

# UAV-Based Drifting Victim Detection in Flash Floods Using Region-Based Convolutional Neural Networks

Prima Kristalina

`prima@pens.ac.id`

Politeknik Elektronika Negeri Surabaya

Aries Pratiarso

Politeknik Elektronika Negeri Surabaya

A. Fauzi Makarim

Politeknik Elektronika Negeri Surabaya

Munawar Munawar

Politeknik Negeri Lhokseumawe

---

## Research Article

**Keywords:** Mask R-CNN, UAV-Based Detection, Search and Rescue, Flash Flood, Drift Trajectory Prediction

**Posted Date:** September 1st, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10854549/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# UAV-Based Drifting Victim Detection in Flash Floods Using Region-Based Convolutional Neural Networks

Prima Kristalina<sup>1</sup> . Aries Pratiarso<sup>1</sup> . A. Fauzi Makarim<sup>1</sup> . Munawar<sup>2</sup>

<sup>1</sup>Department of Electrical Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia

<sup>2</sup>Department of Electrical Engineering, Politeknik Negeri Lhokseumawe, Aceh, Indonesia

Corresponding Author: prima@pens.ac.id

## Abstract

This study proposes a deep learning-based system to assist Search and Rescue (SAR) teams in predicting the drift direction of objects during flash flood events. The proposed system utilizes aerial video data acquired by unmanned aerial vehicles (UAVs) to detect drifting persons in successive video frames using the Mask R-CNN algorithm and to determine their geographic locations based on embedded Global Positioning System (GPS) metadata. To evaluate the effectiveness of the proposed approach, the conventional Mask R-CNN model is compared with a fine-tuned Mask R-CNN model trained on a customized dataset, with performance assessed in terms of detection accuracy and confidence. Experimental results demonstrate an average detection confidence exceeding 90%, particularly after fine-tuning the Mask R-CNN model using the customized dataset. These results indicate that the proposed approach can provide reliable object detection and spatial information for estimating drift trajectories in dynamic river environments. The system has the potential to support SAR operations by providing timely and accurate information on the location and movement direction of drifting persons, particularly in flash flood conditions characterized by substantial variations in river elevation and flow dynamics.

**Keywords—** Mask R-CNN . UAV-Based Detection . Search and Rescue . Flash Flood . Drift Trajectory Prediction

## 1. INTRODUCTION

Flash floods are among the most hazardous hydrometeorological disasters because they can develop rapidly, generate high-velocity water flows, and substantially alter river and surrounding land conditions within a short period. Unlike conventional inundation floods, flash floods are strongly influenced by rainfall intensity, watershed morphology, river gradient, drainage characteristics, soil conditions, and land-cover changes. In mountainous and steep-slope watersheds, intense rainfall can produce rapid surface runoff and a sudden increase in river discharge, leaving limited time for communities and emergency responders to react. Previous studies in Aceh have shown that flash-flood vulnerability is closely associated with watershed topography, river gradients, drainage density, slope stability, sedimentation, and potential natural damming. In particular, steep river gradients can accelerate downstream flow and contribute to sudden flood surges, while abrupt changes in terrain may intensify the propagation of floodwater toward downstream settlements [1], [2].

The importance of these characteristics was again demonstrated by the severe hydrometeorological disasters that affected Aceh, Indonesia, in November 2025. During this period, several districts in Aceh experienced extensive flooding following periods of moderate to heavy rainfall. The National Agency for Disaster Management (BNPB) reported flooding in areas including Aceh Utara, Aceh Timur, Aceh Barat, Aceh Selatan, Aceh Tenggara, Aceh Singkil, and other parts of the province. In Aceh Utara, prolonged rainfall increased surface-water volume and was compounded by inadequate drainage capacity and runoff from surrounding upland areas. Flooding affected numerous villages and disrupted roads, electricity, communications, and public services. In Aceh Tenggara, flooding also resulted in damaged bridges and disrupted road access, demonstrating how rapidly changing river conditions can restrict conventional emergency operations [3].

The severity of the situation became particularly evident in the number of casualties and missing persons. By 29 November 2025, BNPB reported 47 deaths and 51 missing persons in Aceh alone, with victims distributed across several affected districts [4]. Earlier reports had already documented missing persons in Bener Meriah and Gayo Lues, while emergency operations were complicated by damaged infrastructure and inaccessible locations. BNPB subsequently emphasized the need to accelerate emergency response operations in affected areas. These circumstances demonstrate that the effectiveness of search and rescue (SAR) operations depends not only on the availability of personnel and equipment

but also on the ability to rapidly obtain reliable information about the affected environment. When roads are damaged, communication networks are disrupted, and river conditions change rapidly, ground-based assessment may become difficult or dangerous. Consequently, technologies capable of rapidly acquiring spatial information from above the disaster area are increasingly important for emergency response.

Understanding the characteristics of local rivers is therefore essential for improving flash-flood response and SAR planning. Rivers in Aceh are not hydraulically homogeneous; their flow behaviour is influenced by variations in elevation, channel geometry, slope, watershed morphology, sediment deposition, vegetation, and surrounding land use. Research conducted in Aceh has demonstrated that flash-flood hazards can be associated with steep upstream terrain, high rainfall, river gradients, erosion, drainage density, and natural damming processes [1], [2]. Similarly, investigations of the Jambo Aye basin indicate that upstream areas with steep slopes and high rainfall can be particularly susceptible to flash-flood generation, whereas downstream areas may experience inundation because of lower gradients and reduced drainage capacity [5]. Therefore, identifying the spatial characteristics of floodwater and understanding how water propagates through river channels and adjacent areas are important for determining potentially hazardous locations and supporting the prioritization of SAR activities.

Conventional ground surveys provide valuable information but may not be sufficiently rapid or spatially comprehensive during an active disaster. Satellite imagery can provide extensive spatial coverage; however, its usefulness for emergency response may be constrained by spatial resolution, revisit time, cloud cover, and the availability of suitable imagery during rapidly developing events. Unmanned aerial vehicles (UAVs), in contrast, can acquire high-resolution imagery directly over disaster-affected areas with flexible flight planning and relatively rapid deployment. UAV imagery can provide detailed visual information about river channels, inundated areas, damaged infrastructure, debris, and human subjects from an aerial perspective. Recent research has demonstrated the potential of UAV-based imaging combined with deep learning for obtaining detailed spatial information about river channels and improving flood-related analysis [6]. UAV-based deep learning has also been demonstrated to be effective for extracting flooded regions from aerial imagery, providing an alternative to conventional image classification approaches [7]. For SAR applications, however, merely acquiring aerial imagery is insufficient. The large volume of visual information generated by UAVs must be processed efficiently to identify objects or regions that are relevant to emergency operations. Manual inspection of individual UAV frames can be time-consuming and may become impractical when large areas must be assessed rapidly. This challenge motivates the application of computer vision and deep learning techniques capable of automatically identifying relevant objects and delineating their spatial boundaries.

Region-based Convolutional Neural Networks (R-CNNs) provide an effective framework for object detection and image segmentation by combining region proposal mechanisms with deep feature extraction. Among these approaches, Mask R-CNN extends Faster R-CNN by introducing an additional branch for pixel-level instance segmentation. Consequently, the method can simultaneously identify an object, determine its bounding region, and generate a segmentation mask representing its spatial extent. This characteristic is particularly advantageous for UAV-based flood imagery because disaster scenes may contain visually complex backgrounds, including flowing water, vegetation, debris, buildings, boats, and other objects with appearances that can overlap with potential victims. Instead of relying solely on rectangular bounding boxes, instance-level segmentation provides more detailed information about the shape and location of detected objects. The use of Mask R-CNN is therefore well suited to UAV-based drifting-victim detection because it can provide object-level localization together with pixel-level delineation. Such information can be used not only to determine whether a potential victim is present but also to identify the spatial relationship between the detected subject and surrounding floodwater. This capability is particularly relevant in river environments where victims may occupy only a small portion of a high-resolution aerial image and may be visually similar to floating debris. Previous flood-mapping research has demonstrated that convolutional neural networks can effectively extract flooded regions from UAV imagery, indicating the suitability of deep-learning-based computer vision for aerial disaster assessment [8]. However, detecting individual drifting victims presents an additional challenge because

the target objects are often small, partially obscured, and embedded within visually heterogeneous river environments. In this study, Mask R-CNN is fine-tuned using a customized UAV image dataset acquired from a disaster-affected river environment. The study focuses exclusively on image-based information obtained from UAV observations, allowing the proposed approach to investigate the capability of visual deep learning to detect drifting victims without relying on external temporal sensor data. The resulting detection information is intended to support rapid visual assessment of flood-affected river areas and provide actionable information for SAR teams. By focusing on imagery acquired directly from the disaster environment, the study also addresses the practical requirement for localized datasets that reflect the appearance, scale, background characteristics, and environmental conditions encountered during actual flood events.

The main contributions of this study are threefold. 1) We develop a customized UAV-based image dataset representing drifting-victim conditions in a flood-affected river environment, providing localized visual data for training and evaluating deep-learning-based detection. 2) We also develop and fine-tune a Mask R-CNN-based detection framework to identify drifting victims from aerial imagery and evaluate its effectiveness in challenging river and flood environments. The instance-segmentation capability of Mask R-CNN enables the detected targets to be spatially delineated rather than represented only by rectangular bounding boxes. 3) We demonstrate the potential of UAV-based Mask R-CNN detection as a rapid visual assessment tool for SAR operations by providing spatial information on detected victims and the surrounding flood-affected environment. The proposed approach is expected to contribute to more efficient aerial reconnaissance and support decision-making during emergency response, particularly in river environments where high water velocity, complex terrain, and restricted ground access make conventional search operations difficult.

## **2. RELATED WORKS**

### **2.1 UAV-Based Search and Rescue (SAR) Systems**

UAVs have become central to disaster management due to their ability to eliminate the risk of human loss, low cost, and capacity to provide fast, effective surveillance over large and hazardous areas. At the strategic level, Bakir and Kasimoglu proposed a multi-criteria decision-making model for long-term UAV base-location and deployment planning, validated on a real-world earthquake scenario with measurable improvements in distance and cost, highlighting the importance of deployment planning alongside on-board detection capability [9]. At the detection level, Siktari et al. developed a UAV subsystem focusing on people detection, face recognition, and tracking of identified individuals, using YOLOv11, YOLOv11-pose, and dlib CNN models within a ROS2 framework, with keypoint-based tracking used to maintain a safe following distance, demonstrating real-time feasibility in initial experiments [10]. Complementing this individual-level detection focus, Mohd Daud et al.'s scoping review found drones broadly valuable across mapping, search and rescue, transportation, and training, but identified a lack of data on the use of drones specifically in disaster victim identification, recommending greater focus on this area to reduce identification delays [11]. At the coordination level, Kareem et al. proposed a bio-inspired swarm UAV framework combining thermal sensing with optimization-based coordination for real-time, cooperative exploration of a discretized search grid, achieving 80% area coverage within 60% of mission time using Particle Swarm Optimization — prioritizing multi-UAV search efficiency over single-target detection accuracy [12]. Moreover, these studies span deployment planning [13], CNN-based person detection and tracking [10], application-domain gaps [11], and swarm search coordination [12], but none addresses detecting or tracking victims drifting within fast-moving floodwater, that will be a distinct challenge this study addresses.

### **2.2 Object Detection Using Convolutional Neural Networks (CNNs)**

Convolutional Neural Networks (CNNs) have become the foundation of modern object detection, owing to their ability to capture local spatial patterns through convolutional layers, well suited for understanding visual hierarchies and features within images. Issaoui et al. provide a comprehensive review of CNN-based object detection architectures, observing that CNNs demonstrate strong performance in image-based object recognition, while single-stage models such as YOLO are known for their fast and efficient detection in

real-time applications. Reinforcing the accuracy vs speed trade-off that shapes detector selection for time-critical tasks such as SAR [7]. Zhang et al. proposed a Faster R-CNN-based detector for water surface objects, motivated by the observation that classical background-subtraction and image-segmentation methods are highly susceptible to variable scenes, with accuracy greatly reduced by changing sunlight and waves. Their proposed network integrates low-level and high-level features to improve detection accuracy for multi-scale floating objects in complex natural river scenes, directly demonstrating the applicability of region-based CNNs to dynamic aquatic backgrounds relevant to this study [13]. Shah and Tembhurne's review situates CNN-based detectors alongside the more recent transformer-based paradigm, noting that transformers leverage self-attention mechanisms to capture both local and global context in images, and are increasingly used either as replacements for or in conjunction with convolutional networks, indicating that CNN architectures remain a core building block even as detection research evolves toward hybrid designs [14]. Applied specifically to river and water-surface environments. Complementing these detection-focused studies, Kareem et al.'s swarm UAV framework further illustrates the practical need for efficient, real-time processing in field deployments, having achieved 80% area coverage within 60% of mission time through optimized coordination. It is a reminder that detection accuracy must be paired with computational efficiency for operational SAR use [12]. Together, these studies confirm that CNN-based detectors, particularly region-based variants, offer strong accuracy for complex and dynamic visual scenes such as flowing water [13], while broader reviews [14,15] frame the ongoing trade-offs between accuracy, speed, and emerging transformer-based alternatives, trade-offs directly relevant to selecting Mask R-CNN as the detection backbone for this study.

### **2.3 Mask R-CNN for Instance Segmentation**

Mask R-CNN extends the Faster R-CNN architecture by adding a parallel branch for pixel-level instance segmentation alongside conventional bounding-box detection and classification. This capability is particularly valuable in domains where object shape, not just object location, carries meaningful information — such as estimating the orientation, size, or boundary of an irregularly shaped floating object for trajectory prediction. Several studies have adapted the base Mask R-CNN architecture for domain-specific challenges; for example, one study proposed an adaptive lightweight Mask R-CNN system using an upgraded Region Proposal Network and a MobileNetV3 backbone, achieving real-time processing at 30 frames per second for underwater debris segmentation. Such adaptations demonstrate that, while Mask R-CNN's core architecture is computationally heavier than single-stage alternatives, targeted optimization can bring it within range of near-real-time operational requirements [15].

### **2.4 CNN and Mask R-CNN Applications to Water Surfaces and Floating Objects**

A growing body of research has applied CNN-based detectors directly to floating and underwater objects, a domain closely related to drifting-victim detection. Di et al. proposed an enhanced YOLOv8 model for solid floating waste detection, building on prior work in which Minh et al. applied Mask R-CNN to the detection and quantification of floating waste for automated monitoring along riverbanks, while other researchers integrated Mask R-CNN with a dust feature enhancement module based on the hue-lightness-saturation color space to reduce misjudgment of debris regions — illustrating the range of Mask R-CNN adaptations already applied to riverine waste monitoring [16]. For real-time river monitoring, Zhang et al. proposed EYOLOv3, a one-stage detector designed because two-stage CNN models achieve high detection accuracy but struggle to reach real-time detection, while one-stage models are faster but less accurate. Their review of prior region-based approaches notes that Faster R-CNN with class activation showed strong adaptability for detecting small floating objects, and that Mask R-CNN variants have similarly been extended for refined floating-object segmentation, reinforcing the accuracy-speed trade-off central to selecting a detector for time-critical river monitoring [17]. Extending Mask R-CNN beyond surface debris to underwater ecosystem monitoring, Mahmoud et al. constructed a high-performance underwater object detection and segmentation model based on Mask R-CNN using VGG-16 as the feature backbone, incorporating preprocessing such as contrast enhancement and noise reduction alongside data augmentation during training. The model was evaluated on a dataset of 638 labeled underwater images spanning seven classes of sea creatures, demonstrating Mask R-CNN's adaptability for instance segmentation across visually complex, low-visibility aquatic environments beyond simple floating debris [18].

Collectively, these studies confirm that Mask R-CNN and related region-based CNNs are well established for detecting and segmenting objects in river, surface, and underwater settings [16–18], but their application has remained centred on waste, debris, and marine-life monitoring rather than aerial UAV-based detection of drifting flash-flood victims.

### 3. METHODOLOGY

#### 3.1. System Design

The system design was divided into Data collection, data processing, model development, and implementation and evaluation. Regarding dataset collection, we gathered image datasets, sourced from both social media and primary data captured via drones. The requirement of devices in the data collection is listed in Table 1.

Table 1. List of Equipment for data collection

| No. | Device   | Specification                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-----|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1   | Drone    | DJI Mavic Air 2S, a lightweight drone (~200 g) equipped with a 20 MP camera (2.4 $\mu\text{m}$ pixel size) and a gimbal stabilization system                                                                                                                                                                                                                                                                                                                   |
| 2   | Computer | Processor: Intel Core i7 or equivalent, with multi-core capability to enhance data processing efficiency.<br>RAM: At least 16 GB, to handle video data processing and deep learning model training.<br>GPU: An NVIDIA GPU with CUDA support, such as the GTX 1080 or higher, is necessary to accelerate the Mask R-CNN model training process.<br>Storage: An SSD with a minimum capacity of 512 GB to store video data, datasets, and model training outputs. |

#### A. Data Collection

The primary data collection was conducted at the Telocor Marine Tourism area in Sidoarjo on November 23, 2024. There were 26 video files resulting from aerial drone photography taken at various altitudes along the course of the Porong River. The video captured the river area where objects moving in the river, were recorded. There were 4 points we stop for data collecting, ie. Point A, B, C, and D along Porong River, as shown at Figure 1. Using a drone as the data collection tool enabled the recording of videos from wider angles and optimal altitudes to effectively capture object movements on the water surface. Following the video recording, the footage was extracted into multiple image frames. Each frame obtained from the drone video serves as an image input to be further processed for annotation purposes. From 589 images resulted as a dataset, 515 data were used to train the R-CNN model to detect the presence of objects in the river, while the 25 images served as data test and 48 images used for validation data.

The other images were depicted from social media. A total of 463 images were split into 278 images for training, 93 for validation, and 92 for testing. The entire data extraction and processing workflow was carried out on the Roboflow platform. Image annotation was performed using the labels "Severe," "Moderate," and "Mild." Sample images from the dataset are shown in Figure 2.

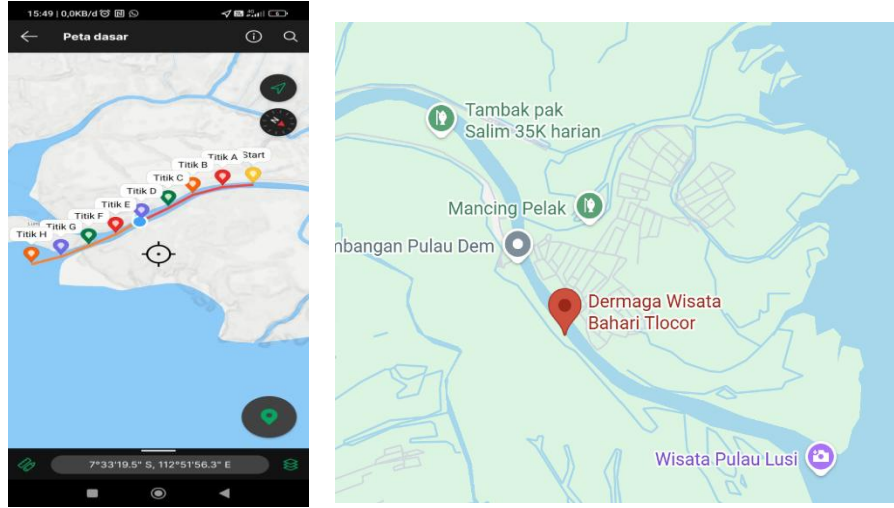


Figure 1. Location of Data Collected in Porong river (-7.550395827122008 Long, 112.84687003611015 Lat)



Figure 2. Sample dataset of flash flood images

### B. Data Processing

Let the drone image dataset be:

$$X_I = \{x_i | x_i = UAV\_Captured(t_i, l_i), i = 1, 2, \dots, N\} \quad (1)$$

where all data set image is  $i=1, 2, \dots, N$ , captured by UAV (drone) in time  $t_i$  in geographical location  $l_i$ .

Image data processing aims to ensure the data was consistent, clean, and suitable for the feature extractor. This process involved five stages: image quality checking, image annotation, image resizing, pixel normalization, and image augmentation.

The quality check involves identifying images with issues such as file corruption, excessively low resolution, blank pages, duplication, incorrect file formats, or irrelevance to the target. Images exhibiting these issues must be removed from the dataset before proceeding to Deep Learning-based training. Examples of images requiring removal from the dataset are shown in Figure 3.



Figure 3. Several images that need to be discarded: blank images, duplicate images, and low-resolution images.

The image annotation process involved labeling specific objects within the image. The objective of this process was to train the deep learning model to recognize object patterns that share similarities and were classified as the same type of object. This labeling was represented by bounding boxes accompanied by labels, as illustrated in Figure 4, which shows labels indicating severe and minor flooding.



Figure 4. Annotations for flood status: Severe and Mild.

Image resizing is necessary for deep learning models to ensure a consistent input size. For Mask R-CNN, a type of CNN model, the input size used is typically  $224 \times 224$  or  $256 \times 256$  pixels. Therefore, at this stage, images of varying dimensions are resized using CNN model functions; this ensures that the model consistently receives images of the same size during training. Pixel normalization is the process of scaling pixel values to a specific range, making the image data consistent and suitable for use as input in deep learning models. For example, pixel values ranging from 0 to 255 might be normalized to a 0–1 range—such as through Min-Max Normalization, as illustrated in Figure 5.

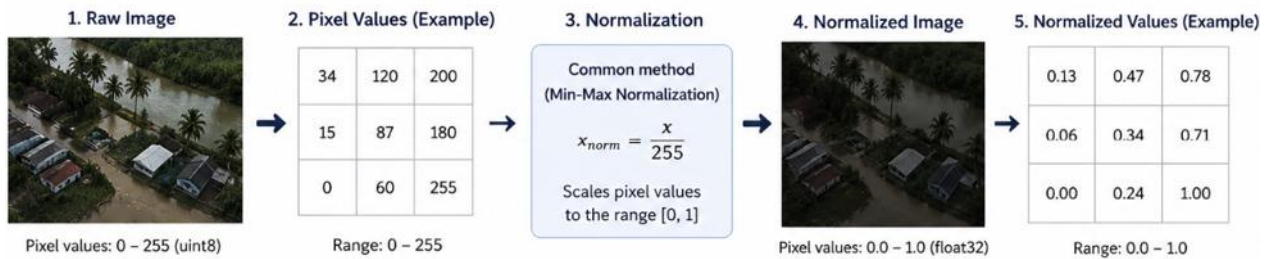


Figure 5. Example of pixel normalization from the 0–255 range to the 0–1 range using the Min-Max Normalization method.

The pixel normalization using the Min-Max method is expressed by Equation (2):

$$x' = \frac{x}{255} \quad (2)$$

This equation converts the pixel range of  $0 < x < 255$  to  $0 \leq x < 1$ .

Image augmentation is a process used to increase the volume of image data available for training and testing Deep Learning models. Acquiring large amounts of image data is challenging due to the limited storage capacity of image-capturing devices (such as camera-connected nodes or drone cameras). This contrasts with textual data, where file sizes are significantly smaller than those of images. Image augmentation can be performed in various ways, such as horizontal flipping, rotation by specific degrees, scaling, cropping, and adjustments to brightness, contrast, blur, and Gaussian noise. An image of a flash flood can undergo the various augmentation techniques mentioned above, as illustrated in Figure 6. The image data processing pipeline is shown in Figure 7.



Figure 6. Various augmentation models for a flash flood image



Figure 7. Image data preprocessing pipeline

### C. Model Development

At this stage, separate feature extraction was performed, involving modality-specific feature extraction for the textual datasets and the image dataset.

For the image dataset, separate feature extraction was carried out using Mask R-CNN. The Mask R-CNN component acts as a specialized feature extractor that transforms each pre-processed drone image into a fixed-dimensional feature vector. The Mask R-CNN architecture, as shown in Figure 8, consists of

a backbone, a Feature Pyramid Network (FPN), a Region Proposal Network (RPN), RoI feature extraction, and a prediction head.

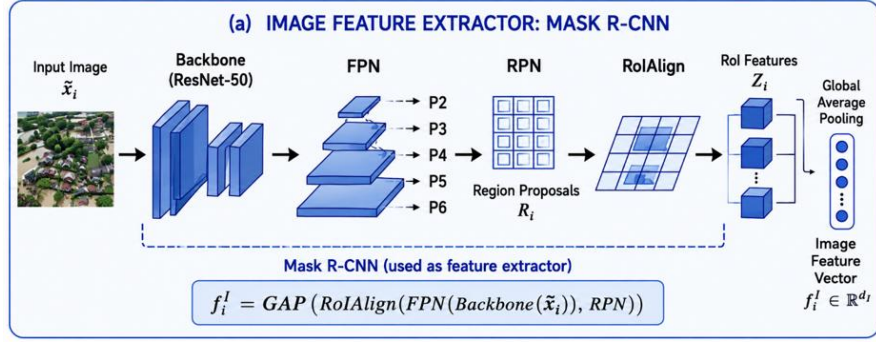


Figure 8. Mask R-CNN architecture with an input dataset of images captured by a drone.

The steps for separate feature processing using Mask R-CNN are as follows:

1. The backbone transforms the image into a feature map as follows:

$$F_i^{backbone} = B(\tilde{x}_i; \theta_B) \quad (3)$$

Where  $B(\cdot)$  = Backbone CNN from Mask R-CNN,  $\theta_B$  = backbone parameter, and  $F_i^{backbone}$  = extracted convolutional feature maps.

2. With the Feature Pyramid Networks (FPN) value  $F_i^{FPN} = \text{FPN}(F_i^{backbone}; \theta_{FPN})$ , then FPN will produce multi-scale feature map as follows:  $F_i^{FPN} = \{P_1, P_2, P_3, P_4, P_5, P_6\}$  which can reveal objects at various scales in drone imagery.
3. The Region Proposal Network (RPN) identifies region candidates corresponding to the detected areas. The RPN formulation is

$$R_i = \text{RPN}(F_i^{FPN}; \theta_{RPN}) \quad (4)$$

where  $R_i$  represents the proposed region of interest (RoI).

4. The features from the RoIs will be extracted using RoIAlign as follows:

$$Z_i = \text{RoIAlign}(F_i^{FPN}; R_i) \quad (5)$$

This yields feature representations corresponding to the detected regions.

In the case of flash flood detection, the resulting RoI representations can capture areas associated with various visual patterns linked to flooding.

5. Next, the extracted RoI features can be combined to obtain a fixed-dimensional image representation. Using Global Average Pooling, a vector is obtained for each image feature:  $f_i^l = \text{GAP}(Z_i)$  atau  $f_i^l \in \mathbb{R}^{d_l}$  where  $d_l$  is the dimension of the represented image.

Overall, the vector for each image feature is expressed as:

$$f_i^l = \text{GAP}\left(\text{RoIAlign}\left(\text{FPN}\left(B(\tilde{x}_i; \theta_B)\right); \text{RPN}\left(\text{FPN}\left(B(\tilde{x}_i; \theta_B)\right)\right)\right)\right) \quad (6)$$

#### D. Model Evaluation and Validation

Model evaluation encompasses: evaluating accuracy based on the training dataset and model accuracy when processing real-world data, comparing accuracy relative to the drone's altitude above the object, and comparing Average Precision (AP) and Average Recall (AR) between standard Mask R-CNN and fine-tuned Mask R-CNN for both bounding box detection and segmentation.

Two accuracy curves are compared:

**Model-1**, trained on the original dataset

**Model-2**, trained on the modified dataset

The TensorBoard visualization displays training logs from the output directory, which include metrics such as accuracy and loss across epochs.

### 1. Accuracy Curves

Model-1 shows a gradual increase in accuracy, with a noticeable jump between epochs 19 to 39, then stabilizing around 0.85–0.89. However, the false positive rate remains high and fluctuating, while false negatives are not yet optimal. This suggests the model sometimes incorrectly detects non-existent objects or misses existing ones in contrast, Model-2, trained with the modified dataset, performs significantly better. Its accuracy rapidly increases early in training and stabilizes at around 0.95. The false positive rate drops to near zero, and false negatives also decrease substantially. This indicates a more accurate and reliable detection model. The comparison of two models is shown in Figure 9.

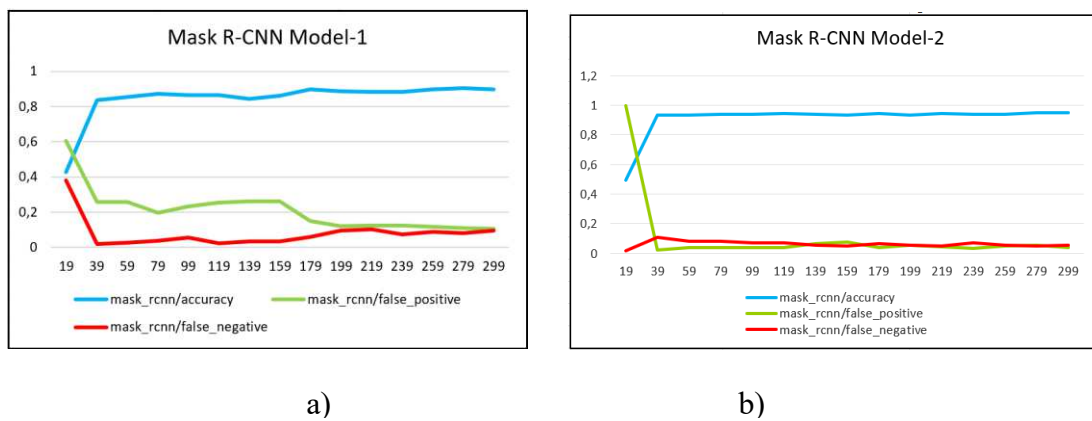


Figure 9. Accuracy Curve of Mask R-CNN on a) Original Dataset, and b) Modified Dataset

### 2. Average Precision

a. Using Bounding Box (bbox) detection.

On Mask R-CNN: This standard model generally shows good performance in bounding box detection. However, its accuracy tends to decrease when applied to datasets with different characteristics than the dataset used for training (e.g., COCO). Meanwhile, in Fine-tuned Mask R-CNN, after the adjustment process the AP value increases significantly. The model becomes more adaptive to that specific dataset, resulting in better localization and object detection.

b. Using Segmentation (segm) detection.

Regarding R-CNN: While the model performs reasonably well at instance segmentation, it may struggle when the dataset contains objects of varying shapes or sizes. In contrast, with the fine-tuned Mask R-CNN, the adjustment process enables the model to better comprehend object shapes, contours, and boundaries, resulting in more accurate segmentation masks. These comparison is shown in Table 2.

Table 2. Average Precision for Mask R-CNN and fine-tuned Mask R-CNN models for two types of detection model

|      | Threshold | Mask R-CNN | Fine Tuned Mask R-CNN |
|------|-----------|------------|-----------------------|
| bbox | 0.50:0.95 | 37.5%      | 55.05%                |
|      | 0.5       | 54.6%      | 94.94%                |
|      | 0.75      | 41.9%      | 51.71%                |
| segm | 0.50:0.95 | 34.3%      | 53.82%                |
|      | 0.5       | 52.8%      | 96.94%                |
|      | 0.75      | 37.4%      | 54.99%                |

### 3. Average Recall

#### a. Using Bounding Box (bbox) detection.

Average recall for bounding boxes using Mask R-CNN reflects the extent to which the model detects all objects present in an image. Standard Mask R-CNN often faces limitations in detecting smaller objects or those situated in specific lighting conditions or orientations. In contrast, with Fine-Tuned Mask R-CNN, bounding box average recall tends to improve because the model becomes more sensitive to variations within the target dataset, enabling it to detect objects that might have previously been missed.

#### b. Using Segmentation (segm) detection.

Average recall for segmentation using Mask R-CNN indicates how effectively the model segments all objects in an image. Standard Mask R-CNN may not be optimal for small or complex objects, or for those with contours that are difficult to recognize. However, with Fine-Tuned Mask R-CNN, segmentation average recall improves, demonstrating that the model becomes more responsive to a wider range of object shapes and sizes. This reflects an enhanced capability for instance segmentation. These Average Recall results are presented in Table 3.

Table 3. Average Recall for Mask R-CNN and fine-tuned Mask R-CNN models for two types of detection model

|      | Threshold | Mask R-CNN | Fine Tuned<br>Mask R-CNN |
|------|-----------|------------|--------------------------|
| bbox | 0.50:0.95 | 37.5%      | 71.3%                    |
|      | 0.5       | 54.6%      | 92.9%                    |
|      | 0.75      | 41.9%      | 92.9%                    |
| segm | 0.50:0.95 | 34.3%      | 71.1%                    |
|      | 0.5       | 52.8%      | 92.9%                    |
|      | 0.75      | 37.4%      | 90.7%                    |

### 4. CONCLUSION

This study developed and evaluated a fine-tuned Mask R-CNN model for detecting and segmenting floating objects in river environments using UAV imagery. Experimental results demonstrate that fine-tuning the Mask R-CNN model using a customized dataset substantially improved its performance compared with the default model. The detection accuracy increased from 87% to 92%, while the model achieved an F1-score of 0.88 and a mean Average Precision (mAP) of 0.82, indicating a favorable balance between precision and recall and a reliable capability for object detection and instance segmentation. Improvements in Average Precision (AP) and Average Recall (AR) further confirm the effectiveness of model adaptation to the characteristics of the target river environment.

In addition to improving detection performance, the fine-tuned model reduced the average inference time from 0.5 to 0.3 s per image, demonstrating improved computational efficiency. Although some detection errors remained under challenging conditions, particularly in low-light scenes and during rapid camera movement, the overall results indicate that the fine-tuned model can effectively identify and segment floating objects from UAV imagery. These findings demonstrate the feasibility of applying UAV-based Mask R-CNN for rapid aerial observation of river environments and provide a foundation for its further development as a computer-vision component of search and rescue operations. Future research should focus on evaluating the model under more diverse environmental and flight conditions, increasing the diversity and scale of the training dataset, and validating its performance using real-world SAR scenarios.

## Acknowledgement

The authors would like to express their gratitude to the Directorate of Research and Community Service, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, for the financial support provided for this work. This article is part of the regular fundamental research funded under contract number 173/C3/DT/DT.05.00/PL-BARU/2026.

## References

- [1] Azmeri A, Isa AH. An analysis of physical vulnerability to flash floods in the small mountainous watershed of Aceh Besar Regency, Aceh province, Indonesia. *Jamba*. 2018 Apr 26;10(1):550. doi: 10.4102/jamba.v10i1.550. PMID: 29955274; PMCID: PMC6014066
- [2] Azmeri, Iwan K. Hadihardaja, Rika Vadiya, Identification of flash flood hazard zones in mountainous small watershed of Aceh Besar Regency, Aceh Province, Indonesia, *The Egyptian Journal of Remote Sensing and Space Science*, Volume 19, Issue 1, 2016, Pages 143-160, <https://doi.org/10.1016/j.ejrs.2015.11.001>.
- [3] Badan Nasional Penanggulangan Bencana, Perkembangan Situasi dan Penanganan Bencana di Tanah Air, published at 27 November 2025, Available at <https://bnpb.go.id/berita/perkembangan-situasi-dan-penanganan-bencana-di-tanah-air-tanggal-27-november-2025>
- [4] Badan Nasional Penanggulangan Bencana, Upaya Percepatan Penanganan Darurat, BNPB Berikan Pendampingan Banjir di Aceh, published at 27 November 2025, Available at <https://bnpb.go.id/berita/upaya-percepatan-penanganan-darurat-bnpb-berikan-pendampingan-banjir-di-aceh>
- [5] Cut Azizah, Nuraina Nuraina, Nuraida Usman, Pengenalan Karakteristik Hidrologi Dan Banjir Das Jambo Aye, *RAMBIDEUN Jurnal Pengabdian Kepada Masyarakat*, 4(2), Pages106-111, <https://doi.org/10.51179/pkm.v4i2.583>
- [6] Jen-Yu Han, Chun-Jia Huang, Yuan-Shun Chang, Yen-Cheng Lin, Hao-Che Ho, A novel approach for utilizing UAV imaging and deep learning techniques for river channel measurement and flood simulation, *Geomorphology*, Volume 496, 2026, 110173, <https://doi.org/10.1016/j.geomorph.2026.110173>.
- [7] H. Issaoui, A. ElAdel and M. Zaied, "Object Detection Using Convolutional Neural Networks: A Comprehensive Review," 2024 IEEE 27th International Symposium on Real-Time Distributed Computing (ISORC), Tunis, Tunisia, 2024, pp. 1-6, <https://doi.org/10.1109/ISORC61049.2024.10551342>
- [8] Gebrehiwot A, Hashemi-Beni L, Thompson G, Kordjamshidi P, Langan TE. Deep Convolutional Neural Network for Flood Extent Mapping Using Unmanned Aerial Vehicles Data. *Sensors (Basel)*. 2019 Mar 27;19(7):1486. <https://doi.org/10.3390/s19071486>. PMID: 30934695; PMCID: PMC6479537.
- [9] Bakir ME, Kasimoglu F. An integrated multi-criteria decision-making model for long-term planning of UAVs in disaster management. *PLoS One*. 2026 Jan 6;21(1):e0340303. doi: 10.1371/journal.pone.0340303. PMID: 41493985; PMCID: PMC12774348.
- [10] Šiktar, L., Čaran, B., Šekoranja, B., Švaco, M. (2025). Autonomous UAV Navigation for Search and Rescue Missions Using Computer Vision and Convolutional Neural Networks. In: Jovanović, K., Rodić, A., Raković, M. (eds) *Advances in Service and Industrial Robotics. RAAD 2025. Mechanisms and Machine Science*, vol 190. Springer, Cham. [https://doi.org/10.1007/978-3-032-02106-9\\_73](https://doi.org/10.1007/978-3-032-02106-9_73)
- [11] Sharifah Mastura Syed Mohd Daud, Mohd Yusmialdil Putera Mohd Yusof, Chong Chin Heo, Lay See Khoo, Mansharan Kaur Chainchel Singh, Mohd Shah Mahmood, Hapizah Nawawi, Applications of drone in disaster management: A scoping review, *Science & Justice*, Volume 62, Issue 1, 2022, Pages 30-42, <https://doi.org/10.1016/j.scijus.2021.11.002>.
- [12] Kareem AA, Abid AJ, Hammood DA, Alubady R, Yaqoob SJ, Almomani O, Rubanenko O. A bio-inspired swarm UAV framework integrating thermal sensing and optimization-based coordination for efficient search and rescue operations. *Sci Rep*. 2025 Dec 27;16(1):3361. doi: 10.1038/s41598-025-33223-z. PMID: 41454066; PMCID: PMC12835151.
- [13] Zhang L, Zhang Y, Zhang Z, Shen J, Wang H. Real-Time Water Surface Object Detection Based on Improved Faster R-CNN. *Sensors*. 2019; 19(16):3523. <https://doi.org/10.3390/s19163523>

- [14] Shah, S., Tembhurne, J. Object detection using convolutional neural networks and transformer-based models: a review. *Journal of Electrical Systems and Inf Technol* 10, 54 (2023). <https://doi.org/10.1186/s43067-023-00123-z>
- [15] Deluxni, N., Sudhakaran, P., Alroobaea, R. et al. Adaptive lightweight mask R-CNN model for underwater debris instance segmentation and classification towards sustainable marine waste management. *Sci Rep* 16, 14057 (2026). <https://doi.org/10.1038/s41598-026-44542-0>
- [16] Di, J., Xi, K. & Yang, Y. An enhanced YOLOv8 model for accurate detection of solid floating waste. *Sci Rep* 15, 25015 (2025). <https://doi.org/10.1038/s41598-025-10163-2>
- [17] Zhang L, Xie Z, Xu M, Zhang Y, Wang G. EYOLOv3: An Efficient Real-Time Detection Model for Floating Object on River. *Applied Sciences*. 2023; 13(4):2303. <https://doi.org/10.3390/app13042303>
- [18] Haitham A. Mahmoud, Ahmed T. Soliman, Mohammed El-Meligy, Azhar Imran, Abdelaty Edrees Abdelgawad, Advanced marine data analytics and underwater object detection using mask R-CNN for ecosystem monitoring, *Sustainable Computing: Informatics and Systems*, Volume 50, 2026, 101342.



**Prima Kristalina** holds a PhD in electrical engineering from Institut Teknologi Sepuluh Nopember, Indonesia. Currently, she is an Associate Professor at the Department of Electrical Engineering, Politeknik Elektronika Negeri Surabaya, Indonesia. Her research interests include localization, object tracking and positioning, artificial intelligence for disaster management system, and mobile crowd sensing. She also serves as Editor in Chief for *Emitter International Journal of Engineering Technology*. In addition, she is a reviewer for reputable journals such as *IEEE Access*, *Ad Hoc Networks*, *Computers and Electrical Engineering*, *Engineering Application of AI*, *Pervasive and Mobile Computing*, *Progress in Disaster Science*, etc. She can be contacted at email: [prima@pens.ac.id](mailto:prima@pens.ac.id).



**Aries Pratiarso** holds a Magister Engineer in electrical engineering from Institut Teknologi Sepuluh Nopember, Indonesia. He is a lecturer at the Department of Electrical Engineering, Politeknik Elektronika Negeri Surabaya, Indonesia. His research interests include data coding, mobile communication, and artificial intelligence and its applications. He can be contacted at email: [aries@pens.ac.id](mailto:aries@pens.ac.id).



**A. Fauzi Makarim** holds a Magister Engineer in electrical engineering from Politeknik Elektronika Negeri Surabaya, Indonesia. Currently, he is a lecturer at the Department of Electrical Engineering, Politeknik Elektronika Negeri Surabaya, Indonesia. His research interests include cloud computing, artificial intelligence and its applications, and data mining. He can be contacted at email: [fauzi@pens.ac.id](mailto:fauzi@pens.ac.id).



**Munawar** holds a Magister Engineer in electrical engineering from Universitas Syiah Kuala, Indonesia. Currently, he is a lecturer at the Department of Electrical Engineering, Politeknik Negeri Lhokseumawe, Indonesia. His research interests include communication systems and data communication.

