Turkish Journal of Science & Technology
21(2), 437-447, 2026

Research Article
https://doi.org/10.55525/tjst.1941049

Foreground-Aware Region of Interest Cropping and CBAM Attention for Excavator Segmentation in UAV Imagery

Salih Enes ŞEKER[1,2], Ali DEĞİRMENCİ[1*], İlyas ÇANKAYA[1]
[1] Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Sciences, Ankara Yıldırım Beyazıt University, Ankara, Türkiye
[2] HAVELSAN, Ankara, Türkiye
[1] salihenesseker23@aybu.edu.tr, *[2] alidegirmenci@aybu.edu.tr, [3] ilyas.cankaya@aybu.edu.tr

(Geliş/Received: 07/05/2026; Kabul/Accepted: 24/09/2026)

Abstract: In disaster situations like earthquakes, excavators have a crucial part to play in the rapid clearing of rubble, which is critical for conducting search and rescue efforts. Effective tracking and localization of excavators can be achieved through accurate monitoring of their presence and distribution in disaster-hit regions, enabling better coordination and resource management. This study presents a deep learning-based approach for the segmentation of excavators in high-resolution Unmanned Aerial Vehicle (UAV) imagery, addressing challenges related to small object representation and complex backgrounds. Specifically, our proposed approach involves using a foreground-aware region of interest cropping mechanism to ensure that there are sufficient excavator pixels in the training dataset, and a convolutional block attention module to improve feature refinement by taking into account channel-wise and spatial relationships between features. The proposed network utilizes a U-Net architecture based on ResNet-34. Experimental evaluations on UAV imagery reveal that the proposed approach achieved an IoU of 0.663, mIoU of 0.830, Dice score of 0.798, Precision of 0.856, Recall of 0.747 and geometric mean (G-Mean) of 0.863. These results demonstrate a notable improvement over the baseline model (U-Net with a ResNet-34 backbone), with the IoU increasing from 0.509 to 0.663 and the Dice score improving from 0.675 to 0.798.

Keywords: Excavator, deep learning, segmentation, UAV, ResNet.

İHA Görüntülerinde Ekskavatör Segmentasyonu için Ön Plan Farkındalıklı İlgi Bölgesi Kırpma ve CBAM Dikkat Mekanizması

Öz: Deprem gibi afet durumlarında, ekskavatörler enkazın hızlı bir şekilde temizlenmesinde hayati bir rol oynar; bu da arama ve kurtarma çalışmalarının yürütülmesi açısından kritik öneme sahiptir. Ekskavatörlerin varlığı ve afet bölgelerindeki dağılımının doğru bir şekilde izlenmesi sayesinde, ekskavatörlerin etkili bir şekilde takip edilmesi ve konumlarının belirlenmesi sağlanabilir; bu da daha iyi koordinasyon ve kaynak yönetimi imkânı sunar. Bu makale, yüksek çözünürlüklü İnsansız Hava Aracı (İHA) görüntülerinde ekskavatörlerin segmentasyonu için derin öğrenme tabanlı bir yaklaşım sunmakta ve küçük nesnelerin temsil edilmesi ve karmaşık arka planlarla ilgili zorlukları ele almaktadır. Özellikle, önerdiğimiz yaklaşım, eğitim veri setinde yeterli sayıda ekskavatör pikseli olmasını sağlamak için ön plana duyarlı ilgi alanı kırpma mekanizmasının ve özellikler arasındaki kanal bazlı ve uzamsal ilişkileri dikkate alarak özellik iyileştirmesini geliştirmek için bir evrişimli blok dikkat modülünün kullanılmasını içerir. Önerilen ağ, ResNet-34'e dayalı bir U-Net mimarisini kullanır. İHA görüntüleri üzerinde yapılan deneysel değerlendirmelerde önerilen yaklaşım, 0.663 IoU, 0.830 mIoU, 0.798 Dice katsayısı, 0.856 Kesinlik, 0.747 Duyarlılık, 0.863 G-Mean değerleri elde etmiştir. Bu sonuçlar, temel model olarak kullanılan ResNet-34 omurgalı U-Net mimarisine kıyasla belirgin bir performans artışı sağlamaktadır; IoU değeri 0.509'dan 0.663'e, Dice katsayısı ise 0.675'ten 0.798'e yükselmiştir. Elde edilen bu iyileştirmeler, önerilen yöntemin segmentasyon başarımını artırdığını ortaya koymaktadır.

Anahtar kelimeler: Ekskavatör, derin öğrenme, bölütleme, İHA, ResNet.

1. Introduction

Unmanned Aerial Vehicles (UAVs) have gained traction as a means of obtaining high-resolution images suitable for different tasks. Compared to satellite-based image acquisition, UAV platforms provide high-resolution images quickly. Unlike satellite-based methods, UAVs enable the acquisition of images with significantly higher spatial resolution. Thus, aerial imagery obtained using UAVs can be applied for infrastructure monitoring, construction site monitoring, environment observations, disasters monitoring, etc. Moreover, UAV imagery combined with computer vision methods and, especially, deep learning, enables the automated monitoring of specific objects over vast geographical areas. [1]. Though considerable progress has been achieved in developing vision systems based on deep learning methods, there still remain some issues related to object detection and segmentation in aerial images. In particular, objects captured by UAVs take only a small part of an image, which

* Corresponding author: alidegirmenci@aybu.edu.tr. ORCID Number of authors: [1] 0009-0000-5245-8564, [2] 0000-0001-9727-8559, [3] 0000-0002-6072-3097

leads to class imbalance between object (foreground) and background classes [2]. Moreover, aerial images usually contain significant variations in terms of the scale of objects, complex backgrounds, and lighting. Such problems may hinder segmentation models from learning discriminative representation of small objects. Improving training methods and data sampling methods appears to be a relevant problem nowadays. Among the targets monitored via UAV-based computer vision systems, heavy construction equipment represents a critical category. The automatic detection of construction equipment facilitates tracking various activities in construction sites, infrastructure areas, etc. [3]. Monitoring construction equipment is especially necessary in disaster response scenarios when it is important to track resource distribution. Heavy machinery is essential across various construction operations. Particularly, heavy construction machines such as excavators are vital for digging, lifting, and debris removal, especially in post-earthquake scenarios. Consequently, such machinery is vital for disaster response operations where time-sensitive tasks are critical. Existing excavator detection methods have demonstrated promising performance in general construction environments; however, they are not well suited for post-disaster remote sensing imagery. Most studies focus on object detection using bounding boxes rather than pixel-level segmentation, limiting their ability to accurately delineate excavator boundaries in complex scenes [4]. Small object segmentation remains a major challenge because excavators often occupy only a few pixels in high-resolution aerial images. Their limited visual information, together with repeated feature downsampling, makes them difficult to distinguish from the surrounding background, resulting in missed detections and inaccurate boundaries [5]. These challenges become more severe in disaster scenarios, where collapsed buildings, debris, smoke, shadows, and occlusions significantly increase scene complexity. Under these conditions, conventional detection and segmentation methods often fail to maintain sufficient robustness and accuracy, highlighting the need for models that can effectively preserve fine-grained features and exploit multi-scale contextual information [6]. In addition, many research studies have focused on excavator detection as well as object segmentation in UAV imaging. For instance, Meng et al. developed a real-time excavator detection system for pipeline safety inspection using a UAV [7]. In this work, the You Only Look Once (YOLO)v3 detection algorithm was optimized by adjusting the parameters of the model to anchor boxes using K-means clustering and adopting multi-scale training to overcome issues related to scale invariance in the aerial perspective view. The final implementation of the system is deployed in an edge computing pipeline to perform real-time inference. According to the experimental results, the detector exhibited high performance, with a recall of 99.4%, a precision of 97.7%, and an Intersection over Union (IoU) of 83.1%. Another study that utilized UAVs for excavator detection is the construction site monitoring framework introduced by Zhang and Zhang [8]. In this study, both one-stage and two-stage detectors are considered in the detection module, which include Region-based Convolutional Neural Networks (Faster R-CNN), YOLOv3, and YOLOv5. It is worth noting that the best performances in real-time detection were provided by YOLOv5x, having a mean average precision (mAP) of 95.0%, although excavator class achieves an average precision (AP) of 93.0% using Faster R-CNN. Hua et al. developed a computer vision-based excavator detector specifically tailored to nighttime construction applications [9]. In this study, the framework consists of an image enhancement module and an object detector module. More specifically, the former implements the Night-Enhancement Algorithm whereas the latter employs the YOLOv7 detector. Additionally, in order to boost performance in terms of object detection in low lighting conditions, several attention mechanisms were implemented such as Squeeze and Excitation, Convolutional Block Attention Module (CBAM) and Efficient Channel Attention. According to experimental results, the use of the image enhancement algorithm boosted the visibility in the images and increased detection accuracy by 8.85%. Furthermore, Kim et al. addressed excavator activity recognition through a vision-based approach combined with temporal behavior analysis using video sequence analysis [10]. To this end, a hybrid CNN-Bidirectional Long Short-Term Memory (BiLSTM) architecture was employed to train the model, which involves pre-trained CNN model that serves for spatial feature extraction, while BiLSTM extracts the temporal features by looking back in time to past frames and forward in time to future frames. Using four basic excavator activities (excavation, hauling, dumping, and return), experimental results show that the precision and recall of this approach are 87.5% and 88.52%, respectively, with 88% test accuracy. Fang et al. focused on the topic of excavator detection and workers detection on construction sites using a CNN-based approach [11]. To detect workers and excavators, an Improved Faster R-CNN (IFaster R-CNN) is introduced in which region proposal network uses five scales and four ratios. With respect to the performance metrics, IFaster R-CNN achieved a mAP score of 93%, with detection accuracy of 95% on test data. Specifically, excavator class exhibits AP of 95%, precision and recall scores of 99% and 81%, respectively. Moreover, Spasev et al. analyzed the application of the SegFormer architecture for segmentation of UAV remote sensing images in the context of the UAVid dataset [12]. In particular, SegFormer is based on a hierarchical transformer encoder model to extract multi-scale features from the images, and a lightweight decoder module that does not require any positional encodings in the form of multilayer perceptron networks. The experimental results

show that SegFormer-B5 achieved a mean Intersection over Union (mIoU) score of 69.55%, whereas an ensemble with test-time augmentation achieved 70.88% mIoU score. Finally, Lyu et al. presented the new UAVid dataset for segmenting UAV remote sensing imagery from oblique viewpoints, in which 300 densely labeled images from 30 sequences were created with 8 classes [13]. In order to evaluate the dataset, they tested several baselines from the literature, and proposed a novel model called Multi-Scale-Dilation that incorporates multi-scale feature extraction with dilated convolution, achieving 50.1% mIoU with pre-training and spatial-temporal regularization. Also, Li et al. introduced a novel architecture for UAV image segmentation, edge feature fusing and multi-level upsampling integrated with Deeplabv3+ (EMNet), which is based on the improved Deeplabv3+, in which an edge detection branch was added to the encoder module, while a multi-level upsampling strategy is incorporated in the decoder [14]. On the UAVid dataset, the EMNet achieved 71.46% mIoU and 80.46% mean pixel accuracy (mPA), improving mIoU by 7.11% over Deeplabv3+, while on the International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen dataset, it reached 91.80% mIoU and 95.42% mPA. Li et al. designed the Attentive Bilateral Contextual Network (ABCNet) for efficient semantic segmentation of fine-resolution remote sensing images [15]. The ABCNet adopts a dual-branch design with a spatial path to preserve low-level details and a contextual path with attention mechanisms to capture global dependencies, followed by a feature aggregation module for fusion. Experimental results showed that the ABCNet achieved 81.83% mIoU and 90.68% overall accuracy on the ISPRS Vaihingen dataset, and 88.56% mIoU and 91.10% overall accuracy on the ISPRS Potsdam dataset. Girisha et al. presented an encoder–decoder-based CNN architecture (UVid-Net) for UAV video semantic segmentation that incorporates temporal information through a two-branch encoder processing consecutive frames, along with a feature-refiner module to improve boundary accuracy [16]. On the ManipalUAVid dataset, the UVid-Net achieved a mIoU of 0.79 and an F1-score of 0.91, outperforming baseline models, with mIoU values of 0.75 for U-Net, 0.64 for FCN-8, and 0.65 for DeepLabV3+, respectively.

Object and vehicle detection in aerial imagery has been extensively studied using deep learning approaches. However, most of these works focus on general object detection tasks or broad categories such as vehicles, and relatively few studies specifically address the segmentation of heavy construction equipment in UAV imagery. In this study, a pipeline is proposed to segment the excavators in high-resolution UAV imagery using deep learning–based models. The dataset consists of high-resolution UAV images that were divided into smaller patches for efficient training. Binary masks corresponding to each image were created, with the excavator being marked as the foreground and all other parts being marked as the background. In this study, the issue of representing small objects is tackled via data-oriented and model-oriented techniques. The Region of Interest (ROI) cropping method has been applied to generate more pixels of the excavator in the training dataset, while the CBAM module is used to improve the feature extraction process by incorporating spatial and channel correlations. The main contributions of this study are summarized as follows:

- A new expert-annotated UAV dataset was constructed from post-earthquake imagery acquired after the 6 February 2023 Kahramanmaraş earthquakes.
- To the best of our knowledge, this is the first study to investigate semantic segmentation of excavators in post-earthquake UAV imagery collected in Türkiye.
- A foreground-aware ROI cropping strategy integrated with a CBAM-enhanced U-Net architecture is proposed to increase the amount of informative foreground content in training samples and improve feature representation.
- Extensive experiments demonstrate that the proposed method consistently outperforms the baseline semantic segmentation architectures across multiple quantitative evaluation metrics.

The structure of this article is organized as follows. Section 2 presents a detailed description of the dataset. Section 3 describes the proposed approach, including the overall pipeline. Section 4 reports and analyzes the experimental results along with the experimental setup. Finally, Section 5 concludes the study by summarizing the main findings and outlining potential directions for future work.

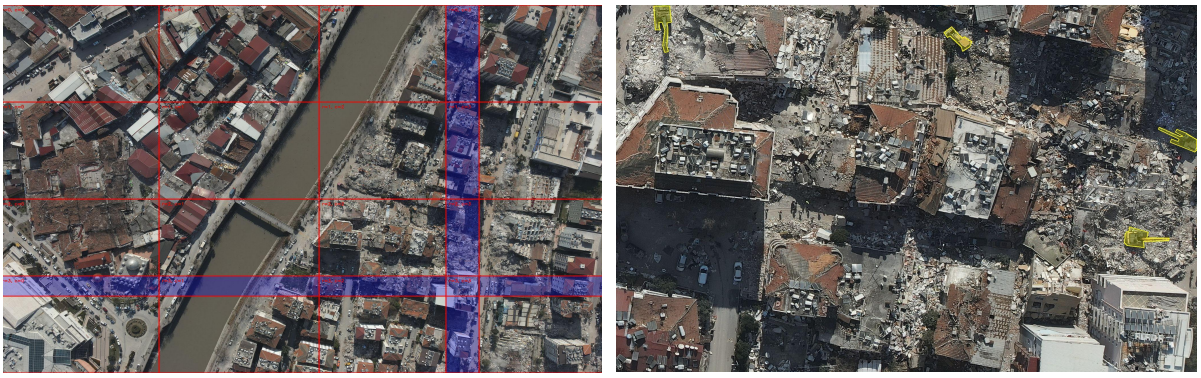
2. Dataset

This study uses UAV imagery from a post-disaster dataset collected in Adıyaman, Gaziantep, Hatay, and Kahramanmaraş, Türkiye, following the earthquakes of 6 February 2023. All UAV images represent real post-earthquake conditions and include diverse urban scenes containing collapsed buildings, debris accumulation, intact structures, and construction equipment involved in post-disaster operations. The original UAV images generally have a resolution of 8192×5460 pixels, providing sufficient spatial detail for identifying excavators and delineating their boundaries. To facilitate manual annotation while preserving the original 3:2 aspect ratio, each 8192×5460-pixel UAV image was divided into a 4×4 grid, producing 16 image patches of 2160×1440 pixels. The first three columns were extracted without horizontal overlap, starting at x-coordinates of 0, 2160, and 4320 pixels. Since

these columns covered only the first 6480 pixels of the image width, the rightmost column was aligned with the right boundary of the original image and started at x = 6032 pixels. Consequently, the rightmost column overlapped with the preceding column by 448 pixels. Similarly, the first three rows were extracted without vertical overlap, starting at y-coordinates of 0, 1440, and 2880 pixels. The bottom row was aligned with the bottom boundary of the original image and started at y = 4020 pixels, resulting in a 300-pixel overlap with the preceding row.

This tiling strategy ensured that the complete spatial extent of each original image was covered without resizing, padding, truncating, or discarding any border regions. The overlapping regions were retained in both adjacent patches and were not removed during dataset preparation. In Figure 1(a), the patch boundaries are indicated by red lines, while the regions in which the rightmost column and bottom row overlap with the preceding column and row are highlighted in semi-transparent blue. Compared with the original 8192 × 5460-pixel images, the resulting patches are less complex and more manageable for manual annotation. The intermediate resolution of 2160×1440 pixels was selected because directly delineating excavator boundaries in the original high-resolution images was impractical. The resulting 2160×1440-pixel image patches constituted the primary units used for manual annotation. This resolution allowed the annotators to recognize excavators clearly while retaining sufficient spatial context for precise boundary delineation. The patches were manually annotated by civil engineers specializing in post-disaster damage assessment. Excavator instances were delineated using polygons rather than bounding boxes, enabling their irregular object boundaries to be represented more precisely. For visualization, the annotated excavators are displayed using semi-transparent yellow masks with yellow boundaries. Representative annotation results are presented in Figure 1(b), demonstrating the precise localization of excavator regions that cannot be represented with equivalent spatial accuracy using bounding boxes. Consequently, the masks generated through this procedure provide detailed pixel-level annotations of the excavator instances.

A total of 1,130 original post-earthquake UAV images, generally having a resolution of 8192×5460 pixels, were used in this study. Applying the described 4×4 tiling procedure generated 16 patches from each original image, resulting in 18,080 image patches of 2160×1440 pixels in total. Among these 18,080 patches, 4,703 patches containing at least one excavator instance were retained for the final experimental dataset, while patches containing no excavator instances were excluded. Although the selected image patches contain a high proportion of background pixels, patches without any excavator instances were excluded because this study focuses on pixel-level segmentation in excavator-containing scenes. Therefore, model behavior in completely negative scenes was not evaluated and should be investigated in future studies. To prevent data leakage between the subsets, the dataset was divided on a city basis rather than through random image-patch splitting. Accordingly, the training subset comprised 1,263 image patches from Hatay and 2,433 from Kahramanmaraş, while the validation and test subsets contained 531 image patches from Adıyaman and 476 from Gaziantep, respectively. The final dataset consists of 4,703 UAV image tiles containing a total of 8,032 annotated excavator instances, corresponding to an average of 1.70 excavators per image. A strong pixel-level class imbalance is observed, where excavator pixels account for only 0.98% of all pixels, while background pixels constitute the remaining 99.02%. Across the entire dataset, the annotated excavators have an average area of approximately 17,859 pixels, with object sizes ranging from 232 to 158,319 pixels, indicating a highly diverse distribution of target scales.



(a) (b)

Figure 1. (a) Tiling of a 8192×5460 post-disaster UAV image into 2160×1440 pixel tiles. (b) Polygon-based annotation of excavators on a 2160×1440 pixel UAV image tile.

3. Methods

3.1 Residual networks-34 (ResNet-34)

ResNet-34 is a deep convolutional neural network designed to address the degradation problem associated with increasing network depth. It employs residual learning through identity shortcut connections, allowing the network to learn residual mappings instead of directly fitting the desired underlying functions. These connections improve gradient flow, mitigate the vanishing gradient problem, and facilitate stable training. ResNet-34 consists of 34 layers and uses basic residual blocks to learn hierarchical and discriminative feature representations while maintaining moderate computational requirements. Due to this balance between feature extraction capability and computational efficiency, ResNet-34 is commonly employed as an encoder backbone in semantic segmentation architectures [17].

3.2 U-Net

The U-Net architecture represents a foundational convolutional network design specifically developed for biomedical image segmentation, characterized by its symmetric encoder-decoder structure [18]. The encoder path contracts spatial resolution to capture high-level contextual features, while the decoder path symmetrically expands feature maps to enable precise pixel-level localization [18]. A defining characteristic of the classic U-Net is the utilization of skip connections that directly transfer feature maps from encoder stages to corresponding decoder stages. These skip connections preserve fine-grained spatial information lost during downsampling, thereby allowing the network to combine low-level structural details with high-level semantic representations. Consequently, this architecture facilitates accurate boundary delineations and effective segmentation performance, even when trained on limited datasets.

3.3 Proposed approach

Semantic segmentation is a basic problem in computer vision, which tries to associate a class label for each pixel of an image. Though there have been many breakthroughs in recent years in deep learning, semantic segmentation still faces some problems like complicated background, scaling problems, and very small object sizes. As a solution to these problems, a ROI cropping method is used based on the foreground detection technique. Moreover, a CBAM model is also used in the network to improve feature extraction based on channel and spatial dependence. The workflow of the proposed approach is given in Figure 2.

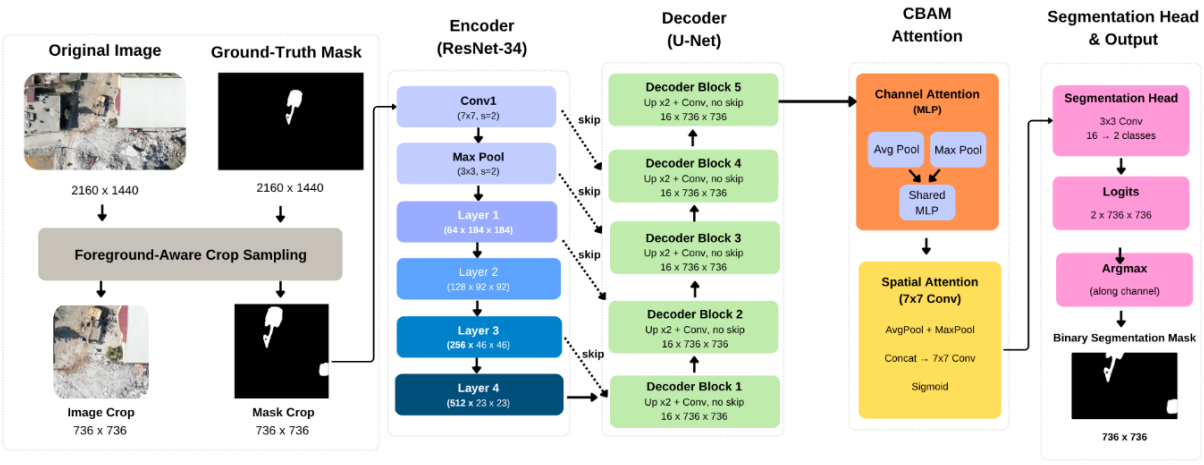


Figure 2. Overview of the proposed approach.

To overcome the foreground and background data imbalance issue in high-resolution UAV imagery, a foreground-aware ROI cropping technique is used to sample training patches. While random cropping of the input images is a common practice used in the semantic segmentation pipeline, it can lead to training samples containing

mostly background pixels, since target objects may be present only in a very small proportion of an image. To tackle this problem, a foreground-aware cropping scheme is utilized in order to bias the selection of foreground-containing samples for training purposes. With some predefined probability, crop centers are selected among foreground pixels; consequently, it ensures that an excavator object will appear in the training patch. On the other hand, in all other cases, standard random cropping is used. This way, while the focus of interest is foreground pixels, it still remains possible to cover various background examples. The use of such a technique is somewhat similar to a combination of class-aware cropping and hard example mining methods [19], which have been studied before. Specifically, prioritizing foreground samples enhances the feature representation of underrepresented target objects in terms of learning more discriminative features of objects that are not well-represented in training data.

Algorithm 1 presents the proposed foreground-aware ROI cropping strategy. Unlike conventional ROI-based cropping, which typically forces the region of interest to be entirely included within the cropped patch, the proposed method randomly selects a foreground pixel and uses it only as the crop center. Consequently, the resulting crop naturally preserves both the target object and its surrounding contextual information. Furthermore, foreground-centered cropping is applied with a probability of 0.5, while the remaining samples are generated using conventional random cropping. This balanced strategy increases the representation of foreground objects without sacrificing background diversity, thereby improving the learning of small and sparsely distributed targets.

Algorithm 1. Foreground-Aware ROI Cropping

Input	UAV image I and corresponding ground-truth mask M
Output	Cropped image Ic and cropped mask Mc
1	Identify all foreground pixels in M.
2	Generate a random number r ∈ [0,1].
3	If foreground pixels exist and r < 0.5, randomly select one foreground pixel.
4	Center the crop window on the selected foreground pixel.
5	Otherwise, randomly select a valid crop location within the image.
6	Extract the image and mask patches using the determined crop window.
7	Return the cropped image Ic and cropped mask Mc.

$$F_{CBAM} = CBAM(F_{decoder}) \quad (1)$$

$$\hat{Y} = SegHead(F_{CBAM}) \quad (2)$$

where F_{decoder} denotes the final decoder feature map, F_{CBAM} is the attention-refined feature representation, and $\hat{Y}$represents the predicted segmentation mask. The CBAM module sequentially applies channel and spatial attention before the segmentation head.

This study considers the U-Net architecture used as the basis for semantic segmentation of excavator pixels in UAV imagery. Specifically, the U-Net neural network uses an encoder-decoder structure where the encoder part encodes hierarchies of feature representations while the decoder part gradually restores spatial resolution in order to generate pixel-level segmentation results [18]. The use of skip connections between encoder and decoder blocks ensures that the U-Net architecture is able to integrate contextual information along with precise localization of objects at various depths. In other words, this technique enables transferring high-resolution feature maps generated by encoder blocks to subsequent blocks of the decoder part of the architecture. Thus, it becomes possible to recover lost spatial information after downsampling operations. Accurate detection of excavator objects on images of UAV video footage becomes difficult owing to the relatively small size of the objects and complex

backgrounds. Therefore, to overcome the problem described above and enhance the discriminative ability of features, a CBAM model is added to the U-Net+ResNet-34 architecture. The chosen attention mechanism is lightweight yet efficient and includes two components of attention that act sequentially. Specifically, CBAM is introduced after the decoder part output but before a segmentation head. The proposed attention mechanism follows the approach developed by Woo et al. [20], and it is introduced into the architecture of the U-Net in such a way that a compromise can be achieved between performance improvement and computational overhead. The addition of CBAM to the architecture enables the refinement of semantic features extracted by the decoder part of the network. This is important because the decoder is capable of integrating information provided by encoder-decoder skip connections. The channel attention component uses global pooling to re-calibrate feature maps in each channel, while the spatial attention component uses convolution to highlight relevant spatial areas.

4. Results and Discussion

4.1 Performance metrics

To thoroughly analyze the performance of the segmentations, several metrics have been calculated: IoU, mIoU, Dice coefficient, Precision, Recall, and G-Mean. These measures cover the assessment of pixel-level similarity, boundary match, accuracy of classification, as well as evaluation under imbalanced classes setting. Confusion matrix for pixel-wise segmentation can be utilized to evaluate the number of true positives, false positives, and false negatives. Using the values from the pixel-wise confusion matrix, it is possible to compute Precision, Recall, Dice score, and G-Mean in order to conduct a thorough evaluation of segmentation performance. In order to assess segmentation performance, a pixel-wise confusion matrix is created using the predictions and actual labels of the image. Each pixel in the matrix is considered a separate sample. True positive is the number of correctly classified pixels as a part of the target class, false positive is the number of pixels misclassified as the target class, true negative is the number of pixels correctly classified as not being in the target class, false negative is the number of pixels belonging to the target class and being misclassified. Evaluating the confusion matrix is essential to assess the model's capacity to discriminate between foreground and background classes.

4.2 Experimental setup

All experiments were conducted on a Linux-based workstation equipped with an Intel® Xeon® Gold 6240R CPU (2.40 GHz), 125 GB of system memory, and a single NVIDIA RTX A6000 GPU with 48 GB of VRAM. The models were implemented in Python 3.11 using PyTorch 2.7.1 with CUDA 12.6 and cuDNN 9.5.0.1. OpenCV 5.0.0, NumPy 2.2.6, scikit-learn 1.7.1, and segmentation-models-pytorch 0.3.3 were used for data processing, model implementation, and evaluation. To ensure a fair comparison, all experiments were conducted using the same hardware platform, software environment, and training configuration. The task was formulated as binary semantic segmentation, where excavators were defined as the foreground class and all remaining pixels as background. A fixed input resolution of 736×736 pixels was used during training, and no data augmentation was applied in order to isolate the effects of architectural and methodological components, such as ROI-based cropping and attention mechanisms. The models were trained using the AdamW optimizer with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-4}. The batch size was set to 16 for all models. Cross-entropy loss was used as the objective function. Training was conducted for up to 150 epochs with an early stopping strategy based on validation performance, using a patience of 15 epochs. During validation and testing, a deterministic center crop of 736 × 736 pixels was extracted from each 2160 × 1440 image tile, without resizing or sliding-window inference. All evaluation metrics were computed over this cropped region, while foreground-aware and random cropping strategies were applied only during training.

4.3 Results

Figure 3 presents the training and validation loss and IoU curves for (a) the baseline model, (b) the baseline model with ROI sampling, (c) the baseline model with CBAM, and (d) the proposed approach combining ROI sampling and CBAM. In all configurations, the training and validation losses decrease sharply during the initial epochs and subsequently stabilize, indicating successful convergence. Similarly, the training IoU increases consistently as optimization progresses. The validation IoU curves exhibit greater fluctuations throughout the training process than the training curves, which can be attributed to the limited foreground proportion, substantial variation in excavator size, and differences between the training and validation scenes. Nevertheless, the validation

losses remain generally low without a persistent upward trend. The combined ROI-CBAM configuration maintains stable loss behavior while achieving competitive validation IoU values, suggesting that foreground-focused sampling and attention-based feature refinement provide complementary benefits. Overall, the curves demonstrate that all variants converge effectively, while the proposed approach offers the most favorable balance between optimization stability and segmentation performance.

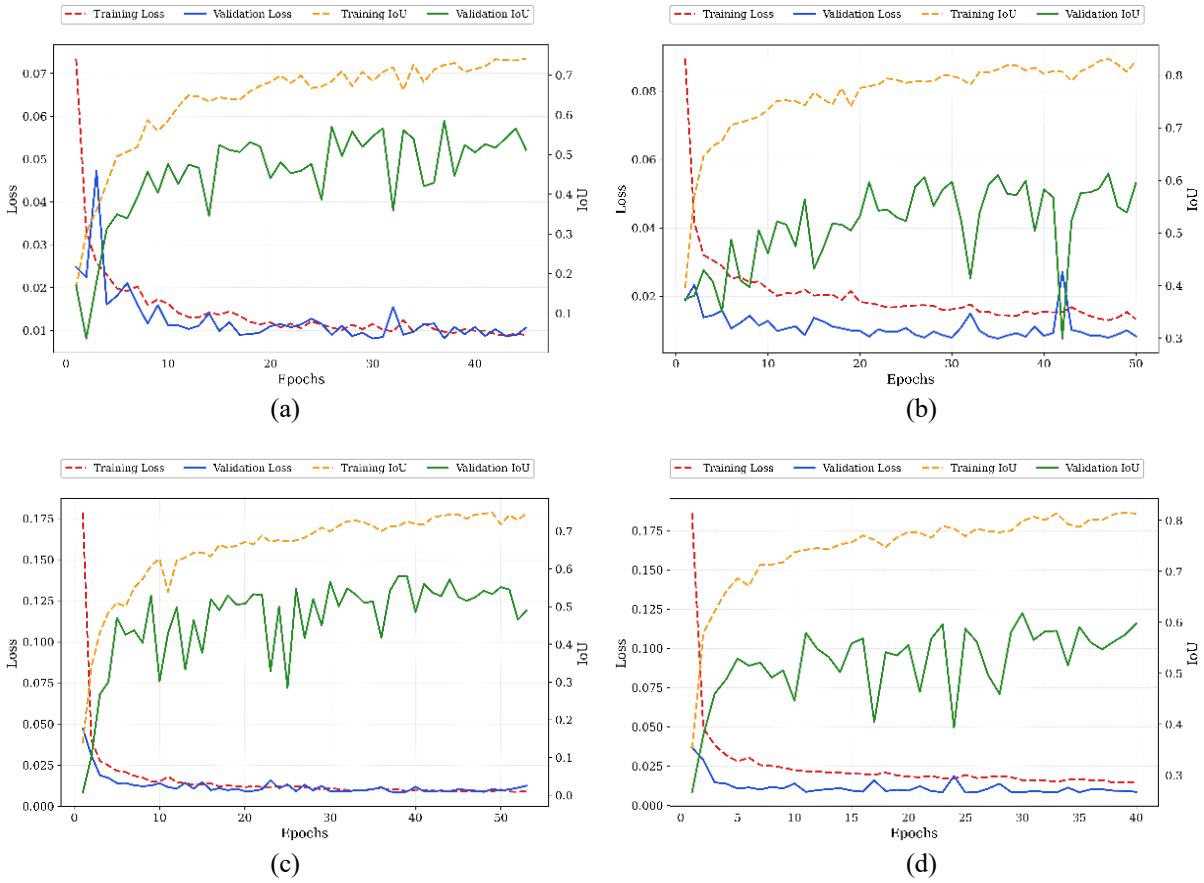


Figure 3. Loss and IoU curves over epochs for (a) Baseline, (b) Baseline + ROI, (c) Baseline + CBAM, and (d) Proposed Approach.

Table 1. The performance metrics of the ablation studies

Backbone	IoU	Dice	Precision	Recall	G-Mean
Baseline	0.509	0.675	0.864	0.554	0.744
Baseline + ROI	0.635	0.776	0.854	0.712	0.843
Baseline + CBAM	0.580	0.734	**0.866**	0.637	0.770
Proposed Approach	**0.663**	**0.798**	0.856	**0.747**	**0.863**

As shown in Table 1, the proposed approach achieves the best overall performance, with an IoU of 0.663, mIoU of 0.830, Dice score of 0.798, recall of 0.747, and G-Mean of 0.863. Compared with the baseline, it improves IoU, Dice, recall, and G-Mean by 0.154, 0.123, 0.193, and 0.119 respectively. ROI sampling alone substantially improves recall and overlap-based metrics, demonstrating greater sensitivity to excavator regions. CBAM alone achieves the highest precision of 0.866, indicating fewer false-positive foreground predictions, although its recall remains lower than that of the ROI-based configurations. Combining ROI sampling with CBAM provides a better balance between precision and recall and consequently produces the highest Dice and G-Mean scores. Overall, the

results presented in Table 1 demonstrate that ROI sampling and CBAM provide complementary benefits, with their combination yielding the most effective segmentation performance.

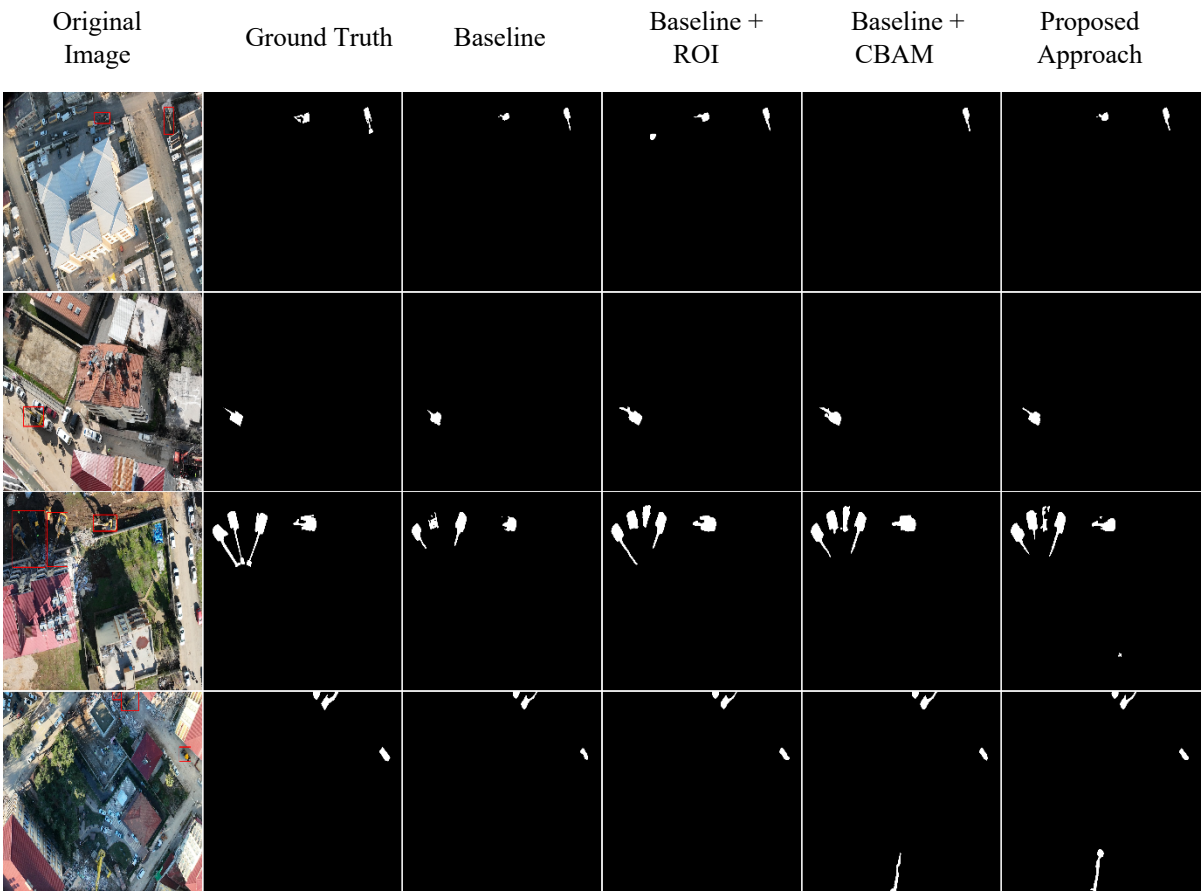


Figure 4. Qualitative comparison of segmentation results for excavator segmentation. From left to right: original UAV images, ground truth masks (GT), Baseline, Baseline + ROI, Baseline + CBAM and the proposed approach.

Figure 4 presents qualitative segmentation results for representative UAV images, arranged as the original image, ground truth, baseline prediction, baseline with ROI sampling, baseline with CBAM, and the proposed ROI–CBAM approach, respectively. All model variants successfully localize most excavators despite substantial variations in object size, orientation, shape, and surrounding background complexity. The baseline model generally identifies the main target regions but occasionally produces incomplete boundaries or misses smaller components. ROI sampling improves the representation of foreground regions, whereas CBAM helps refine spatial responses by emphasizing task-relevant features. The proposed approach generally produces masks that more closely follow the ground-truth shapes, particularly for scenes containing multiple, elongated, or closely positioned excavators. Nevertheless, minor boundary deviations and occasional small false-positive regions remain visible in challenging scenes [10]. Overall, the qualitative comparisons in Figure 4 demonstrate the complementary effects of ROI sampling and CBAM on excavator localization and boundary delineation.

Table 2 compares the trainable parameter counts and peak GPU memory usage of the evaluated ablation configurations. All configurations contain approximately 24.43 million trainable parameters, indicating that the ROI cropping and CBAM modules introduce no substantial increase in model complexity at the reported precision. While the ROI-based configuration maintains the same peak GPU memory usage as the baseline, the inclusion of CBAM raises the peak memory consumption to 15.81 GB. The combined ROI and CBAM configuration requires the same peak GPU memory as the CBAM-only configuration, suggesting that the additional memory overhead mainly originates from the attention mechanism.

Table 2. Comparison of trainable parameters and peak GPU memory usage across ablation configurations

Model	Trainable Parameters (M)	Peak GPU Memory (GB)
Baseline	24.43	14.58
Baseline + ROI	24.43	14.58
Baseline + CBAM	24.43	15.81
Baseline + ROI + CBAM	24.43	15.81

5. Conclusion

This study focuses on excavators because they are the major heavy equipment utilized in such operations as debris removal. It aims at examining the efficiency of ResNet based U-Net architectures and their application to debris recognition in UAV imagery, especially the effects of applying ROI-based cropping and CBAM attention mechanisms in segmentation models. The problem of detecting and segmenting small objects is approached both from the perspective of data-centric improvements and architectural changes in the design of the model itself. The experimental results show that combining foreground-aware data cropping with CBAM attention improves excavator segmentation performance. The proposed approach achieves the highest IoU, mIoU, Dice, Recall, and G-Mean values, while the highest Precision is obtained by the baseline with CBAM. The qualitative results also indicate that the proposed approach produces more accurate and complete segmentation masks, particularly for small and sparsely distributed excavators.

References

[1] Zhu X.X., Tuia D., Mou L., Xia G.S., Zhang L., Xu F., Fraundorfer F. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci Remote Sens Mag 2017; 5(4): 8–36.

[2] Xia G.S., Bai X., Ding J., Zhu Z., Belongie S., Luo J., Zhang L. DOTA: A large-scale dataset for object detection in aerial images. Proc IEEE Conf Comput Vis Pattern Recognit 2018; 3974–3983.

[3] Fang W., Ding L., Luo H., Love P.E.D. Falls from heights: A computer vision-based approach for safety harness detection. Autom Constr 2018; 91: 53–61.

[4] Li Q, Sun M, Dong L, Gao X, Wang Z, Zhang H. HCD-Mask: A multi-task model for small object detection and instance segmentation in high-resolution UAV images. 2022 IEEE International Conference on Industrial Technology (ICIT); 22–25 August 2022; Shanghai, China. pp. 1–8.

[5] Yu A, Quan Y, Yu R, Guo W, Wang X, Hong D, Zhang H, Chen J ve diğerleri. Deep Learning Methods for Semantic Segmentation in Remote Sensing with Small Data: A Survey. Remote Sens 2023; 15(20): 4987.

[6] Qiu Z, Gong T, Cong R, Zhao Y, Bai H. UESDNet: UAV Visual Multitask Perception Via Expert-Guided Prior With Mutual Optimization of Segmentation and Detection. IEEE J Sel Top Appl Earth Obs Remote Sens **2026**; 19: 3157–3174.

[7] Meng L., Peng Z., Zhou J., Zhang J., Lu Z., Baumann A., Du Y. Real-time detection of ground objects based on unmanned aerial vehicle remote sensing with deep learning: Application in excavator detection for pipeline safety. Remote Sens 2020; 12(1): 182.

[8] Zhang S., Zhang L. Construction site safety monitoring and excavator activity analysis system. Constr Robot 2022; 6(2): 151–161.

[9] Hua T., Lang B., Sun J., Lin G., Wang J., Daping L., Wang X. Excavator identification method based on computer vision technology in nighttime scenes. Eng Constr Archit Manag 2025; 1–22.

[10] Kim I.S., Latif K., Kim J., Sharafat A., Lee D.E., Seo J. Vision-based activity classification of excavators by bidirectional LSTM. Appl Sci 2022; 13(1): 272.

[11] Fang W., Ding L., Zhong B., Love P.E.D., Luo H. Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. Adv Eng Inform 2018; 37: 139–149.

[12] Spasev V., Dimitrovski I., Chorbev I., Kitanovski I. Semantic segmentation of unmanned aerial vehicle remote sensing images using SegFormer. Int Conf Intell Syst Pattern Recognit; 2024 June; 108–122.

[13] Lyu Y., Vosselman G., Xia G.S., Yilmaz A., Yang M.Y. UAVid: A semantic segmentation dataset for UAV imagery. ISPRS J Photogramm Remote Sens 2020; 165: 108–119.

[14] Li X., Li Y., Ai J., Shu Z., Xia J., Xia Y. Semantic segmentation of UAV remote sensing images based on edge feature fusing and multi-level upsampling integrated with Deeplabv3+. PLoS One 2023; 18(1): e0279097.

[15] Li R., Zheng S., Zhang C., Duan C., Wang L., Atkinson P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. ISPRS J Photogramm Remote Sens 2021; 181: 84–98.

[16] Girisha S., Verma U., Pai M.M., Pai R.M. Uvid-net: Enhanced semantic segmentation of UAV aerial videos by embedding temporal information. IEEE J Sel Top Appl Earth Obs Remote Sens 2021; 14: 4115–4127.

[17] He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. Proc IEEE Conf Comput Vis Pattern Recognit 2016; 770–77.

[18] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI); 5–9 October 2015; Munich, Germany. pp. 234–241.

[19] Shrivastava A., Gupta A., Girshick R. Training region-based object detectors with online hard example mining. Proc IEEE Conf Comput Vis Pattern Recognit 2016; 761–769.

[20] Woo S., Park J., Lee J.Y., Kweon I.S. CBAM: Convolutional Block Attention Module. Proc Eur Conf Comput Vis 2018; 3–19.