SAFE-Net: Scale-Aware Feature Enhancement for Aerial Person Detection in Flood Disaster Imagery

Anonymous CVPR submission

Paper ID 5

Abstract

Detecting stranded persons from Unmanned Aerial Vehicles (UAVs) is crucial for autonomous search and rescue operations during flood disasters. However, reliable detection in aerial flood imagery remains challenging due to extreme scale variations, strong water reflections and motion blur. In many cases, human targets occupy less than 0.1% *of the image area, making them difficult to detect using conventional object detectors. To address these challenges, we propose SAFE-Net (Scale-Aware Feature Enhancement Network), a lightweight detection framework which builds upon the YOLOv8 architecture by replacing standard C2f blocks with a Scale-Aware Feature Enhancement (SAFE) module. The SAFE module improves the representation of tiny human targets through two mechanisms: scale-aware spatial weighting to emphasize extremely small objects, and texture enhancement using depthwise convolutions to recover fine edge information while suppressing noise from water surfaces. We also introduce UAV-based Flood Survivor Detection Dataset (UAV-SURV), a dataset of 6,122 annotated aerial images collected from real flood monitoring videos. Experimental results show that SAFE-Net improves detection accuracy by 10.9% in mAP@0.5:0.95 over YOLOv8n while reducing model parameters by 33.8%, demonstrating an effective and lightweight solution for UAV-assisted disaster response.*

1. Introduction

Flood disasters are among the most destructive natural hazards worldwide, often leaving individuals stranded in inaccessible urban and rural regions. Rapid identification of survivors is critical during search-and-rescue operations, where early detection significantly improves survival rates. In recent years, Unmanned Aerial Vehicles (UAVs) have emerged as effective tools for disaster response [1, 3, 14], providing rapid aerial reconnaissance and real-time situational awareness over large flooded areas. However, manually inspecting large volumes of drone footage is time-consuming, cognitively demanding, and prone to human error. Automated computer vision systems capable of detecting stranded individuals from aerial imagery can therefore play a vital role in accelerating rescue operations.

Figure 1. Examples of aerial flood scenes captured by UAVs during floods. The images illustrate the complexity of flooded environments, including water reflections, partially submerged structures, debris, and extremely small human instances.

Detecting stranded persons in aerial flood imagery remains challenging due to the complex visual characteristics of flooded environments. As shown in Figure 1, flood scenes often contain strong reflections, ripples, glare, shadows, and floating debris that introduce severe visual noise. In addition, due to the altitude of UAV platforms, human targets frequently appear extremely small, sometimes occupying less than 0.1% of the total image area. These conditions significantly degrade the performance of conventional object detectors, which are typically trained on ground-level datasets containing larger and clearer objects [2, 4, 6, 9].

Recent deep learning–based object detectors [4, 12], particularly single-stage architectures such as YOLO, have demonstrated strong performance in real-time detection tasks. However, these models still struggle in flood scenarios due to extreme scale variations, low contrast, and

complex background clutter. Moreover, publicly available datasets rarely capture the unique visual conditions present during real flood disasters, limiting the ability of detection models to generalize effectively in such environments. These challenges highlight the need for specialized datasets and detection architectures capable of handling extremely small targets and suppressing water-induced visual noise in aerial flood imagery.

To address these challenges, we introduce *UAV-SURV*, a dataset consisting of 6,122 annotated aerial images extracted from real flood monitoring videos. The dataset captures diverse flood conditions, including variations in illumination, water patterns, camera altitude, and object scale. Each image is annotated with bounding boxes for human instances and divided into training and testing subsets. Furthermore, we propose SAFE-Net (Scale-Aware Feature Enhancement Network), a modified YOLOv8 architecture designed to improve aerial person detection in flood disaster scenes. SAFE-Net replaces the standard C2f modules in the YOLOv8 backbone with a Scale-Aware Feature Enhancement (SAFE) module. The SAFE module enhances feature representation through scale-aware spatial weighting and texture enhancement mechanisms that emphasize weak human features while suppressing noise caused by water reflections and clutter.

The main contributions of this work are summarized as follows:

- We introduce *UAV-SURV*, a dataset of 6,122 annotated aerial images extracted from real flood monitoring videos, designed to support research on human detection in flood disaster environments.
- We propose SAFE-Net, a modified YOLOv8 architecture that incorporates a Scale-Aware Feature Enhancement module to improve feature representation for extremely small human instances.
- The proposed SAFE module enhances detection performance through scale-aware spatial weighting and texture enhancement mechanisms that suppress water-induced visual noise.
- Experimental results show that SAFE-Net improves detection accuracy by 10.9% in mAP@0.5:0.95 compared to YOLOv8n while reducing model parameters from 3.2M to 2.12M.

2. Related Work

Unmanned Aerial Vehicles (UAVs) have become increasingly important for disaster monitoring and emergency response due to their ability to capture high-resolution imagery from difficult-to-access regions. UAV-based vision systems enable rapid assessment of affected areas and assist in locating stranded individuals during search-and-rescue operations. Several works have explored deep learning approaches for analyzing UAV imagery in disaster scenarios. Ijaz et al. [5] proposed a UAV-assisted edge computing framework that deploys compressed convolutional neural networks on embedded devices for real-time disaster classification. Similarly, Bashir et al. [1] introduced an efficient CNN-based model for disaster event classification using UAV imagery, demonstrating that deep learning models can effectively analyze aerial scenes captured during emergencies. Although these works demonstrate the usefulness of UAV imagery for disaster analysis, they primarily focus on scene-level classification rather than detecting individual victims in complex environments.

The performance of deep learning models for disaster analysis largely depends on the availability of high-quality annotated datasets. Several datasets have been introduced to support flood monitoring and disaster scene understanding. FloodNet [8] provides high-resolution UAV imagery collected after Hurricane Harvey and supports multiple computer vision tasks such as image classification, semantic segmentation, and visual question answering. While FloodNet enables comprehensive flood scene understanding, it does not specifically address the detection of people in flood environments. More recently, the DeepFlood dataset [3] was introduced to support accurate flood mapping and segmentation using high-resolution aerial imagery with detailed annotations of inundated vegetation and flood-affected areas. Similarly, UrbanSARFloods [14] provides a benchmark dataset for large-scale flood mapping using Sentinel-1 SAR imagery, enabling research on flood detection across urban and open-area environments. Despite these advances, existing flood datasets mainly focus on flood extent estimation and segmentation tasks rather than identifying stranded individuals in aerial flood imagery. Beyond disaster-specific datasets, several large-scale UAV perception datasets have been proposed to support aerial scene understanding. For example, UAVScenes [13] introduces a multi-modal UAV dataset that includes synchronized camera images and LiDAR data with semantic annotations to enable tasks such as segmentation, depth estimation, and localization. While such datasets advance UAV perception research, they are primarily designed for general aerial scene understanding and do not focus on disaster response or victim detection scenarios.

Remote sensing imagery presents several challenges compared to conventional computer vision datasets. Objects often appear at very small scales due to high imaging altitude, and scenes contain complex background structures with clutter, occlusions, and illumination variations. In addition, obtaining pixel-level annotations for remote sensing images is time-consuming and expensive. To address limited labeled data, Lu et al. [7] proposed an uncertainty-aware semi-supervised learning framework that improves segmentation performance by leveraging both labeled and unlabeled remote sensing data. Despite these advances, de-

tecting extremely small objects such as humans in aerial flood imagery remains challenging due to scale variations, water reflections, and background clutter.

In contrast to existing works, this paper focuses specifically on detecting stranded individuals in flood scenarios using aerial imagery. We introduce the UAV-SURV dataset together with SAFE-Net, a scale-aware feature enhancement network designed to improve the detection of extremely small human instances in complex flood environments.

3. UAV-SURV Dataset Curation and Statistics

Reliable detection of stranded persons in aerial flood imagery requires datasets that capture realistic flood conditions. However, publicly available datasets rarely contain aerial imagery recorded during real flood disasters. To address this limitation, we construct a dataset called *UAV-SURV*, specifically designed for aerial person detection in flood environments.

3.1. Dataset Curation

Constructing a dataset for aerial person detection in flood environments is challenging due to the extremely small size of human instances, dynamic water surfaces, and the high cost of manually annotating long UAV video sequences. Similar challenges have been observed in aerial small-object detection datasets such as RealDroneVision [11], which highlights the importance of specialized datasets and semi-automatic labeling strategies. To efficiently generate reliable annotations, we adopt a semi-automatic dataset creation pipeline inspired by the Self-Annotated Labeling from Videos (SA-LfV) framework [10, 11], combined with a human-in-the-loop verification process.

The UAV-SURV dataset is derived from drone videos recorded during real flood events in Amaravathi, India. As shown in Figure 1 these videos capture large flooded regions with complex visual conditions including water reflections, ripples, floating debris, and varying illumination. Such environmental factors significantly increase the difficulty of detecting stranded individuals from aerial viewpoints.

The dataset construction process follows four stages:

1. Initial tracking-based proposals: Raw UAV videos were first processed using an object tracker to generate candidate bounding boxes across frames. These tracker outputs provided an initial set of pseudo-labels representing potential human instances. Although efficient, the automatically generated proposals often contained false positives caused by background clutter or reflections and occasionally missed true human instances.
2. Human verification of tracker outputs: Human annotators reviewed the tracker-generated proposals to remove false positives and identify missed human instances. This step produced a cleaner initial annotation set while significantly reducing the manual effort compared to exhaustive frame-by-frame labeling.
3. False negative recovery via detector training: To further improve annotation coverage, a detection model was trained using the refined tracker annotations. The trained detector was then applied to the original video sequences to identify additional candidate detections that were previously missed.
4. Iterative human-in-the-loop refinement: The new detections generated by the trained model were again inspected by human annotators. Remaining false positives were removed and additional missed instances were corrected. This iterative refinement process gradually improved the annotation quality while minimizing manual labeling effort.

Using this semi-automatic pipeline, we generated a high-quality dataset with reliable bounding box annotations while significantly reducing annotation cost. The final UAV-SURV dataset contains 6,122 annotated aerial images representing diverse flood scenarios suitable for training and evaluating aerial person detection models.

Table 1. Statistical analysis of object area ratio and object aspect ratio in the dataset.

Split	Object Area Ratio			Object Aspect Ratio		
	Min	Mean	Max	Min	Mean	Max
Train	0.000042	0.000967	0.091016	0.023	0.479	14.063
Test	0.000039	0.000997	0.079982	0.062	0.478	7.031

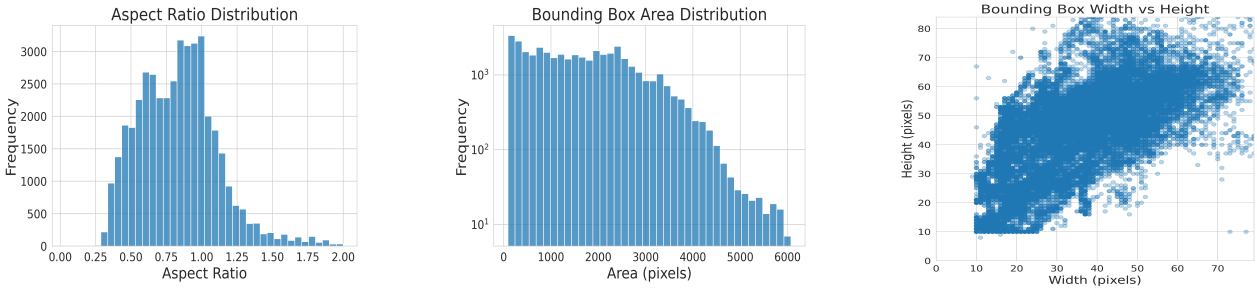
3.2. Dataset Statistics

The final UAV-SURV dataset contains a total of 6,122 annotated images. The dataset is divided into training and testing subsets to enable reliable model evaluation, with 4,898 images used for training and 1,224 images used for testing.

A key characteristic of aerial flood imagery is the extremely small size of human instances. Due to the altitude of the UAV, individuals often occupy only a very small fraction of the image area. Analysis of the bounding box statistics shows that most human instances occupy less than 0.1% of the image area, making the detection task particularly challenging for conventional object detectors. Representative samples from the UAV-SURV dataset are shown in Figure 4.

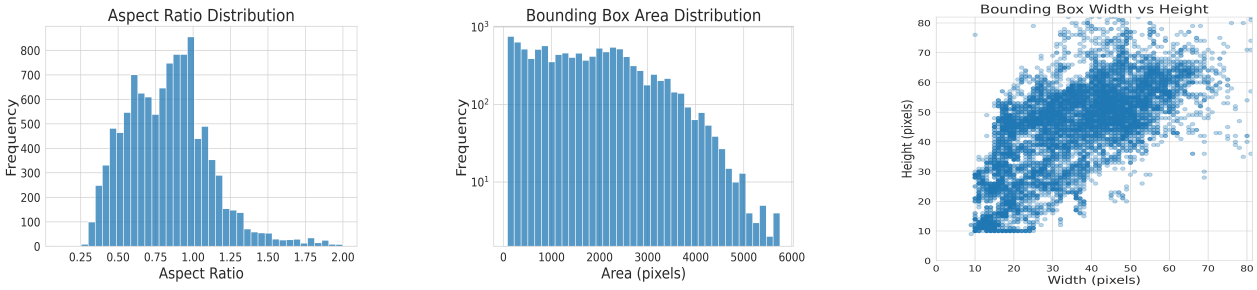
To further analyze the dataset characteristics, we examine several statistical properties of the annotations, including aspect ratio distribution, bounding box area distribution, the number of boxes per image, and the spatial distribution of human instances. Figures 2 and 3 illustrate these statistics for the training and testing subsets, and detailed statistics are provided in Table 1.

The aspect ratio distribution shows that most human bounding boxes fall within a narrow range, reflecting the



(a) Aspect Ratio Distribution of Bounding Boxes (b) Bounding Box Area Distribution (c) Width vs Height of Bounding Boxes

Figure 2. Statistical analysis of bounding box characteristics in the training dataset.



(a) Aspect Ratio Distribution of Bounding Boxes (b) Bounding Box Area Distribution (c) Width vs Height of Bounding Boxes

Figure 3. Statistical analysis of bounding box characteristics in the testing dataset.

Figure 4. Samples from the UAV-SURV dataset. The images illustrate aerial flood environments where stranded individuals appear at extremely small scales and diverse spatial locations. The scenes exhibit challenging visual conditions including cluttered backgrounds, water reflections, varying illumination, and complex urban layouts.

typical upright structure of human figures observed from aerial viewpoints. The bounding box area distribution indicates that the majority of human instances correspond to extremely small objects in the image. The distribution of boxes per image shows variability in the number of persons present in different frames, ranging from sparse scenes with only one or two individuals to frames containing multiple persons. These statistics confirm that UAV-SURV presents a challenging dataset characterized by small objects, spatial sparsity, and complex environmental conditions.

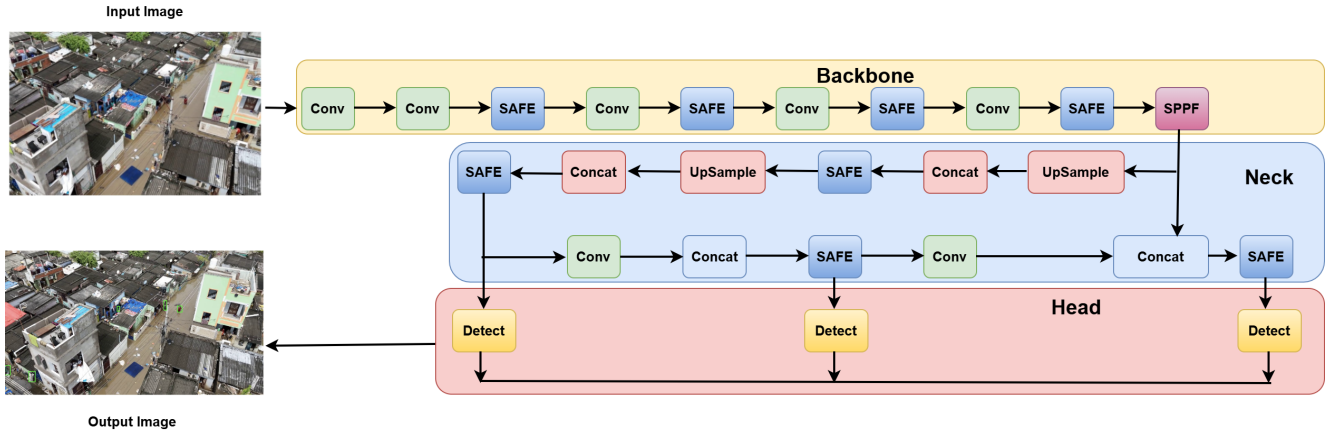


Figure 5. Overview of the proposed SAFE-Net architecture. SAFE-Net follows the YOLOv8 pipeline with backbone, neck, and detection head, where the standard C2f modules are replaced with SAFE blocks to enhance feature representations of small human instances in aerial flood imagery.

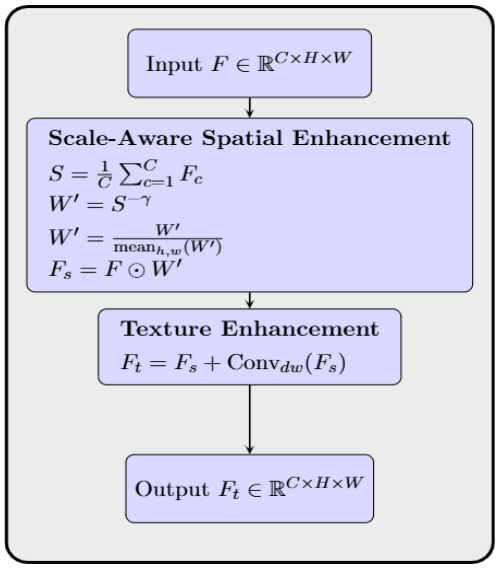


Figure 6. Overview of the proposed SAFE module used in SAFE-Net. The module refines backbone feature maps through scale-aware spatial enhancement and texture enhancement to improve the detection of small human instances in aerial flood imagery.

Overall, UAV-SURV provides a realistic benchmark for aerial person detection in flood disaster environments and supports the evaluation of detection models designed for UAV-assisted search-and-rescue applications. The overall SAFE-Net architecture is illustrated in Figure 5. The overall SAFE-Net architecture is illustrated in Figure 5.

4. Proposed Methodology

This section presents SAFE-Net, a modified YOLOv8 architecture designed for aerial person detection in flood disaster environments. Detecting humans in aerial flood imagery is challenging due to extremely small object sizes, water reflections, and cluttered backgrounds. To address these challenges, SAFE-Net enhances feature representations using scale-aware spatial enhancement and texture enhancement mechanisms integrated into the backbone network. Specifically, the standard C2f modules in YOLOv8 are replaced with a lightweight SAFE module that strengthens the representation of small human instances while maintaining computational efficiency.

4.1. SAFE-Net Architecture

SAFE-Net follows the general YOLOv8 architecture consisting of a backbone for feature extraction, a neck for multi-scale feature aggregation, and a detection head for bounding box prediction. The backbone extracts hierarchical feature representations from the input image, while the neck aggregates features across multiple scales to improve small object detection.

In the proposed framework, the conventional C2f modules in the YOLOv8 backbone are replaced with SAFE blocks that integrate scale-aware spatial enhancement and texture enhancement operations. The overall structure of the SAFE module is illustrated in Figure 6.

Given an input image I, the backbone extracts feature maps

$$F \in \mathbb{R}^{C\times H\times W} \tag{1}$$

where C, H, and W denote the number of channels, height, and width of the feature map. These feature maps are subsequently refined by the SAFE modules before being forwarded to the detection head.

4.2. Scale-Aware Spatial Enhancement

In aerial flood imagery, humans often occupy only a small fraction of the image area, leading to weak feature responses that may be suppressed during feature aggregation. To address this issue, SAFE-Net introduces a scale-aware spatial enhancement mechanism that emphasizes spatial regions corresponding to small objects.

Given the feature map F, a spatial importance map is computed by averaging the feature responses across channels:

$$S = \frac{1}{C}\sum_{c=1}^{C} F_c \tag{2}$$

where $F_c \in \mathbb{R}^{H\times W}$ denotes the c-th channel of the feature map. The resulting spatial map $S \in \mathbb{R}^{1\times H\times W}$ captures the average activation across channels.

A scale-aware weighting map is then computed as

$$W^{'} = S^{-\gamma} \tag{3}$$

where γ controls the strength of the enhancement. To stabilize training, the weight map is normalized as

$$W^{'} = \frac{W^{'}}{\text{mean}_{h,w}(W^{'})}. \tag{4}$$

The normalized weight map is broadcast along the channel dimension and applied to the feature map through element-wise multiplication:

$$F_s = F \odot W^{'} \tag{5}$$

where $\odot$ denotes element-wise multiplication. This operation amplifies weak spatial responses associated with small-object regions while suppressing dominant background responses.

4.3. Texture Enhancement

Flood scenes often contain low-contrast human silhouettes that blend with surrounding water surfaces. To enhance local structural information, SAFE-Net incorporates a lightweight texture enhancement operation based on depthwise convolution.

The enhanced feature representation is computed as

$$F_t = F_s + \text{Conv}_{dw}(F_s) \tag{6}$$

where Conv_{dw} denotes a depthwise convolution applied independently to each channel. This residual operation enhances edge and structural information, enabling the network to better distinguish human shapes from water patterns and background clutter.

The resulting feature map F_t is forwarded to the YOLOv8 neck and detection head for final object prediction.

4.4. Integration with YOLOv8

The SAFE module is integrated into the YOLOv8 backbone by replacing the standard C2f blocks with the proposed feature refinement block. This modification allows the detector to better capture small human features in aerial flood imagery without significantly increasing model complexity. The refined feature maps are then passed through the neck and detection head for multi-scale object prediction.

By combining scale-aware spatial enhancement and texture refinement, SAFE-Net improves the detection of small human instances in challenging flood environments while maintaining the efficiency required for UAV-based search-and-rescue applications.

5. Implementation Details

The proposed SAFE-Net model is implemented using the Ultralytics YOLOv8 framework in PyTorch. The standard C2f modules in the YOLOv8 backbone are replaced with the proposed SAFE module, which incorporates scale-aware spatial enhancement and texture enhancement to improve feature representation for small human instances in aerial flood imagery.

5.1. Training Setup

The model is trained on the UAV-SURV dataset, which consists of 6,122 annotated aerial images collected from real flood monitoring videos. The dataset is divided into 4,898 images for training and 1,224 images for testing. Each image contains bounding box annotations corresponding to the person class.

All experiments are conducted using an NVIDIA RTX 2080 Ti GPU. The input image resolution is set to 640×640, and the model is trained for 100 epochs with a batch size of 16. The training procedure follows the standard YOLOv8 optimization pipeline with stochastic gradient descent (SGD) optimizer, momentum of 0.937, and weight decay of 5×10^{-4}. Data augmentation techniques including mosaic augmentation, random scaling, flipping, and color jitter are applied during training to improve model generalization.

5.2. Evaluation Metrics

Model performance is evaluated using standard object detection metrics including precision, recall, mean Average Precision at IoU threshold 0.5 (mAP@0.5), and mean Average Precision across multiple IoU thresholds (mAP@0.5:0.95). These metrics provide a comprehensive assessment of detection accuracy and localization performance.

5.3. Implementation Efficiency

SAFE-Net maintains a lightweight architecture suitable for aerial deployment scenarios. The final model contains ap-

Table 2. Comparison of lightweight detectors on the UAV-SURV dataset. SAFE-Net achieves the best detection performance while maintaining a lightweight architecture.

Model	Params (M)	FLOPs (G)	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5n	2.6	7.7	0.963	0.807	0.880	0.710
YOLOv8n	3.2	8.7	0.967	0.813	0.889	0.725
YOLOv10n	2.3	6.7	0.962	0.804	0.889	0.740
RT-DETR	20	60	0.953	0.79	0.82	0.723
SAFE-Net (Ours)	2.12	5.9	**0.980**	**0.895**	**0.950**	**0.834**

Table 3. Ablation study of SAFE-Net components on UAV-SURV.

Method	Scale	Texture	Params (M)	FLOPs (G)	Precision	Recall	mAP@0.5:0.95
YOLOv8n + Scale Enhancement	✓	–	2.06	5.6	0.972	0.872	0.778
YOLOv8n + Texture Enhancement	–	✓	2.08	5.7	0.973	0.892	0.821
SAFE-Net (Scale + Texture)	✓	✓	2.12	5.9	**0.980**	**0.895**	**0.834**

proximately 2.12 million parameters and requires about 5.9 GFLOPs. During inference, the model achieves an average inference time of approximately 3 ms per image on an RTX 2080 Ti GPU, enabling efficient processing of aerial imagery for disaster response applications.

6. Experiments and Results

This section evaluates the performance of the proposed SAFE-Net architecture on the UAV-SURV dataset and compares it with several lightweight YOLO-based detectors. We also conduct an ablation study to analyze the contribution of each component in the proposed SAFE module.

6.1. Comparison with YOLO Baselines

To demonstrate the effectiveness of SAFE-Net, we compare it with three lightweight object detection models: YOLOv5n, YOLOv8n, and YOLOv10n. All models are trained using the same dataset split and training configuration described in Section 5. Performance is evaluated using precision, recall, and mean Average Precision at IoU threshold 0.5 (mAP@0.5).

Table 2 summarizes the detection performance of different models on the UAV-SURV testing set. Qualitative detection examples are shown later in Figure 8.

The results show that SAFE-Net consistently outperforms all baseline detectors. In particular, SAFE-Net achieves a 10.9-point improvement in mAP@0.5:0.95 compared to YOLOv8n while maintaining a lightweight architecture. This improvement demonstrates the effectiveness of the proposed SAFE module in enhancing feature representations for extremely small human instances in aerial flood imagery, as visualized in Figure 7.

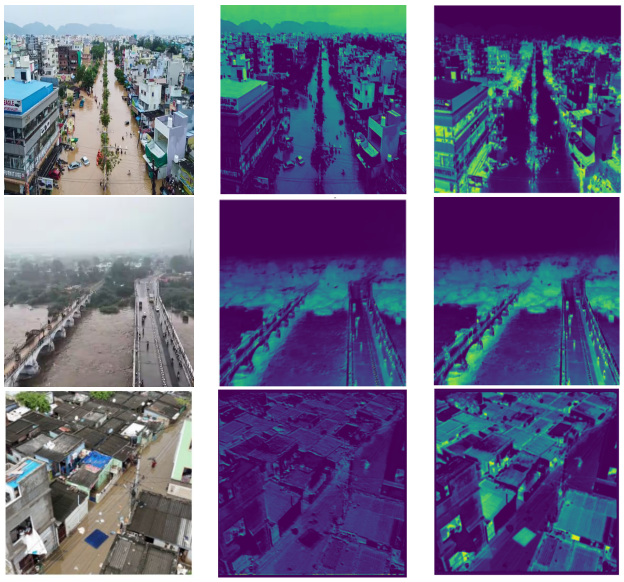


Figure 7. Visualization of feature responses produced by YOLOv8 and SAFE-Net. Each row shows the original image (left), YOLOv8n activation map (middle), and SAFE-Net activation map (right). The SAFE module produces more focused responses around human regions while suppressing background noise.

6.2. Ablation Study

To analyze the contribution of the implemented SAFE components, we perform an ablation study by evaluating scale-aware enhancement and texture enhancement. All ablation experiments are conducted using the YOLOv8 backbone while keeping the training settings unchanged.

Table 3 presents the ablation results.

The ablation results show that both implemented SAFE components contribute to improved detection performance. Scale-aware enhancement improves the detection of ex-

Figure 8. Visual results on the UAV-SURV dataset using SAFE-Net model.

tremely small human instances by amplifying weak spatial responses. Texture enhancement further improves boundary information, which helps distinguish human shapes from complex water textures.

SAFE-Net achieves the highest detection performance while maintaining a lightweight architecture. As shown in Table 2, the proposed model improves mAP@0.5 by 6.1 points and mAP@0.5:0.95 by 10.9 points compared to YOLOv8n while keeping low parameter count and computational complexity.

7. Conclusion

This work presents UAV-SURV, a real flood-disaster aerial dataset, and SAFE-Net, a lightweight YOLOv8-based detector with scale-aware and texture enhancement modules. Experiments show consistent gains over lightweight baselines, particularly for tiny-person detection, while preserving efficient inference suitable for UAV-assisted search-and-rescue deployment.

References

[1] Munzir Hubiba Bashir, Musheer Ahmad, Danish Raza Rizvi, and Ahmed A Abd El-Latif. Efficient cnn-based disaster events classification using uav-aided images for emergency response application. *Neural Computing and Applications*, 36(18):10599–10612, 2024. 1, 2

[2] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 1

[3] Mulham Fawakherji, Jeffrey Blay, Matilda Anokye, Leila Hashemi-Beni, and Jennifer Dorton. Deepflood for inundated vegetation high-resolution dataset for accurate flood mapping and segmentation. *Scientific Data*, 12(1):271, 2025. 1, 2

[4] Muhammad Hussain. Yolov1 to v8: Unveiling each variant–a comprehensive review of yolo. *IEEE access*, 12:42816–42833, 2024. 1

[5] Haris Ijaz, Rizwan Ahmad, Rehan Ahmed, Waqas Ahmed, Yan Kai, and Wu Jun. A uav-assisted edge framework for real-time disaster management. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 2

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[7] Xiaoqiang Lu, Lingling Li, Licheng Jiao, Xu Liu, Fang Liu, Wenping Ma, and Shuyuan Yang. Uncertainty-aware semi-supervised learning segmentation for remote sensing images. *IEEE Transactions on Multimedia*, 2025. 2

[8] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 2

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1

[10] Arun Kumar Sivapuram, Prashanth Komuravelli, and Rama Krishna Sai Gorthi. Sa-lfv: self-annotated labeling from videos for object detection. *Machine Learning*, 114(1):21, 2025. 3

[11] Arun Kumar Sivapuram, Pranav RT Peddinti, Harish Puppala, Komuravelli Prashanth, Jaladi Sri Harsha, and Rama Krishna Sai Gorthi. Realdronevision: Dataset and architecture advancements for small-object drone detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6687–6695, 2026. 3

[12] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE transactions on pattern analysis and machine intelligence*, 44(4): 1922–1933, 2020. 1

[13] Sijie Wang, Siqi Li, Yawei Zhang, Shangshu Yu, Shenghai Yuan, Rui She, Quanjiang Guo, JinXuan Zheng, Ong Kang Howe, Leonrich Chandra, et al. Uavscenes: A multi-modal dataset for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28946–28958, 2025. 2

[14] Jie Zhao, Zhitong Xiong, and Xiao Xiang Zhu. Urbansarfloods: Sentinel-1 slc-based benchmark dataset for urban and open-area flood mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 419–429, 2024. 1, 2