IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 10, October, 2025, pp. 3483-3492
Journal homepage: http://iieta.org/journals/mmep

Deep Learning Based Remote Sensing Image Annotation Based on Semantic and Contextual Information

Prajakta Ugale[1,2*], Poonam Railkar[1]

[1] Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune 411041, India
[2] Department of Computer Engineering, MIT Academy of Engineering, Pune 412105, India

Corresponding Author Email: pvugale@mitaoe.ac.in

https://doi.org/10.18280/mmep.121015

Received: 14 August 2025
Revised: 23 September 2025
Accepted: 28 September 2025
Available online: 31 October 2025

Keywords:
remote sensing image annotation, content-based image retrieval, deep learning, YOLOv8

ABSTRACT

Remote sensing images are widely used in various applications, including military, agriculture, surveillance, urban planning, forestry, land cover classification, vegetation mapping, and land mapping. Image annotation is crucial for content-based remote sensing image retrieval (CBRSIR) for enhancing the retrieval performance. This paper presents the CBRSIR based on an improved YOLOv8 framework that considers the semantic, contextual, and spatial correlation properties of objects. It utilizes the Deformable Convolutional Network (DCN) to enhance and acquire finer details of objects and spatial context by adaptively adjusting the receptive fields. The spatial channel-wise convolution module (SCConv) is utilized to capture the channel-wise semantic and correlation information. The effectiveness of the algorithm is assessed on the UC merged dataset. The inclusion of DCN and SCConv modules in the YOLOv8 framework enhances its performance, achieving a recall of 98.50%, a precision of 98.40%, an F1-score of 98.45%, and an accuracy of 98.15%. These improvements mark a significant advancement in remote sensing object detection and annotation compared to conventional approaches.

1. INTRODUCTION

Content Based Image Retrieval (CBIR) is an advanced method for finding and retrieving images from extensive collections by analyzing the actual visual elements within the images. Instead of relying on keywords or manually added tags, CBIR analyzes features such as color patterns, textures, shapes, and the arrangement of objects within the image to identify similar content [1]. This makes the search process more accurate and automated. It is particularly valuable in fields such as medical diagnostics, biometric identification, digital archives, and security systems, where retrieving the right image quickly and precisely is essential. By focusing on the actual content of an image, CBIR significantly enhances the relevance and effectiveness of image searches, making it a vital tool in modern image processing and computer vision applications [2].

In CBIR, image annotation plays a crucial role in enhancing both the accuracy and utility of the system. CBIR typically relies on features such as color, texture, and shape to compare and identify images [3]. However, these features are low-level and don't always reflect the actual meaning or content of the image. Image annotation helps close this gap by adding descriptive tags or labels to images, thereby providing the system with a better understanding of their content. This makes it easier for the system to match user queries with the right images [4]. Additionally, annotated images provide valuable training data for supervised learning, enabling the system to become smarter and more accurate over time. In short, annotation adds meaning to visual data, making CBIR systems more effective and relevant in delivering the right results [5].

Over time, image annotation has evolved from a slow, manual process to a smart, automated one, dramatically improving the functionality of CBIR systems. Initially, annotating images involved manually adding descriptive tags, which was tedious, inconsistent, and unsuitable for managing growing image collections [6]. To ease this burden, researchers introduced semi-automated techniques that allowed machines to assist humans, making the process faster and more reliable [7]. As machine learning advanced, primarily through supervised learning, it became possible to train systems to label images automatically using example data. The rise of deep learning, particularly convolutional neural networks (CNNs), has taken things further by enabling machines to understand complex image patterns and generate accurate labels directly from image content [8]. More recently, the use of natural language processing (NLP) has enabled systems to describe images using complete sentences or captions, thereby facilitating a more effective connection between visual data and human language. These advancements have made image annotation smarter, more efficient, and more meaningful, which is particularly valuable in fields such as healthcare, security, and digital archiving [9].

Image annotation in CBIR faces several key challenges that impact the system's overall performance. One major issue is the semantic gap, which highlights the difference between low-level image features (such as color or texture) and the high-level concepts that users desire [10]. Manual annotation is time-consuming, labour-intensive, and prone to human error, especially when dealing with large datasets. Automated annotation methods often struggle with accuracy, particularly when images are complex or contain multiple objects [11]. There is also a lack of standardized vocabularies or ontologies, which can lead to inconsistent labeling. Additionally, subjective interpretation of images can lead to variations in annotations among different users. Handling ambiguous or noisy images further complicates the annotation process [12]. Scalability becomes a problem when millions of images need to be labeled or processed. Moreover, keeping annotations updated as users' needs evolve is another challenge. These factors make image annotation a critical yet complex component of CBIR systems [13].

Various YOLO frameworks have been utilized for object detection and annotation in images, demonstrating notable improvements over traditional annotation methods. YOLOv8 is a crucial tool for object detection in videos and images. However, when it comes to detecting the remote sensing objects, YOLOv8 fails to detect the smaller objects or occluded objects in the images. The spatial connectivity and correlation between the objects are limited. The existing techniques fail to annotate overlapping and occluded objects in images accurately. The conventional YOLO fails to provide the connectivity and spatial relationship between different channels. To tackle the above problems, the proposed system offers the following advancements in YOLOv8:

- Development of remote sensing image annotation using novel YOLOv8 framework to detect and annotate the remote sensing images (RSIs).
- Incorporation of the Deformable Convolution module (DCN) into the Concatenate-Conv-Concatenate-Fusion (C2F) module that forms the DCN_C2F module. The DCN_C2F helps to acquire finer object details and capture spatial context by adaptively adjusting the receptive fields.
- Integration of the Spatial Channel-wise Convolution module (SCConv) with the C2F module to acquire the channel-wise semantic information and correlation to form the SCConv_C2F module.
- Integration of a coordinate attention (CA) mechanism to improve the connectivity between position and inter-channel relationship.

The remaining paper is structured as follows: Section 2 offers the literature review of recent work on remote sensing image annotation. Section 3 provides the details about the proposed methodology. Section 4 gives the discussions on the experimental results and analytical findings. Finally, Section 5 presents the conclusions and future scope of the work.

2. RELATED WORK

Image annotation is a crucial process that underpins numerous computer vision applications, from medical imaging and remote sensing to smart agriculture and mobile visual search. The reviewed studies leverage deep learning and hybrid methods to automate or accelerate this process, aiming to reduce manual effort and improve annotation precision. Despite these advancements, several limitations and challenges persist that necessitate further exploration.

Fernandes et al. [11] presented an innovative annotation approach for video capsule endoscopy (VCE) images, which is particularly beneficial while manual labeling is time-consuming and prone to error. They employed a Siamese network with ResNet-18 for image similarity matching, achieving an accuracy of over 97%. Their method streamlines medical annotation by reducing reliance on manual tools and enabling efficient CBIR in clinical workflows. Gautam and Khanna [12] addressed the limitations of text-based image retrieval by utilizing VGG16 and ResNet-50 to extract deep features for CBIR. Their system enhances precision, narrows the semantic gap, and provides an intuitive interface suitable for applications such as medical retrieval and visual search on mobile devices. Palekar [13] introduced a hybrid deep learning model that combines CNNs with biologically inspired feature selection for automatic image annotation. Using techniques such as LBP, Slantlet transform, and a red colobus monkey-inspired algorithm, the system achieved up to 99.2% accuracy on standard datasets, providing a creative and effective annotation pipeline. Meenakshi et al. [14] focused on lightweight CBIR for mobile devices using AdaBoost. By refining weak classifiers and emphasizing misclassified images, their model achieved a precision of 95.6% while remaining computationally efficient, making it ideal for real-time mobile image search. Zhang et al. [15] developed an interactive medical image annotation framework using Attention U-Net and geodesic distance guidance. Their two-stage segmentation process reduces manual effort while maintaining high accuracy on prostate MRI data, supporting faster and more efficient dataset labeling.

Liu et al. [16] proposed AIO², a system that corrects noisy labels in remote sensing segmentation using adaptive feedback and a mean-teacher model. Applied to building datasets, it improved annotation quality without requiring manual re-labeling, making it ideal for large-scale Earth observation. Song et al. [17] introduced a robust training method for noisy medical labels called Confidence Regularized Co-Teaching. The model filters uncertain labels by focusing on confident disagreements, yielding strong results even with imperfect annotations, which is key for scalable clinical AI. Mamat et al. [18] developed a YOLO-based annotation tool for agricultural images, which classifies oil palm ripeness with an accuracy of over 98%. The system supports smart farming by enabling fast and reliable fruit classification using transfer learning on RGB images. He et al. [19] developed PhenoLearn, a toolkit designed for biologists to annotate images of natural specimens. With a GUI and a deep learning backend, it enables non-experts to accurately label and segment traits, bridging the gap between biology and AI tools, as tested on bird plumage. Beck et al. [20] investigated the combination of ChatGPT-4V with human annotators for labeling satellite images. Their hybrid system reduced costs by 50% with minimal loss of accuracy, demonstrating how large language models can support scalable and cost-effective annotation workflows.

Lotfi et al. [21] proposed a graph-based method for automatic image annotation, modeling relationships between tags using GNNs. Tested on Corel5k and ESP Game datasets, the method improved annotation precision by capturing tag dependencies, enhancing multi-label performance. Wang et al. [22] developed AIDE, a framework that achieves near-supervised segmentation quality using only 10% labeled data. Applied to breast tumor datasets, AIDE reduces reliance on costly manual annotations while maintaining high diagnostic

accuracy. Neptune and Mothe [23] used scientific literature to enrich satellite deforestation image annotations. By combining deep change detection with semantic embedding, their system generates context-aware labels, enhancing the interpretability of environmental monitoring tools.

Huang et al. [24] developed a deep learning framework for multi-label remote sensing annotation that utilizes attention and feature fusion. The model performs well on complex scenes by capturing label correlations and detecting small objects, boosting annotation accuracy in RS imagery. Hua et al. [25] introduced FESTA, a weakly supervised segmentation method that works with sparse scribble annotations. It combines spatial and feature regularization to achieve strong performance with minimal input, making it a cost-effective solution for labeling aerial images. To address the noise, Zhang et al. [26] introduced Deep Multi-Similarity Hashing (DMSH) for RSI retrieval by fusing hashing and spatial information. For the UCM dataset, it yielded a mean average accuracy of 0.97. For cross-modal retrieval, Cheng and Zhang [27] presented a Deep Semantic Alignment Network (DSAN), which was evaluated on many captioned datasets, such as NWPU-RESISC45-Captions and UCMerced-LandUse-Captions. Through semantic alignment, it improves cross-modal retrieval and attains above 90% classification accuracy on the UCMerced-LandUse dataset. Its efficacy may be limited by its poorer accuracy in visually comparable categories, despite its strong performance. For RSI retrieval, Maurya et al. [28] proposed an adaptive DL-based model to address the drawbacks of low retrieval, rigidity, and inefficiency. They examined how well the various transfer learning models performed in image retrieval on the UCM dataset. For the 21-class classification, the accuracy of 95.07% for VGG19 was higher than 93% for VGG16 and 91% for ResNet.

Various YOLO based schemes has been presented for object detection and annotation in RSIs. Blushtein-Livnon et al. [29] examined how annotation tactics, dataset imbalances, and annotators' experience impact the accuracy of remote sensing labelling in their lab experiment. Professionals and non-experts labeled aerial images in ArcGIS Pro to identify and segment small PV panels. It is found that people are better at detection than segmentation. Annotators were more likely to make Type II mistakes (missing items) than Type-I errors (false detections) under different situations, suggesting a cautious approach that favored underestimating. Extreme imbalance negatively affected accuracy, whereas datasets with balanced target-to-background ratios performed more effectively. Interestingly, experience had no influence and occasionally led to overconfidence in segmentation. These results suggest that remote sensing requires improved annotation methodologies to meet the growing demand for high-quality training data.

Wang et al. [30] introduced AG-YOLO, a remote sensing-optimized attention-guided object identification system. Adding a rotation parameter to YOLOv10 and expanding the dual label assignment mechanism improves oriented object recognition. A separate attention branch suppresses complicated background noise and concentrates on key foreground details. Additionally, a three-stage curriculum learning technique enables the model to learn from easier examples before processing more complex data. Zhao et al. [31] introduced SCENE-YOLO. This novel detection framework integrates scene supervision into the YOLOv8 architecture, resulting in reduced processing latency and improved detection accuracy compared to other state-of-the-art approaches on the DOTA dataset. Their technique enhances detection by modeling object instance connections and utilizing background information that other approaches overlook. A transformer-based scene information network adds semantic context to feature extraction, and an omnidimensional dynamic convolution method modifies feature weighting. In complex scenarios, a prototype-based scene label generation method and a scene-assisted detection head improve classification accuracy. DOTA and DIOR experiments showed the model outperformed standard detection networks.

Hu et al. [32] employed CM-YOLO, a cloud- and mist-affected remote sensing image model, to enhance detection in challenging weather conditions. A component-decoupling background suppression module improves target-background contrast, while a local-global semantic mining module combines convolutional networks and selective attention to extract rich contextual data. CM-YOLO exceeded multiple detectors in terms of accuracy, recall, and mAP. Qian et al. [33] proposed IPS-YOLO, a pseudo-fully supervised training system, to address issues in weakly supervised object detection (WSOD). Their technique solves partial object identification and static pseudo ground truth concerns in prior methods. They employed category confidence-guided filtering and weighted synthesis to create a more comprehensive PGT and utilized iterative refinement during training. Proposal creation, increased sample selection, and pseudo-label definition further boosted performance. Comparative investigations demonstrated that IPS-YOLO outperformed WSOD models on numerous benchmark datasets.

Transfer learning for remote sensing object identification was examined by Pandilova et al. [34] using YOLOv8. They compared models trained from scratch with COCO-trained and DIOR- and Ships-tuned models. Transfer learning enhanced detection accuracy, precision, and recall, especially for tiny and similar items. Confusion matrix analysis revealed higher category discrimination, indicating that fine-tuning pre-trained models is beneficial for remote sensing. Finally, Jin et al. [35] introduced MTGS-YOLO, a task-balanced object identification method for remote sensing photography with complicated backdrops and high-resolution data. Their concept utilizes a multi-transformer architecture to capture both global and local information, and employs a GELAN for enhanced flexibility and computational performance. Small item recognition is improved by removing irrelevant background characteristics using a Spatial Context-Aware Module (SCAM). In the DIOR and NWPU dataset experiments, MTGS-YOLO outperformed previous algorithms in terms of detection accuracy, robustness, and generalization.

From the survey of recent techniques of image annotations, the following research gaps are identified:

- While current methods improve annotation accuracy and efficiency, many rely heavily on pre-trained networks without sufficient domain adaptation, which limits their generalization across tasks.
- Few approaches address annotation in low-resource environments or explore how to reduce model complexity for edge deployment.
- Though some works focus on noisy labels, robust handling of label noise under limited supervision remains underexplored.
- Most solutions are dataset-specific, with inadequate

validation on diverse or real-world datasets.

- There's also limited integration of multimodal cues (e.g., text, metadata) for enhanced annotation quality.
- Scalability and usability for non-experts, particularly in fields such as biology or agriculture, are often overlooked.
- Additionally, semi-supervised and active learning-based annotation pipelines are still evolving and lack standardization.
- Interactive systems exist, but they often fail to optimize user feedback integration or reduce cognitive load.
- Ethical considerations, such as annotation bias and privacy in medical datasets, are rarely discussed.
- Finally, real-time, explainable annotation frameworks with human-AI collaboration capabilities remain an open challenge.

3. METHODOLOGY

The framework of the improved YOLOv8-based remote sensing image annotation is illustrated in Figure 1, comprising four components: input, backbone, neck, and head modules. The input model comprises basic data preprocessing steps, including image normalization and resizing to a fixed size. The backbone network provides hierarchical features that depict texture, shape, edges, and objects. This module offers the multiscale features of objects, encompassing both semantic and spatial information. The neck model acquires attributes from the backbone to detect objects at multiple scales, including small, medium, and large. The head module predicts the bounding boxes, class probabilities, and objectness scores of the detected objects.

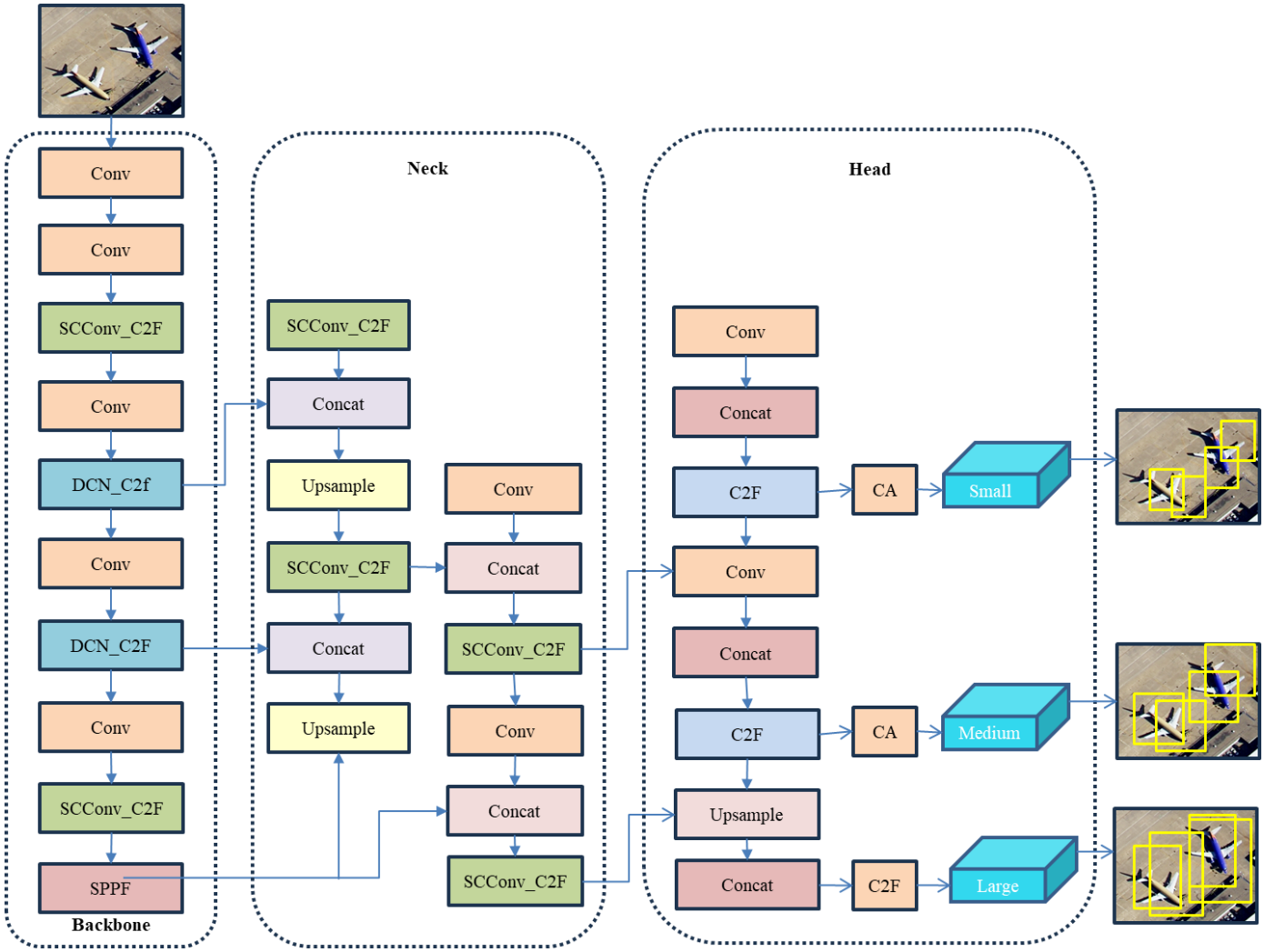

Figure 1. Architecture of improved YOLOv8-based image annotation

The improved YOLOv8 utilizes two SCConv_C2F and two DCN_C2F in the backbone layer to enhance the spatial correlation in the hierarchical texture, shape, and edge patterns of finer objects in RSIs. The four SCConv_C2F layers are utilized in the neck layer to improve the spatial connectivity between different channels, thereby acquiring hierarchical information to characterize medium, small, and large objects in the images. The head layer encompasses two CA layers to improve the contextual connectivity relationship between interchannel positions.

3.1 DCN_C2F

The generalized convolution is given by Eq. (1), where x indicate the input features, y depicts the output features, n signifies the sampling point of the neurons, $w(p_n)$ symbolizes the weights, $x(p_n)$ offers the pixel value of individual features and p_0 denotes the centralized sampling factor of the feature map. In the generalized convolution, the filtering calculations are performed over a fixed grid (R), where every sampling point is utilized for computing the weight using a convolution filter. In deformable convolution, the convolution operation

considers the offset (Δp_n) at position p_n computed using interpolation of sampling points. The deformable convolution operation is provided by Eq. (2).

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_n + p_0) \tag{1}$$

$$y(p_0) = \sum_{p_n \in R} w(p_n) . x(p_n + p_0 + \Delta p_n) \tag{2}$$

The stacking of multiple DCN_C2F layers increases the receptive fields for the image area, enabling the capture of salient gradient flow in objects. The DCN_C2F offers multilevel hierarchical features and acquires invariant attributes, increasing its adaptability to detect objects with varying geometric shapes.

3.2 SCConv_C2F

The SCConv_C2F addresses the issue of redundant features and lower spatial connectivity by utilizing the CNN compression method. The SCConv_C2F consists of the channel reconstruction unit (CRU) and spatial reconstruction unit (SRU). The SRU increases the spatial connectivity by decomposing feature maps into distinct spatial blocks and utilizing separate convolution filters for each block. The CRU helps enhance channel dimensions by utilizing lightweight, dense layers that combine records from similar channels and reduce redundancy. The structure of the SSConv is given in Figure 2 which consists of SRU and CRU as primary building blocks.

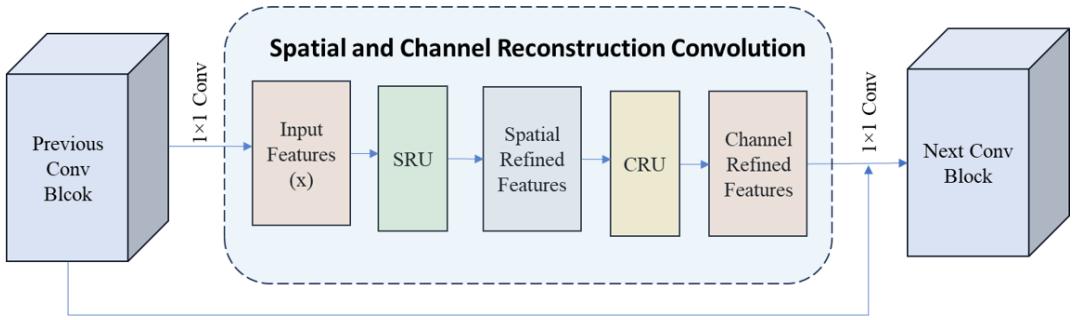

Figure 2. Structure of SCConv with CRU and SRU

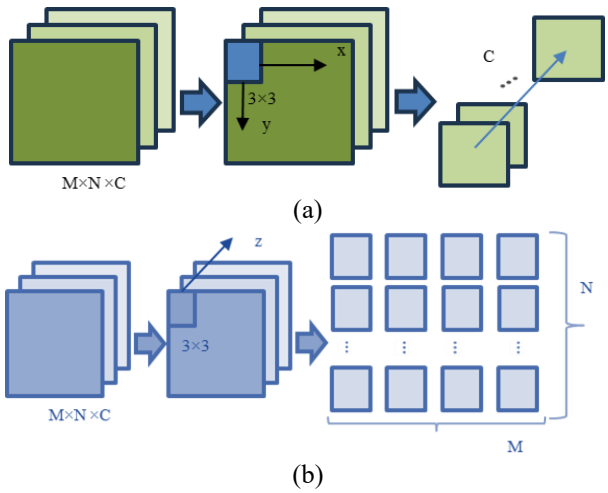

Figure 3. (a) Process of general convolution; (b) Process of spatial channel-wise convolution

In standard convolution, the kernel moves across the (x, y)-plane, and the produced feature maps are stacked along the z-axis, which corresponds to the channel dimension. In contrast, spatial channel-wise convolution shifts the kernel along the z-axis, generating feature maps that are positioned in the (x, y)-plane, following the same spatial layout as the input feature maps, as given in Figure 3(a) and Figure 3(b), respectively. The SCConv is provided as given in Eq. (3) where $O(i, j)$ denotes the SCConv output, C denotes total channels, M stands for rows, N denotes columns and K_c indicates the convolution filter for channel c.

$$O(i, j) = \sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{n=1}^{N} x(i+m, j+n) * K_c(m, n) \tag{3}$$

Furthermore, the CA mechanism helps to improve the inter-channel relationship and position correlation across different feature channels to facilitate object detection. This attention mechanism helps to detect the finer objects with occluded and overlapping conditions.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 System setup

The suggested system is implemented using Python programming on an NVIDIA GPU with 128 GB of RAM, 32 GB of graphics memory, and a Windows operating environment. The parameter specifications for YOLOv8 are listed in Table 1.

Table 1. Parameter specification of YOLOv8

Parameter	Specification
Learning algorithm	adam
Batch size	64
Learning rate	0.001
Loss function	cross entropy
Momentum	0.85
Epoch	200

4.2 Performance metrics

The effectiveness of the proposed system is evaluated based on recall, precision, F1-score, accuracy, and mean average precision (mAP) as given in Eqs. (4)-(8), respectively. Here, TP_V, TN_V, FP_V, FN_V, and AP depict the actual positive value, true negative value, false positive value, false negative value, and area under the recall-precision curve for N classes.

The recall and precision offer the quantitative and qualitative measures of the retrieval results. The accuracy represents the overall accuracy; the F1-score offers a quality-quantity balance; and the mAP provides a rank-based balance between precision and recall.

$$Recall = \frac{TP_V}{TP_V + FN_V} \tag{4}$$

$$Precision = \frac{TP_V}{TP_V + FP_V} \tag{5}$$

$$Accuracy = \frac{TP_V + TN_V}{TP_V + FN_V + FP_V + TN_V} \tag{6}$$

$$F1-score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{7}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{8}$$

4.3 Dataset

To assess the performance of the proposed CBIR approach, experiments were conducted using the UC Merced Land Use Dataset [1]. This dataset comprises 21 distinct categories of land use, each containing 100 images with a resolution of 256 × 256 pixels. The model uses 70% of the data for training and 30% for testing purposes. The images were sourced from the USGS National Map's Urban Area Imagery, capturing various urban regions across the United States. A few example images from the dataset are shown in Figure 4.

Figure 4. UCM sample images

4.4 Discussions on results

The visualizations of the object annotation and detection are shown in Figures 5-7, respectively, for YOLOv8, YOLOv8+DCN-C2F, and YOLOv8+DCN_C2F+SCConv-C2F. The YOLOv8 detects the salient object but misses finer objects with occluded positions, as shown in Figure 5. The DCN and SCConv help detect fine and multiscale objects in images by providing spatial connectivity in the inter-channel positions of the objects, as illustrated in Figure 6 and Figure 7, respectively.

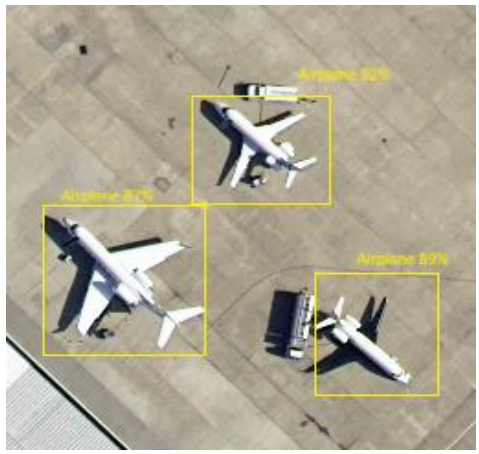

Figure 5. Object detection for sample image using YOLOv8

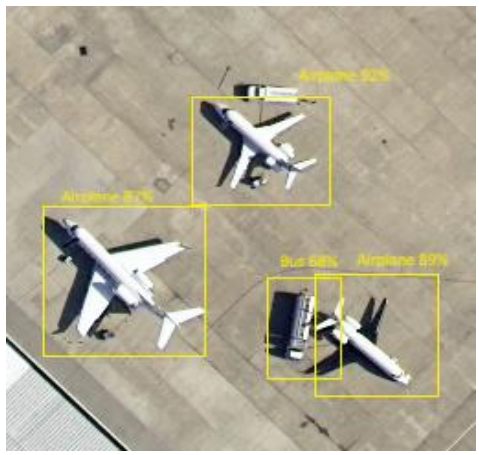

Figure 6. Object detection for sample image using YOLOv8+DCN_C2F

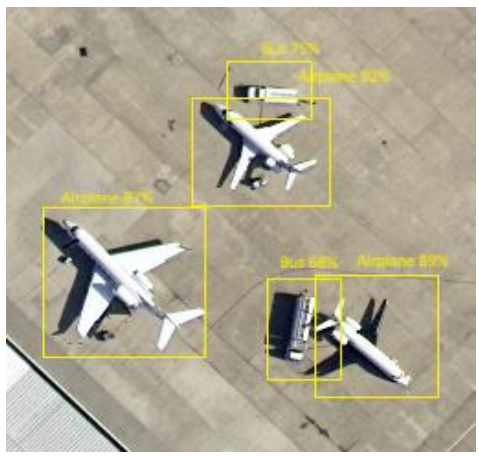

Figure 7. Object detection for sample image using YOLOv8+DCN_C2F+SCConv_C2F

Table 2. Comparative analysis of the proposed method

Algorithm	Recall	Precision	F1-Score	Accuracy
YOLOv3	77.50	85.80	81.41	80.10
YOLOv5	86.20	87.50	86.84	87.00
YOLOv8	89.20	94.20	91.62	91.80
YOLOv8 +DCN_C2F	94.40	94.30	94.35	94.20
YOLOv8+SCConv_C2F	97.50	97.10	97.30	97.10
YOLOv8+DCN_C2F +SCConv_C2F	98.50	98.40	98.45	98.15

The effectiveness of image object annotation is evaluated across various YOLO architectures, including YOLOv3, YOLOv5, and YOLOv8, as shown in the Table 2. The traditional YOLOv3 achieves a recall rate of 77.50%, a precision of 85.80%, an F1-score of 81.41%, and an accuracy of 80.10% on the UCM dataset. The YOLOv5 provides improved accuracy compared to YOLOv3, resulting in a recall of 86.20%, a precision of 87.50%, an F1-score of 86.84%, and an accuracy of 87%. The YOLOv8 achieved a recall rate of 89.20%, a precision of 94.20%, an F1-score of 91.62%, and an accuracy of 91.80%, which demonstrates a notable improvement over YOLOv3 and YOLOv5, owing to its ability to learn the finer details of image objects. The DCN-C2F allows the model to adaptively adjust its receptive fields, enabling it to capture finer object details and better understand spatial context. The YOLOv8+DCN_C2F achieves a recall rate of 94.40%, precision of 94.30%, F1-score of 94.35%, and accuracy of 94.20%. The SCConv-C2F enhances the network's ability to extract and utilize semantic relationships across different channels. The YOLOv8+SCConv_C2F achieves an overall recall of 97.50%, precision of 97.10%, F1-score of 97.30%, and accuracy of 97.10%. The integration of DCN and SCConv in YOLOv8 results in improved recall of 98.50%, precision of 98.40%, F1-score of 98.45% and accuracy of 98.15% which shows a crucial boost for remote sensing object detection and annotation over traditional models. The visualizations of the comparative results for the various models are shown in Figures 8-11.

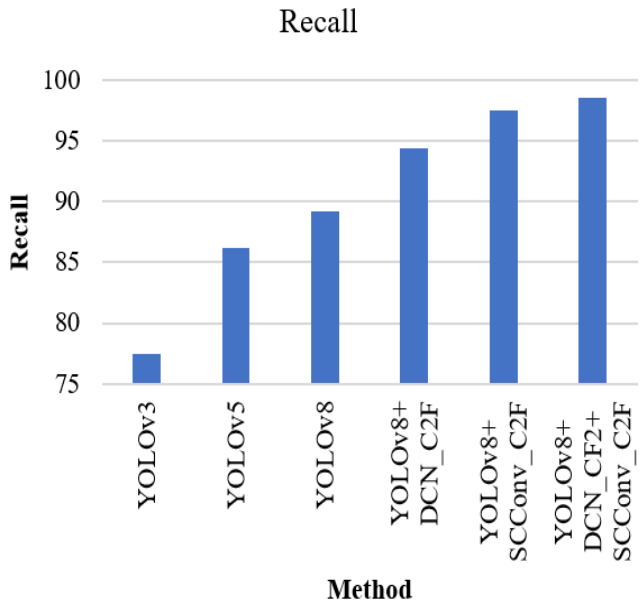

Figure 8. Recall comparison of different methods

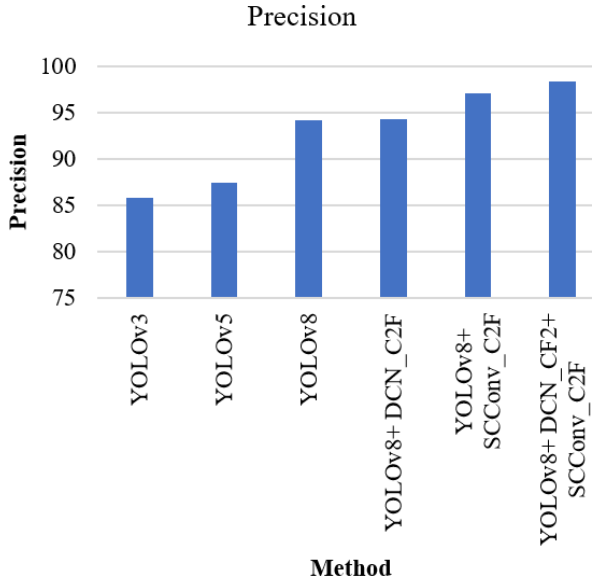

Figure 9. Precision comparison of different methods

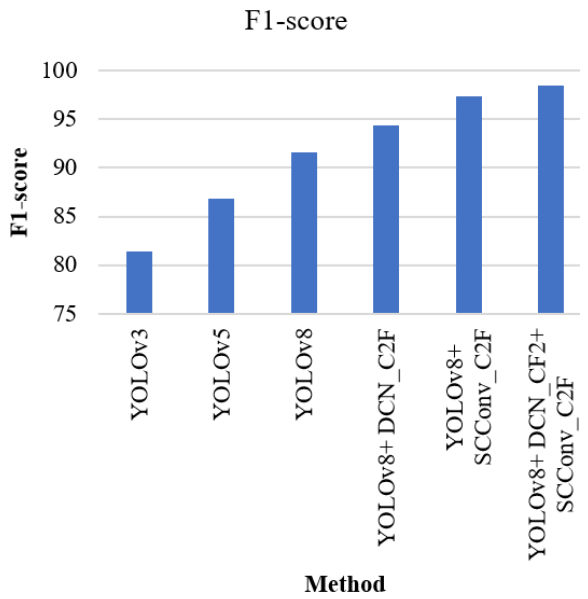

Figure 10. F1-score comparison of different methods

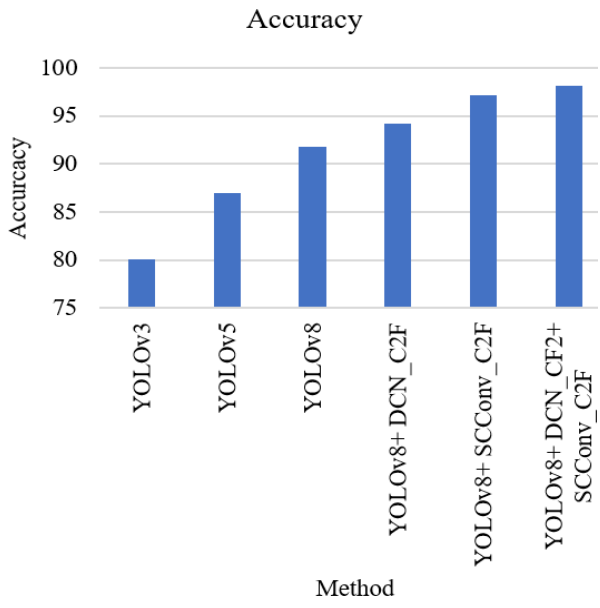

Figure 11. Accuracy comparison of different methods

Table 3. Performance comparison of the proposed YOLOv8 for different learning algorithms

Learning Algorithm	Method	Accuracy	MAP
Adam	YOLOv8	91.80	0.92
	YOLOv8+ DCN_C2F	94.20	0.94
	YOLOv8+ SCConv_C2F	97.10	0.97
	YOLOv8+ DCN_C2F+ SCConv_C2F	98.15	0.99
SGDM	YOLOv8	88.80	0.89
	YOLOv8+ DCN_C2F	91.00	0.91
	YOLOv8+ SCConv_C2F	93.50	0.94
	YOLOv8+ DCN_C2F+ SCConv_C2F	95.00	0.96
RMSProp	YOLOv8	86.00	0.86
	YOLOv8+ DCN_C2F	88.50	0.89
	YOLOv8+ SCConv_C2F	91.00	0.91
	YOLOv8+ DCN_C2F+ SCConv_C2F	92.50	0.93

Table 3 presents the performance comparison of the proposed YOLOv8 framework under different learning algorithms for remote sensing image object detection and annotation. With the Adam optimizer, the baseline YOLOv8 achieved an accuracy of 91.80% with a mAP of 0.92, which was further enhanced to 94.20% (mAP 0.94) when deformable convolution layers (DCN_C2F) were integrated. The inclusion of spatial channel-wise convolution (SCConv_C2F) produced a significant boost, reaching 97.10% accuracy and 0.97 mAP, while the combined DCN_C2F + SCConv_C2F configuration delivered the best results at 98.15% accuracy and 0.99 mAP. In contrast, the SGDM optimizer yielded slightly reduced performance, with YOLOv8 scoring 88.80% accuracy and 0.89 mAP, improving to 95.00% accuracy and 0.96 mAP for the hybrid configuration. The RMSProp optimizer showed the lowest overall results, with the baseline YOLOv8 at 86.00% accuracy and 0.86 mAP, while its best performance under DCN_C2F + SCConv_C2F reached 92.50% accuracy and 0.93 mAP. These outcomes indicate that while architectural enhancements such as deformable and spatial convolutions substantially improve detection accuracy across all optimizers, Adam consistently outperforms SGDM and RMSProp by margins of 3–5% in accuracy and up to 0.03 in mAP, demonstrating its robustness for precise object detection and annotation in complex remote sensing imagery.

The effectiveness of the system is compared with the traditional state-of-the-art using the UCM dataset for object retrieval and annotation, as shown in Table 4, based on average accuracy and mAP. The proposed method shows a remarkable improvement over existing techniques in terms of both accuracy and mAP. Compared to DSAN, which achieved 90% accuracy, and Maurya et al.'s use of traditional CNN architectures, such as VGG19 (95.07%), the advanced integration of SCConv and DCN modules into the YOLOv8 architecture results in a significant boost. Specifically, the final configuration—YOLOv8 combined with DCN_C2F and SCConv_C2F achieves 98.15% accuracy and 0.99 mAP. This marks an 8.15% improvement in accuracy over DSAN and a 3.08% increase over the best-performing CNN baseline (VGG19)—moreover, the mAP increases by 2% over the YOLOv8 baseline and by 2.06% over DMSH. The novel YOLOv8 offered an improved mAP compared to existing YOLO-based frameworks, such as AG-YOLO, CM-YOLO, and YOLO-COCO, which provided mAPs of 0.80, 0.85, and 0.76 for remote sensing images. These enhancements demonstrate the effectiveness of combining deformable convolution and selective context convolution modules, enabling the network to capture spatial context and object deformations more effectively, resulting in more accurate detection and annotation.

Table 4. Comparison of proposed RSI retrieval and annotation with traditional techniques

Authors	Method	Accuracy	MAP
Zhang et al. [26]	DMSH	-	0.97
Cheng and Zhang [27]	DSAN	90 %	-
Maurya et al. [28]	VGG19	95.07%	-
	VGG16	93%	
	ResNet	91%	
Wang et al. [30]	AG-YOLO	-	0.80
Hu et al. [32]	CM-YOLO	89.4%	0.85
Pandilova et al. [34]	YOLO-COCO	-	0.76
Proposed Method	YOLOv8	91.80	0.92
	YOLOv8+ DCN_C2F	94.20	0.94
	YOLOv8+ SCConv_C2F	97.10	0.97
	YOLOv8+ DCN_C2F+ SCConv_C2F	98.15	0.99

5. CONCLUSION AND FUTURE SCOPE

This paper presents a DL based remote sensing image annotation that shows significant improvement in the multiscale image object detection. The improved YOLOv8 incorporates the DCN and SCConv modules to capture the contextual and semantic information of objects during object detection and annotation. By integrating DCN and SCConv into YOLOv8, the model achieves remarkable results, with 98.50% recall, 98.40% precision, 98.45% F1-score, and 98.15% accuracy. The improved YOLOv8 helps detect finer, more complex, occluded, and overlapping objects in intricate images by providing enhanced spatial connectivity and correlation across multiple channels. This notable performance uplift highlights the model's strong potential to outperform traditional methods in remote sensing-based object detection and annotation tasks.

However, the model's performance is limited for multispectral and hyperspectral remote sensing images due to its complexity. It lacks in proving “Interpretability and Explainability” because of the black-box nature of YOLOv8. In the future, the texture and shape attributes can be annotated along with the class label. The annotation labels can be embedded in the images, which can be utilized during real-time CBRSIR. In the future, the performance of YOLOv8 can be improved by optimizing the model's hyperparameters and providing the “Interpretability and Explainability” so as to increase trust in the system.

DATA AVAILABILITY STATEMENT

The open access UC Merced Land Use Database can be found at http://weegee.vision.ucmerced.edu/datasets/landuse.html.

REFERENCES

[1] Dewi, C., Bilaut, F.Y., Christanto, H.J., Dai, G. (2024). Deep learning for the classification of rice leaf diseases using YOLOv8. Mathematical Modelling of Engineering Problems, 11(11): 3025-3034. https://doi.org/10.18280/mmep.111115

[2] Sudjud, S., Jamil, M., Melati, R. (2025). Development of cocoa pod rot disease identification model using You Only Look Once (YOLO-v9). Mathematical Modelling of Engineering Problems, 12(4): 1268-1274. https://doi.org/10.18280/mmep.120418

[3] Chembian, W.T., Senthilkumar, G., Prasanth, A., Subash, R. (2025). K-means pelican optimization algorithm-based search space reduction for remote sensing image retrieval. Journal of the Indian Society of Remote Sensing, 53(1): 101-115. https://doi.org/10.1007/s12524-024-01994-z

[4] Soni, P., Singh, M., Sharma, P., Kumar, T., Cheng, X., Kumar, R., Paliwal, M. (2025). Multifeature fusion for enhanced content-based image retrieval across diverse data types. Journal of Electrical and Computer Engineering, 2025(1): 3889925. https://doi.org/10.1155/jece/3889925

[5] Sornalatha, P., Mahesh, K., Panneerselvam, K. (2025). Multi-class remote sensing image retrieval using optimized convolution neural network with weighted distances. Journal of the Indian Society of Remote Sensing, 53(5): 1343-1358. https://doi.org/10.1007/s12524-024-02036-4

[6] Adnan, M.M., Rahim, M.S.M., Rehman, A., Mehmood, Z., Saba, T., Naqvi, R.A. (2021). Automatic image annotation based on deep learning models: A systematic review and future challenges. IEEE Access, 9: 50253-50264. https://doi.org/10.1109/ACCESS.2021.3068897

[7] Aljabri, M., AlAmir, M., AlGhamdi, M., Abdel-Mottaleb, M., Collado-Mesa, F. (2022). Towards a better understanding of annotation tools for medical imaging: A survey. Multimedia Tools and Applications, 81(18): 25877-25911. https://doi.org/10.1007/s11042-022-12100-1

[8] Bello, R.W., Owolawi, P.A., van Wyk, E.A., Tu, C. (2025). Image annotation tools and dataset: A comparative analysis in brief. In International Conference on AI-Generated Content (AIGC 2024). https://doi.org/10.1117/12.3065175

[9] Fernandes, R., Pessoa, A., Salgado, M., Paiva, A.D., Pacal, I., Cunha, A. (2024). Enhancing image annotation with object tracking and image retrieval: A systematic review. IEEE Access, 12: 79428-79444. https://doi.org/10.1109/ACCESS.2024.3406018

[10] Horvath, A.S., Pouliou, P. (2024). AI for conceptual architecture: Reflections on designing with text-to-text, text-to-image, and image-to-image generators. Frontiers of Architectural Research, 13(3): 593-612. https://doi.org/10.1016/j.foar.2024.02.006

[11] Fernandes, I., Fernandes, R., Pessoa, A., Salgado, M., Paiva, A., Paçal, I., Cunha, A. (2025). A deep learning approach to annotating endoscopic capsule videos via CBIR. Procedia Computer Science, 256: 1108-1115. https://doi.org/10.1016/j.procs.2025.02.218

[12] Gautam, G., Khanna, A. (2024). Content based image retrieval system using CNN based deep learning models. Procedia Computer Science, 235: 3131-3141. https://doi.org/10.1016/j.procs.2024.04.296

[13] Palekar, V. (2024). Adaptive optimized residual convolutional image annotation model with bionic feature selection model. Computer Standards Interfaces, 87: 103780. https://doi.org/10.1016/j.csi.2023.103780

[14] Meenakshi, K., Singh, K.D., Batta, P., Chauhan, R., Bhargava, D., Lohani, B.P. (2024). Enhanced ML based content-based image retrieval for mobile devices. In 2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit), Greater Noida, India, pp. 213-218. https://doi.org/10.1109/GlobalAISummit62156.2024.10947892

[15] Zhang, Y., Chen, J., Ma, X., Wang, G., Bhatti, U.A., Huang, M. (2024). Interactive medical image annotation using improved Attention U-net with compound geodesic distance. Expert Systems with Applications, 237: 121282. https://doi.org/10.1016/j.eswa.2023.121282

[16] Liu, C., Albrecht, C.M., Wang, Y., Li, Q., Zhu, X.X. (2024). AIO2: Online correction of object labels for deep learning with incomplete annotation in remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-17. https://doi.org/10.1109/TGRS.2024.3373908

[17] Song, Y., Liu, Y., Lin, Z., Zhou, J., Li, D., Zhou, T., Leung, M.F. (2024). Learning from AI-generated annotations for medical image segmentation. IEEE Transactions on Consumer Electronics. https://doi.org/10.1109/TCE.2024.3474037

[18] Mamat, N., Othman, M.F., Abdulghafor, R., Alwan, A.A., Gulzar, Y. (2023). Enhancing image annotation technique of fruit classification using a deep learning approach. Sustainability, 15(2): 901. https://doi.org/10.3390/su15020901

[19] He, Y., Cooney, C.R., Maddock, S., Thomas, G.H. (2025). PhenoLearn: A user-friendly toolkit for image annotation and deep learning-based phenotyping for biological datasets. Journal of Evolutionary Biology, 38(8): 1152-1162. https://doi.org/10.1093/jeb/voaf058

[20] Beck, J., Kemeter, L.M., Duerrbeck, K., Abdalla, M.H.I., Kreuter, F. (2025). Towards integrating ChatGPT into satellite image annotation workflows. A comparison of label quality and costs of human and automated annotators. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 4366-4381. https://doi.org/10.1109/JSTARS.2025.3528192

[21] Lotfi, F., Jamzad, M., Beigy, H. (2021). Automatic image annotation using quantization reweighting function and graph neural networks. In International Conference on Service-Oriented Computing, pp. 46-60. https://doi.org/10.1007/978-3-031-14135-5_4

[22] Wang, S., Li, C., Wang, R., Liu, Z., et al. (2021). Annotation-efficient deep learning for automatic medical image segmentation. Nature Communications, 12(1): 5915. https://doi.org/10.1038/s41467-021-26216-9

[23] Neptune, N., Mothe, J. (2024). Enriching satellite image

annotations of forests with keyphrases from a specialized corpus. Multimedia Tools and Applications, 84: 37637-37653. https://doi.org/10.1007/s11042-024-20015-2

[24] Huang, R., Zheng, F., Huang, W. (2021). Multilabel remote sensing image annotation with multiscale attention and label correlation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 6951-6961. https://doi.org/10.1109/JSTARS.2021.3091134

[25] Hua, Y., Marcos, D., Mou, L., Zhu, X.X., Tuia, D. (2021). Semantic segmentation of remote sensing images with sparse annotations. IEEE Geoscience and Remote Sensing Letters, 19: 1-5. https://doi.org/10.1109/LGRS.2021.3051053

[26] Zhang, H., Qin, Q., Ge, M., Huang, J. (2024). Deep multi-similarity hashing with spatial-enhanced learning for remote sensing image retrieval. Electronics, 13(22): 4520. https://doi.org/10.3390/electronics13224520

[27] Cheng, L., Zhang, Z. (2022). Deep semantic alignment network for cross-modal remote sensing image retrieval. Journal of Remote Sensing, 15(3): 189-205. https://doi.org/10.1109/JSTARS.2021.3070872

[28] Maurya, A., Akashdeep, Kumar, R. (2024). Classification of University of California (UC): Merced land-use dataset remote sensing images using pre-trained deep learning models. In Deep Learning Techniques for Automation and Industrial Applications, pp. 45-67. https://doi.org/10.1002/9781394234271.ch4

[29] Blushtein-Livnon, R., Svoray, T., Dorman, M. (2025). Performance of human annotators in object detection and segmentation of remotely sensed data. IEEE Transactions on Geoscience and Remote Sensing. https://doi.org/10.13140/RG.2.2.34395.17446

[30] Wang, X., Han, C., Huang, L., Nie, T., Liu, X., Liu, H., Li, M. (2025). AG-Yolo: Attention-guided Yolo for efficient remote sensing-oriented object detection. Remote Sensing, 17(6): 1027. https://doi.org/10.3390/rs17061027

[31] Zhao, T., Feng, R., Wang, L. (2025). SCENE-YOLO: A one-stage remote sensing object detection network with scene supervision. IEEE Transactions on Geoscience and Remote Sensing, 63: 1-15. https://doi.org/10.1109/TGRS.2025.3539698

[32] Hu, J., Wei, Y., Chen, W., Zhi, X., Zhang, W. (2025). CM-YOLO: Typical object detection method in remote sensing cloud and mist scene images. Remote Sensing, 17(1): 125. https://doi.org/10.3390/rs17010125

[33] Qian, X., Zhang, B., He, Z., Wang, W., Yao, X., Cheng, G. (2025). IPS-YOLO: Iterative pseudo-fully supervised training of YOLO for weakly supervised object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 63: 1-14. https://doi.org/10.1109/TGRS.2025.3586239

[34] Pandilova, E., Petrov, M., Spasev, V., Dimitrovski, I., Kitanovski, I. (2024). Transfer learning with Yolo for object detection in remote sensing. In International Conference on ICT Innovations, pp. 121-135. https://doi.org/10.1007/978-3-031-86162-8_9

[35] Jin, Z., Duan, J., Qiao, L., He, T., Shi, X., Yan, B. (2025). MTGS-Yolo: A task-balanced algorithm for object detection in remote sensing images based on improved Yolo. The Journal of Supercomputing, 81(4): 542. https://doi.org/10.1007/s11227-025-07003-5