, B. L., Lapedriza, A., Xiao, J. X., Torralba, A., Oliva, A. Learning Deep Features for Scene Recognition Using Places Database. In Proceedings of Advances in Neural Information Processing Systems (NIPS 2014), Montreal, Quebec, Canada, December 8-13, 2014, 1-9. http://s.dic.cool/S/0jmWJf55.
46. Zhou, S., Irvin, J., Wang, Z. C., Zhang, E., Aljubran, J., Deadrick, W., Rajagopal, R., Ng, A. Deep-Wind: Weakly Supervised Localization of Wind Turbines in Satellite Imagery. In Proceedings of the 33rd Conference on Neu-ral Information Processing Systems (NeurIPS 2019), Vancouver, Canada, December 8-14, 2019, 1-5. http://s.dic.cool/S/hXNsrATN.