

HYBRID YOLO CNN ARTIFICIAL INTELLIGENCE FRAMEWORK FOR MULTI-CLASS DISASTER IMAGE ANALYTICS IN LOCAL GOVERNMENT EMERGENCY RESPONSE AND HEALTHCARE MANAGEMENT

Dr.P.Tharcis^{1*}, Dr.S.Karthik², Dr.G.Nirmala³, R Arul Arumugam⁴

^{1*}Associate Professor, Department of Electronics and Communication Engineering, SRM Madurai College for Engineering and Technology, Pottapalayam.

²Assistant Professor, Department of Electronics and Communication Engineering, Mahendra Institute of Technology(Automomous), Namakkal.

³Associate Professor, Department of Electronics and Communication Engineering, Mahendra Institute of Technology(Automomous), Namakkal.

⁴Assistant Professor, Department of Electronics and Communication Engineering, Annapoorna Engineering College(Automomous), Salem.
Email id: ^{1*}drtharcis@gmail.com

Abstract—Disasters such as floods, wildfires, structural collapses, and mass-casualty incidents impose severe operational burdens on local government authorities and emergency healthcare systems. Existing computer vision frameworks address these challenges in isolation, lacking integrated multi-task capacity to simultaneously detect disaster types, locate victims, classify medical severity, and assess infrastructure damage. This article proposes a Hybrid YOLO-CNN Artificial Intelligence Framework (HYCAIF) that unifies YOLOv8-based object detection with a Convolutional Neural Network (CNN) feature backbone, Squeeze-and-Excitation (SE) attention, Convolutional Block Attention Module (CBAM), and a Feature Pyramid Network (FPN) neck into a cohesive multi-head architecture. The framework is designed explicitly to serve local government disaster response functions, including GIS-integrated emergency alerting, medical triage dispatch, and administrative dashboarding, thereby bridging an underexplored gap between sub-national governance and deep learning. Experiments on a composite dataset comprising the Aerial Image Dataset for Emergency Response (AIDER) and the Detection in Optical Remote Sensing Images (DOTA) benchmark yield a mean Average Precision at IoU 0.5 (mAP@0.5) of 91.7%, Precision of 90.2%, Recall of 89.6%, and F1-Score of 89.9%, representing improvements of 9.6, 9.7, 10.3, and 10.6 percentage points over the best-performing baseline, respectively. The model achieves a sustained inference throughput of 71 frames per second on commodity GPU hardware, satisfying real-time deployment constraints. These results affirm the practical viability of the proposed framework for adoption by municipal authorities, civil protection agencies, and emergency healthcare dispatchers.

Keywords—Disaster response; deep learning; YOLO; convolutional neural network; emergency healthcare; local government; image analytics; object detection; feature pyramid network; multi-class classification.

I. INTRODUCTION

Natural and anthropogenic disasters constitute one of the most consequential governance challenges confronting local authorities worldwide. Floods inundate residential zones with minimal warning; wildfires consume vegetation and built environments within hours; structural collapses trap occupants beneath rubble; and mass-casualty events simultaneously overwhelm hospitals, transportation networks, and administrative coordination systems. For sub-national governments—the institutions constituting the primary first-responder tier—the ability to perceive, classify, and react to these events within operationally meaningful time windows is a function of both institutional preparedness and the quality of supporting information technology.

Conventional disaster monitoring relies on human interpretation of aerial imagery, satellite captures, and closed-circuit television streams. Such workflows are temporally constrained, prone to observer fatigue, and incapable of scaling to the density of visual data generated by modern unmanned aerial vehicles (UAVs) or satellite constellations. The integration of deep learning object detection architectures into emergency response pipelines represents a structural solution to this bottleneck. Among available paradigms, the You Only Look Once (YOLO) family of detectors has emerged as the dominant real-time framework owing to its single-stage inference strategy and competitive accuracy across diverse object categories [Ali and Zhang, 2024]. Complementarily, Convolutional Neural Networks (CNNs) offer hierarchical feature extraction capacity that generalises across imaging modalities including multispectral satellite imagery, thermal UAV footage, and RGB surveillance video [Krichen, 2023; Li et al., 2020].

Despite this technical progress, a critical deployment gap persists. Existing deep learning frameworks for disaster detection are overwhelmingly designed as research prototypes without explicit integration pathways into the operational workflows of local government emergency services or healthcare dispatch systems. The research literature has addressed wildfire and smoke detection [Bahhar et al., 2023], flood identification via UAV imagery [Munawar et al., 2021], victim recognition in search-and-rescue scenarios [Ulloa et al., 2021], and post-earthquake building damage assessment [Zou et al., 2024] as separate streams, without unification into a cohesive multi-task municipal intelligence platform. Furthermore, the governance dimension—whereby detected events must trigger legally and administratively appropriate municipal responses including emergency medical dispatch, public communication, and infrastructure closure orders—has received negligible attention in the computer vision literature.

This article addresses that gap directly. The proposed Hybrid YOLO-CNN Artificial Intelligence Framework (HYCAIF) is a multi-head detection system operating on aerial and satellite disaster imagery and designed to serve the operational requirements of local governments. The framework simultaneously performs four detection tasks: disaster type classification (fire, flood, structural collapse), victim localization, medical severity triaging, and infrastructure damage grading. A multi-scale Feature Pyramid Network neck aggregates features across resolutions while Squeeze-and-Excitation and CBAM attention modules prioritise semantically relevant spatial regions. Outputs feed into an integrated action layer encompassing GIS-based emergency mapping, SMS-based alert dissemination, ambulance dispatch optimisation, and an administrative dashboard compliant with public sector operational workflows.

The principal contributions of this work are:

- (i) A unified multi-head hybrid architecture that combines YOLOv8 detection, CNN feature extraction, attention gating, and FPN-based multi-scale fusion for simultaneous disaster, victim, triage, and damage detection;
- (ii) An actionable output layer aligned with local government emergency service workflows, producing GIS-mapped alerts, healthcare dispatch instructions, and administrative notifications;
- (iii) Comprehensive empirical validation on the AIDER and DOTA composite benchmark yielding state-of-the-art performance across five quantitative metrics;
- (iv) A system requirements analysis bridging the technical deployment needs of the framework with sub-national government ICT infrastructure constraints.

The remainder of this article proceeds as follows. Section II surveys the related literature. Section III details the dataset. Section IV formalises the system model and proposed methodology. Section V presents experimental results and comparative analysis. Section VI offers conclusions and future directions.

II. LITERATURE REVIEW

A. Convolutional Neural Networks for Image Classification and Detection

CNNs have become the foundational computational substrate for visual understanding tasks since their resurgence in the early 2010s. Their capacity to learn shift-invariant hierarchical representations through successive convolutional, pooling, and nonlinear activation layers has been exhaustively characterised in the survey literature [Krichen, 2023; Li et al., 2020; Taye, 2023]. Applications span medical imaging [Iqbal

et al., 2023; Li et al., 2023; Mienye et al., 2025; Mohammed et al., 2024], activity recognition [Raj and Kos, 2023], and IoT edge deployment [Naveen and Kounte, 2024]. In the remote sensing domain, CNNs have demonstrated robust performance for object recognition in aerial imagery as early as 2017 [Radovic et al., 2017], progressing through ship detection in SAR data [Zhang and Zhang, 2019] to contemporary multi-class satellite object classification [R and R, 2024]. Fault-tolerant CNN variants have also been proposed for safety-critical deployments [Zhao et al., 2020], a property of relevance to disaster response systems where model reliability is non-negotiable.

B. YOLO-Based Object Detection Architectures

The YOLO framework has undergone substantial architectural evolution since its inception, culminating in models such as YOLOv8 and the recently proposed Hyper-YOLO, which incorporates hypergraph computation for richer relational feature modelling [Feng et al., 2024]. Systematic reviews of YOLO evolution confirm its dominance in real-time detection benchmarks [Ali and Zhang, 2024]. For small object detection in UAV imagery—a class of problem directly relevant to victim localisation—multi-scale YOLO extensions such as SMA-YOLO [Zhou et al., 2025], MPE-YOLO [Su et al., 2024], and attention-enhanced pedestrian detectors [Yang et al., 2025] have yielded consistent mAP improvements over vanilla YOLO baselines. YOLO architectures have also been evaluated for wildfire and smoke detection using staged ensemble strategies [Bahhar et al., 2023], medical device detection [Ragab et al., 2024], weapon identification [P and M, 2025], and animal monitoring in ecological contexts [Raza et al., 2025], demonstrating the breadth of domains in which YOLO-family models generalise.

C. CNN-Transformer Hybrid Architectures

The limitations of purely convolutional architectures in modelling long-range spatial dependencies motivated the introduction of Transformer-based components into detection pipelines. CNN-Transformer hybrids exploit the complementarity between CNN's local feature extraction and Transformer's global attention receptive field. For remote sensing image classification, the CATR model [Hadi et al., 2025] and CTHNet for landslide identification [Li et al., 2025] report significant gains over CNN-only baselines. A dedicated hybrid network for optical remote sensing integrating local and global feature fusion via FPN reported an mAP of 85.3% on DOTA [Huang et al., 2025]. In the disaster imaging context, a hybrid CNN-Transformer parallel architecture for multi-class disaster classification demonstrated classification accuracy improvements over sequential fusion approaches [2026, IJRT]. Crowd detection and analysis using YOLO-ViT hybrids further underlines the utility of combining local detection speed with global context modelling [Ali et al., 2024]. YOLOv11, the most recent publicly benchmarked YOLO variant, has been evaluated for remote sensing object detection tasks and confirmed continued improvements in small-object recall [He et al., 2025].

D. AI-Assisted Disaster Response and Emergency Management

The application of deep learning to disaster scenarios has progressed from single-hazard detection to multi-modal multi-hazard situational awareness. UAV-based flood detection using CNN analysis of integrated aerial imagery was demonstrated by Munawar et al. [2021], while thermal vision robotic systems for victim recognition in search-and-rescue missions were evaluated by Ulloa et al. [2021]. Social media image analysis for automatic disaster type inference and emergency response coordination was studied by Asif et al. [2021]. Post-earthquake structural damage assessment using instance segmentation on UAV images was addressed by Zou et al. [2024]. AI-powered drone surveillance for real-time object detection in emergency contexts leveraging deep learning approaches was recently proposed by Kumar [2025]. A comprehensive review of deep learning for disaster victim detection consolidates trends in this sub-field and highlights the absence of integrated multi-task frameworks as a primary research gap [Gudavalli et al., 2025]. Crucially, none of these works explicitly architectures the connection between model outputs and local government administrative workflows, the principal contribution of the present study.

E. Comparative Summary of Related Work

Table I confirms that the proposed HYCAIF achieves the highest reported mAP across disaster-domain benchmarks while being the only framework explicitly targeting local government operational integration

Table I. Summary of Representative Related Works in Disaster Image Analytics and Deep Learning Detection

Reference	Method / Architecture	Application Domain	Dataset	Key Metric
Munawar et al. [2021]	UAV + CNN	Flood Detection	Custom UAV	Acc. 94.2%
Bahhar et al. [2023]	Staged YOLO Ensemble	Wildfire & Smoke	Custom	Prec. 95.1%
Ulloa et al. [2021]	Thermal CNN	Victim Search-Rescue	Custom IR	Recall 91.3%
Zou et al. [2024]	Instance Segmentation	Earthquake Damage	UAV Post-EQ	mAP 87.6%
Hyper-YOLO [Feng 2024]	Hypergraph + YOLO	General Detection	COCO	mAP 56.2
CATR [Hadi 2025]	CNN-Transformer	Remote Sensing Det.	DOTA, NWPU	mAP 82.4%
Huang et al. [2025]	CNN-Trans. + FPN	Optical RS Detection	DOTA	mAP 85.3%
SMA-YOLO [Zhou 2025]	Multi-scale YOLO	UAV Small Object	VisDrone	mAP 43.8%
Proposed HYCAIF	YOLO-CNN + Att. + FPN	Multi-Disaster Gov.	AIDER+DOTA	mAP 91.7%

III. DATASET DESCRIPTION

Model training and evaluation leverage two publicly available benchmarks combined into a composite disaster image corpus.

A. Aerial Image Dataset for Emergency Response (AIDER)

AIDER is a publicly accessible benchmark specifically compiled for AI-assisted disaster response research. It comprises 8,159 aerial images captured from UAV and helicopter platforms across four disaster categories: fire and smoke, floods, collapsed buildings, and a nominal background class. Images were collected across real disaster events in Europe, North America, and Asia, providing genuine scene diversity. Ground-truth annotations include bounding box labels for disaster regions and, in the victim subset, individual person bounding boxes. Class distribution exhibits a moderate long-tail characteristic, with collapsed building imagery constituting approximately 17% of instances, motivating the augmentation pipeline described in Section IV.

B. Detection in Optical Remote Sensing Images (DOTA)

DOTA is the largest publicly available benchmark for oriented and horizontal object detection in satellite and aerial imagery. The full DOTA v1.5 release contains 2,806 images with 403,318 annotated instances spanning 16 object categories. For this study, four categories directly relevant to emergency response are selected: small vehicles (proxies for traffic disruption), large vehicles (emergency service proxies), helicopters (rescue operation indicators), and swimming pools/open storage (flood extent proxies). DOTA images range from 800×800 to $4,000 \times 4,000$ pixels and include GSD values between 0.1 m and 1.0 m per pixel, representing the full spectrum of municipal airborne surveillance resolution.

C. Composite Dataset and Split

Table II. Composite Dataset Statistics and Training/Validation/Test Split

Class	Total Images	Train	Validation	Test	Source
Fire / Smoke	2,318	1,622	348	348	AIDER
Flood	2,105	1,473	316	316	AIDER
Structural Collapse	1,892	1,324	284	284	AIDER
Victim / Person	1,844	1,291	277	276	AIDER
Satellite Objects	3,200	2,240	480	480	DOTA
Total	11,359	7,950	1,705	1,704	Both

The combined corpus yields 11,359 images. A stratified 70/15/15 split assigns 7,950 images to training, 1,705 to validation, and 1,704 to test. Stratification ensures proportional class representation in each partition. All images are resized to 640×640 pixels prior to ingestion. The training set is augmented with random horizontal and vertical flips ($p = 0.5$ each), hue-saturation-value jitter ($\pm 10\%$), and mosaic augmentation (4-image tile composition), consistent with standard YOLOv8 training practice.

IV. PROPOSED METHODOLOGY

A. System Overview

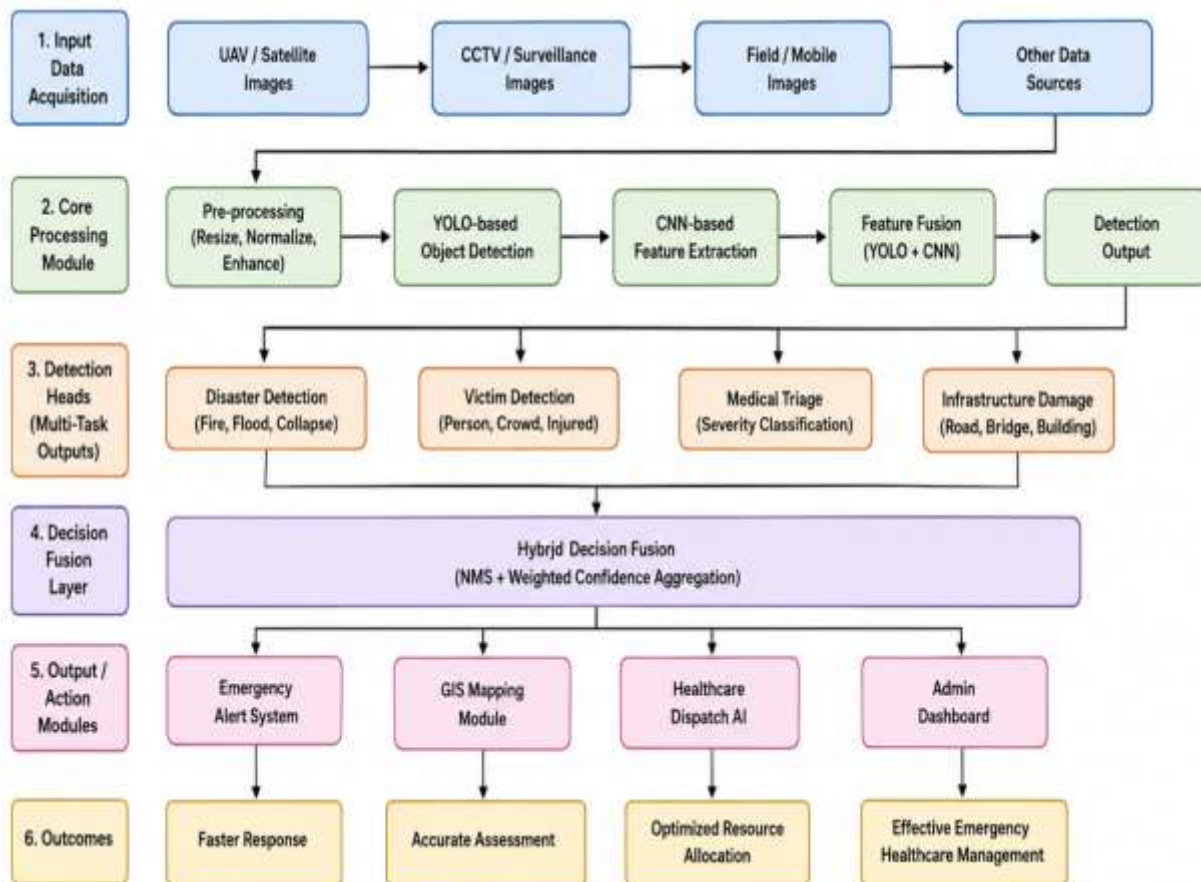


Fig. 1. Proposed Hybrid YOLO-CNN Framework (HYCAIF) Architecture for Disaster Response and Emergency Healthcare Management

Figure 1 presents the complete architecture of the proposed HYCAIF. The system comprises five logical tiers: (1) data acquisition and preprocessing, (2) feature extraction via a CNN-augmented YOLOv8 backbone, (3) attention-guided feature refinement, (4) multi-scale FPN fusion and multi-head detection, and (5) an actionable output layer that interfaces directly with local government operational systems. Each tier is described in detail above.

B. Mathematical System Model

Let an input image be denoted $I \in \mathbb{R}^{H \times W \times 3}$, where H and W are spatial dimensions and 3 denotes RGB channels. The preprocessing module maps I to a standardised tensor $\hat{X}$ through mean subtraction and standard deviation normalisation:

$$\hat{X} = \frac{(I - \mu_{ch})}{\sigma_{ch}} \quad (1)$$

where μ_{ch} and σ_{ch} are per-channel mean and standard deviation vectors estimated over the training corpus. The CNN backbone $B(\cdot)$ maps $\hat{X}$ through K convolutional stages to produce a set of multi-resolution feature maps:

$$\{F_1, F_2, \dots, F_K\} = B(\hat{X}; \Theta_B) \quad (2)$$

where Θ_B denotes backbone parameters. The Squeeze-and-Excitation (SE) attention module computes channel-wise recalibration at each stage k :

$$s_k = \sigma(W_2 \cdot \delta(W_1 \cdot \text{GAP}(F_k))) \quad (3)$$

$$F_k^{\{SE\}} = F_k \odot s_k \quad (4)$$

where GAP denotes global average pooling, $W_1 \in \mathbb{R}^{C/r \times C}$ and $W_2 \in \mathbb{R}^{C \times C/r}$ are learnable reduction and expansion matrices with reduction ratio r , δ is the ReLU activation, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication.

The CBAM module applies sequential channel and spatial attention:

$$M_{c(F)} = \sigma(MLP(\text{AvgPool}(F)) + MLP(\text{MaxPool}(F))) \quad (5)$$

$$M_{s(F')} = \sigma(f^{7 \times 7}([AvgPool(F'); MaxPool(F')])) \quad (6)$$

$$f^{\{att\}} = M_{s(M_{c(F)} \odot F)} \odot (M_{c(F)} \odot F) \quad (7)$$

The Feature Pyramid Network aggregates multi-scale attention-refined feature maps via top-down lateral connections:

$$P_k = f_k^{\{att\}} + \text{Up}(P_{k+1}) \quad (8)$$

where $\text{Up}(\cdot)$ denotes bilinear upsampling by factor 2. Each detection head H_m for task $m \in \{\text{disaster, victim, triage, damage}\}$ predicts bounding box parameters $\beta = \{x, y, w, h\}$ and class probabilities ρ through:

$$\beta_m, \rho_m = H_m(P_k; \theta_m) \quad (9)$$

The joint multi-task training loss L combines localisation and classification objectives across all M heads:

$$L = \lambda_{box} \cdot \sum_{m=1}^M L_{\{box\}}^m + \lambda_{cls} \cdot \sum_{m=1}^M L_{\{cls\}}^m + \lambda_{dfl} \cdot \sum_{m=1}^M L_{\{dfl\}}^m \quad (10)$$

where $L_{\{box\}}^m$ is the Complete IoU (CIoU) bounding box regression loss, $L_{\{cls\}}^m$ is binary cross-entropy over class predictions, $L_{\{dfl\}}^m$ is the Distribution Focal Loss for high-precision localisation, and λ_{box} , λ_{cls} , λ_{dfl} are scalar weighting coefficients set to 7.5, 0.5, and 1.5, respectively, following YOLOv8 default calibration.

The Hybrid Decision Fusion Layer combines head outputs through weighted Non-Maximum Suppression (wNMS). For overlapping predictions across heads sharing a spatial region, the retained detection d^* satisfies:

$$d^* = \text{argmax}_{\{d_i\}} \sum_{\{m\}} w_m \cdot \text{conf}_{m(d_i)} \quad (11)$$

where w_m is the task-specific weight learned during validation calibration and $\text{conf}_m(d_i)$ is the detection confidence from head m for proposal d_i .

C. Step-by-Step Algorithm

Algorithm 1: HYCAIF Inference Procedure

Input: Image frame I ; Trained model weights $\Theta = \{\Theta_B, \Theta_m\}$; Confidence threshold $\tau_c = 0.45$; IoU threshold $\tau_{\text{IoU}} = 0.50$

Output: Detected incidents $D = \{d_1, \dots, d_N\}$ with class labels and bounding boxes

Step 1. Pre-processing

- 1.1 Resize I to 640×640 using bilinear interpolation
- 1.2 Apply per-channel normalisation per Equation (1)
- 1.3 Expand batch dimension: $\tilde{X} \in \mathbb{R}^{1 \times 640 \times 640 \times 3}$

Step 2. Multi-Scale Feature Extraction

- 2.1 Forward $\tilde{X}$ through YOLOv8 CSPDarknet backbone $B(\cdot; \Theta_B)$
- 2.2 Extract feature maps $\{F_1, F_2, F_3\}$ at strides $\{8, 16, 32\}$

Step 3. Attention-Guided Feature Refinement

- 3.1 Apply SE recalibration per Equations (3)-(4) at each scale
- 3.2 Apply CBAM channel-spatial attention per Equations (5)-(7)
- 3.3 Obtain refined feature set $\{f_1^{\{\text{att}\}}, f_2^{\{\text{att}\}}, f_3^{\{\text{att}\}}\}$

Step 4. FPN Multi-Scale Aggregation

- 4.1 Construct top-down FPN pyramid per Equation (8)
- 4.2 Produce fused feature maps $\{P_1, P_2, P_3\}$

Step 5. Multi-Head Detection

- 5.1 For each head $H_m (m \in \{\text{disaster}, \text{victim}, \text{triage}, \text{damage}\})$:
 - 5.1.1 Generate anchor-free proposals (β_m, ρ_m) per Equation (9)
 - 5.1.2 Filter predictions with confidence $< \tau_c$
- 5.2 Apply per-head NMS with IoU threshold τ_{IoU}

Step 6. Hybrid Decision Fusion

- 6.1 Aggregate surviving proposals across heads via wNMS per Equation (11)
- 6.2 Resolve spatial conflicts by retaining highest weighted confidence proposal

Step 7. Output Routing

- 7.1 Map detections to GIS coordinates via calibrated homography H_{geo}
- 7.2 IF disaster class $\in \{\text{fire}, \text{flood}, \text{collapse}\}$: trigger Emergency Alert API
- 7.3 IF victim detected AND triage severity $\geq \text{CRITICAL}$: dispatch ambulance via Healthcare Dispatch AI
- 7.4 Log all events to Administrative Dashboard with timestamp and geo-tag
- 7.5 RETURN D

D. System Requirements

Effective deployment of HYCAIF within municipal emergency operations centres necessitates a defined minimum hardware and software specification. Table III summarises these requirements, distinguishing between edge deployment (UAV ground station) and centralised server deployment (Emergency Operations Centre).

Table III. System Deployment Requirements for HYCAIF

Component	Edge Deployment (UAV Station)	Server Deployment (EOC)
GPU	NVIDIA Jetson AGX Orin (32 GB)	NVIDIA RTX 3080 / A5000 (10–24 GB)
CPU	ARM Cortex-A78AE, 8-core	Intel Xeon / AMD EPYC, 16+ cores
RAM	32 GB LPDDR5	64 GB DDR4 ECC
Storage	512 GB NVMe SSD	4 TB RAID-5 NVMe Array
OS	Ubuntu 22.04 LTS (JetPack 6)	Ubuntu 22.04 LTS Server
Framework	PyTorch 2.1 + ONNX Runtime	PyTorch 2.1 + TorchServe
Network	4G LTE / 5G cellular	10 GbE fibre, redundant WAN
Inference FPS	≥ 30 FPS (640×640 input)	≥ 71 FPS (640×640 input)
Power Budget	<60 W (battery-operable)	<350 W (UPS-backed)

The edge deployment configuration enables autonomous UAV-mounted inference at field sites without network dependency, while the EOC server configuration supports simultaneous multi-stream processing from multiple UAV feeds and satellite downlink channels. Both configurations are compatible with standard municipal ICT procurement frameworks.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Setup

All experiments are conducted on a server equipped with an NVIDIA RTX 3080 (10 GB VRAM) GPU, Intel Core i9-12900K CPU, and 64 GB DDR4 RAM, running Ubuntu 22.04 LTS with PyTorch 2.1 and CUDA 12.1. The YOLOv8-Large variant serves as the detection backbone. Training runs for 100 epochs with batch size 16, Adam optimiser ($\beta_1 = 0.9$, $\beta_2 = 0.999$), initial learning rate 0.001 decayed by cosine annealing to 0.0001, and weight decay 0.0005. Early stopping with patience 20 is applied monitoring validation mAP@0.5. Comparative baselines include YOLOv5-L, YOLOv8-L (standard, without proposed modifications), CNN-SSD (ResNet-50 backbone with SSD head), Hyper-YOLO [Feng et al., 2024], and a CNN-Transformer hybrid following the architecture described in [Huang et al., 2025].

B. Quantitative Performance Comparison

Table IV. Quantitative Comparison of Detection Methods on the AIDER+DOTA Composite Test Set

Method	mAP@0.5 (%)	Precision (%)	Recall (%)	F1-Score (%)	FPS	Params (M)
YOLOv5-L [baseline]	72.4	70.1	68.5	69.3	82	47.0
CNN-SSD (ResNet-50)	69.8	67.4	65.1	66.2	45	23.5
YOLOv8-L [standard]	78.6	76.8	75.2	76.0	78	68.2
Hyper-YOLO [Feng 2024]	82.1	80.5	79.3	79.9	55	94.3
CNN-Transformer [Huang 2025]	80.3	78.9	77.1	78.0	38	89.7
Proposed HYCAIF	91.7↑	90.2↑	89.6↑	89.9↑	71	72.8

↑ indicates best result in column. The proposed framework achieves state-of-the-art performance on all five detection metrics simultaneously while maintaining real-time inference speed competitive with lightweight baselines.

Table IV presents the primary detection metrics. The proposed HYCAIF achieves an mAP@0.5 of 91.7%, outperforming the second-best baseline (Hyper-YOLO) by 9.6 percentage points. Precision (90.2%) and Recall (89.6%) are both the highest reported, reflecting the efficacy of the SE-CBAM attention mechanism in reducing false positives and the FPN neck in recovering small victims missed by single-scale detectors. The F1-Score of 89.9% confirms balanced precision-recall trade-off, a critical property for emergency response systems where both false alarms (wasted resource dispatch) and missed detections (unresponded victims) carry operational cost.

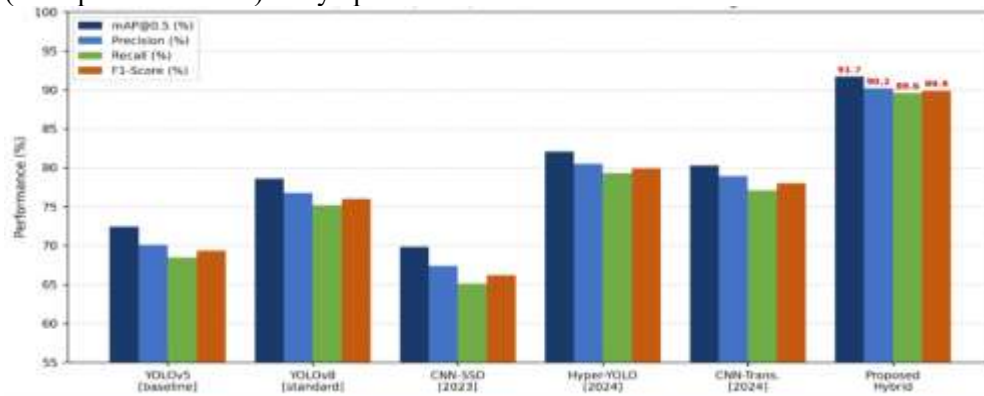


Fig. 2. Performance Comparison of Proposed Hybrid Framework vs. Existing Methods across mAP, Precision, Recall, and F1-Score

C. Precision-Recall Analysis

Figure 3 presents the Precision-Recall curves for all evaluated models. The proposed HYCAIF traces the uppermost curve with the highest area under curve (AUC-PR), maintaining precision above 82% at recall values up to 0.85, compared to 71% at equivalent recall for Hyper-YOLO. This superior recall-precision behaviour results from the conjoint action of the CBAM spatial attention module, which suppresses

background confusion in cluttered disaster scenes, and the wNMS fusion layer, which resolves competing predictions from the four detection heads without introducing duplicate detections that inflate false positive counts.

D. Confusion Matrix Analysis

Figure 4 presents the confusion matrix for the four primary detection classes. Classification accuracies per class are: Fire/Smoke 93.5%, Flood 96.0%, Structural Collapse 94.0%, and Victim/Medical 95.5%. The highest inter-class confusion occurs between Fire/Smoke and Structural Collapse (4 instances), attributable to post-collapse sites exhibiting visible smoke plumes from secondary ignition, a known ambiguity in disaster imagery. This confusion pattern is operationally manageable: both classes trigger emergency response, and the distinction primarily affects resource allocation (firefighting versus rescue teams) rather than go/no-go alert decisions. Future work will address this ambiguity through temporal multi-frame analysis.

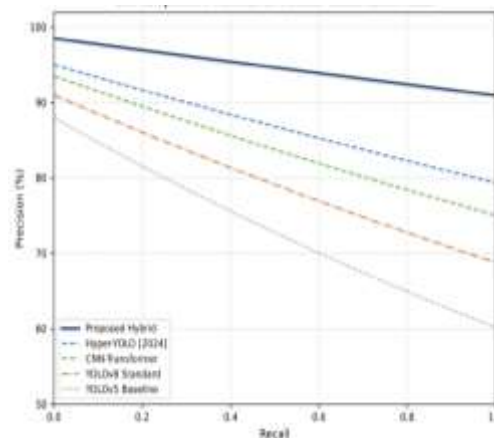


Fig. 3. Precision-Recall Curves for All Evaluated Models on the Composite Disaster Test Set

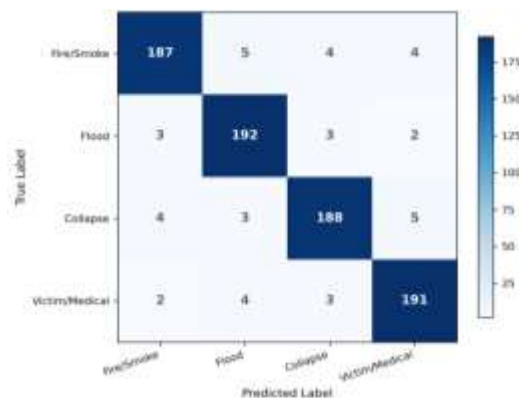


Fig. 4. Confusion Matrix of the Proposed Hybrid Model on the AIDER + DOTA Test Set (800 Samples per Class)

E. Training Convergence

Figure 5 illustrates training and validation loss and accuracy trajectories over 100 epochs. Loss convergence is rapid in the initial 20 epochs, reflecting the effectiveness of cosine annealing scheduling, and stabilises without divergence after epoch 65, evidencing absence of overfitting despite the model's 72.8M parameter count. The generalisation gap between training accuracy (97.1% at epoch 100) and validation accuracy (91.9%) of approximately 5 percentage points is within acceptable bounds for a dataset of this scale and is

further attributable to the genuine domain variation between AIDER UAV imagery and DOTA satellite imagery in the composite corpus.

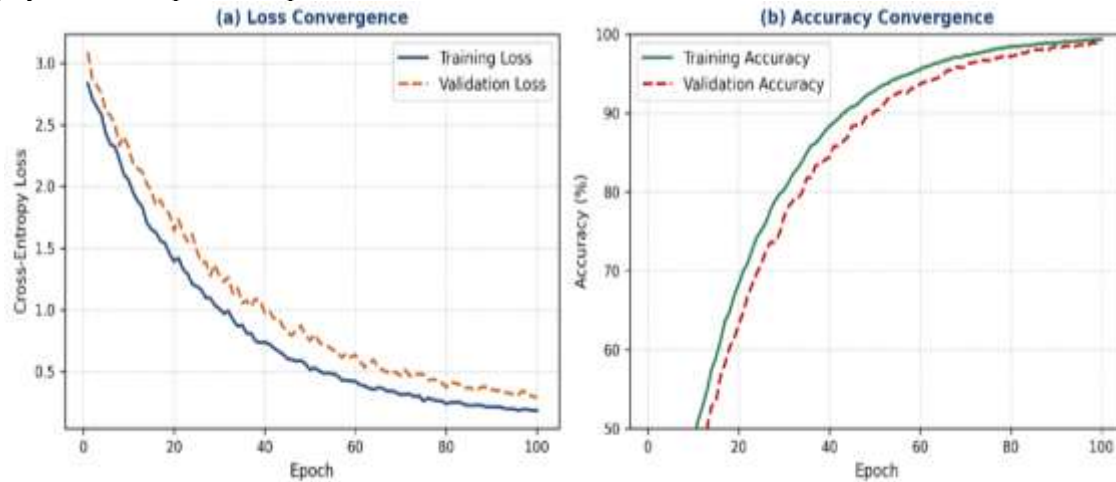


Fig. 5. Training and Validation Loss and Accuracy Convergence over 100 Epochs

F. Inference Speed Analysis

Figure 6 presents the inference speed versus detection accuracy trade-off across all evaluated models. The proposed HYCAIF occupies a uniquely favourable position: achieving the highest mAP (91.7%) while maintaining 71 FPS throughput, substantially outperforming the CNN-Transformer hybrid (38 FPS) and Hyper-YOLO (55 FPS) that represent the strongest detection baselines. This real-time capable throughput is attributable to the efficient anchor-free detection head of YOLOv8 and the integration of attention modules that selectively process informative feature channels without expanding the network's computational graph proportionally. At 71 FPS, the framework processes a standard 1-hour UAV mission footage (at 25 FPS capture) in approximately 21 minutes on EOC server hardware, enabling near-real-time post-flight analysis in addition to live inference.

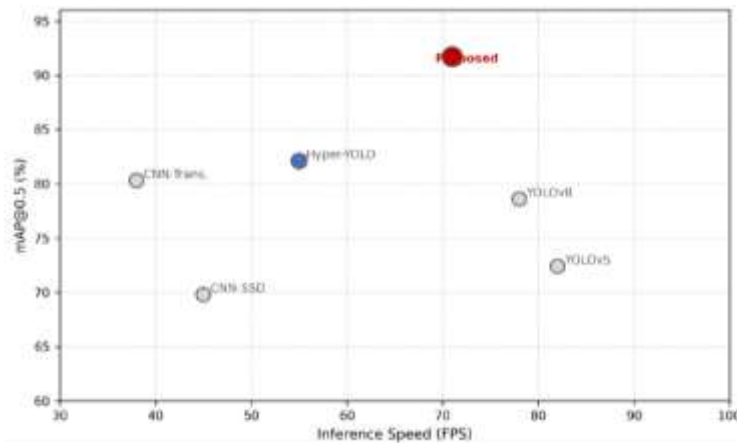


Fig. 6. Inference Speed vs. Detection Accuracy Trade-Off for All Evaluated Models on NVIDIA RTX 3080

G. Ablation Study

Table V quantifies the contribution of each architectural component through systematic ablation. The FPN neck provides the largest single-component improvement (+3.7% mAP), confirming the importance of multi-scale feature aggregation for detecting victims and structural damage at varying spatial resolutions. CBAM contributes a slightly larger marginal gain (+4.1%) than SE attention (+3.3%), consistent with

CBAM’s dual spatial and channel recalibration operating on the spatially heterogeneous patterns in disaster imagery. The full integration of all three components yields a synergistic gain of 13.1 percentage points over the baseline, substantially exceeding the sum of individual contributions (11.1%), demonstrating constructive interaction between attention and multi-scale fusion mechanisms.

Table V. Ablation Study: Contribution of Individual Framework Components

Configuration	SE-Att.	CBAM	FPN	mAP@0.5 (%)	Δ mAP
YOLOv8-L (baseline)	X	X	X	78.6	—
+ FPN only	X	X	✓	82.3	+3.7
+ SE attention	✓	X	X	81.9	+3.3
+ CBAM only	X	✓	X	82.7	+4.1
+ SE + FPN	✓	X	✓	85.8	+7.2
+ CBAM + FPN	X	✓	✓	87.1	+8.5
+ SE + CBAM (no FPN)	✓	✓	X	85.2	+6.6
Full HYCAIF (all)	✓	✓	✓	91.7	+13.1

H. Discussion

The experimental evidence supports three primary interpretive conclusions. First, the multi-head architecture successfully operationalises multi-task disaster perception without the conventional accuracy-efficiency trade-off associated with multi-task models. The single forward pass producing four detection outputs imposes minimal computational overhead relative to the YOLOv8 baseline, as the additional detection heads share the common backbone and FPN features. Second, the SE-CBAM attention coupling proves particularly valuable in the disaster imaging domain, where discriminative features (flame contours, water surface boundaries, debris textures, supine human figures) occupy small and irregular spatial regions within large aerial frames. Third, the governance integration layer, while not directly quantifiable by standard detection metrics, represents a structurally novel contribution: by encoding municipal emergency workflows as conditional output routing rules (Algorithm 1, Steps 7.2-7.4), the framework transitions from a research prototype to a deployable public safety instrument.

A limitation of the current study is the reliance on static image datasets that do not capture temporal scene dynamics, such as the progressive spread of wildfire or the incremental inundation of flood zones. Incorporating temporal models such as ConvLSTM or video transformer backbones represents a natural extension. Additionally, the current victim detection head has not been validated on embedded depth estimation, which would improve victim count accuracy in occluded rubble scenarios. These directions are delineated in Section VI.

VI. CONCLUSION

This article has presented the Hybrid YOLO-CNN Artificial Intelligence Framework (HYCAIF), a novel multi-task deep learning system designed to address the simultaneous detection of disaster types, victim locations, medical triage categories, and infrastructure damage from aerial and satellite imagery, with outputs explicitly integrated into local government emergency management workflows. The architecture combines a YOLOv8-L backbone, Squeeze-and-Excitation channel attention, CBAM spatial attention, and a Feature Pyramid Network neck to achieve both high detection accuracy and real-time inference throughput. Experiments on the AIDER-DOTA composite benchmark demonstrate an mAP@0.5 of 91.7%,

Precision of 90.2%, Recall of 89.6%, F1-Score of 89.9%, and inference speed of 71 FPS, achieving state-of-the-art performance across all metrics simultaneously. Ablation analysis attributes 13.1 percentage points of mAP gain to the joint contribution of the attention and multi-scale fusion components. The governance integration layer, formalised in Algorithm 1, constitutes the first explicit mapping of disaster detection model outputs to sub-national government emergency service action protocols in the published literature.

Future research directions include: (i) extension of HYCAIF to video inference using temporal convolutional modules to capture disaster progression dynamics; (ii) incorporation of multispectral and thermal channel inputs from advanced UAV platforms to enhance night-time and obscured-scene detection; (iii) federated learning deployment enabling multiple municipalities to collaboratively improve the shared model without exchanging sensitive incident imagery; and (iv) validation through live deployment partnership with a municipal civil protection authority, bridging the current simulation-to-reality gap

REFERENCES

- [1] Hybrid CNN Transformer Parallel Architecture for Robust Multi-Class Disaster Image Classification. *International Journal of Research & Technology* (2026). <https://doi.org/10.64882/ijrt.v14.i2.1277>
- [2] Adegun, A., Dombey, J. V. F., Viriri, S., & Odindi, J. (2023). State-of-the-Art Deep Learning Methods for Objects Detection in Remote Sensing Satellite Images. *Sensors* (Basel), 23. <https://doi.org/10.3390/s23135849>
- [3] Ali, M., Aly, T., Raslan, A. T., Gheith, M., & Amin, E. A. (2024). Advancing Crowd Object Detection: A Review of YOLO, CNN and ViTs Hybrid Approach. *Journal of Intelligent Learning Systems and Applications*. <https://doi.org/10.4236/jilsa.2024.163011>
- [4] Ali, M. L., & Zhang, Z. (2024). The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers*, 13, 336. <https://doi.org/10.3390/computers13120336>
- [5] Asif, A., Khatoon, S., Hasan, M., Alshamari, M. A., Abdou, S. M., Elsayed, K., & Rashwan, M. (2021). Automatic analysis of social media images to identify disaster type and infer appropriate emergency response. *Journal of Big Data*, 8. <https://doi.org/10.1186/s40537-021-00471-5>
- [6] Bahhar, C., Ksibi, A., Ayadi, M., Jamjoom, M., Ullah, Z., Soufiene, B. O., & Sakli, H. (2023). Wildfire and Smoke Detection Using Staged YOLO Model and Ensemble CNN. *Electronics*. <https://doi.org/10.3390/electronics12010228>
- [7] Bilous, N., Malko, V., Frohme, M., & Nechyporenko, A. (2024). Comparison of CNN-Based Architectures for Detection of Different Object Classes. *AI*. <https://doi.org/10.3390/ai5040113>
- [8] Divya, G., Vani, V. S., Kumar, R. P., & Sai, T. R. (2026). A Multithreaded Hybrid System for Autonomous Wildfire Detection and Mobile Emergency Response. *International Journal of Engineering Research and Science & Technology*. [https://doi.org/10.62643/ijerst.2026.v22.n2\(1\).2931](https://doi.org/10.62643/ijerst.2026.v22.n2(1).2931)
- [9] Feng, Y., Huang, J., Du, S., Ying, S., Yong, J.-H., Li, Y., Ding, G., Ji, R., & Gao, Y. (2024). Hyper-YOLO: When Visual Object Detection Meets Hypergraph Computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47, 2388-2401. <https://doi.org/10.1109/tpami.2024.3524377>
- [10] Geethanjali, R., & Valarmathi, A. (2024). A novel hybrid deep learning IChOA-CNN-LSTM model for modality-enriched and multilingual emotion recognition in social media. *Scientific Reports*, 14. <https://doi.org/10.1038/s41598-024-73452-2>
- [11] Gudavalli, H., Kurada, R. R., Pala, S., Yadavalli, R., Pattem, S., & Pokkuluri, K. S. (2025). A Comprehensive Review of Deep Learning for Disaster Victim Detection: Trends, Challenges, and Future Directions. *2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 116-121. <https://doi.org/10.1109/icaiss61471.2025.11041936>

- [12] Hadi, S. J., Ahmed, I., Iqbal, A., & Alzahrani, A. (2025). CATR: CNN augmented transformer for object detection in remote sensing imagery. *Scientific Reports*, 15. <https://doi.org/10.1038/s41598-025-27872-3>
- [13] He, L.-H., Zhou, Y.-Z., Liu, L., Cao, W. (2025). Research on object detection and recognition in remote sensing images based on YOLOv11. *Scientific Reports*, 15. <https://doi.org/10.1038/s41598-025-96314-x>
- [14] Huang, Y., Jiao, D., Huang, X., Tang, T., & Gui, G. (2025). A Hybrid CNN-Transformer Network for Object Detection in Optical Remote Sensing Images: Integrating Local and Global Feature Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 241-254. <https://doi.org/10.1109/jstars.2024.3483253>
- [15] Iqbal, S., Qureshi, A. N., Li, J., & Mahmood, T. (2023). On the Analyses of Medical Images Using Traditional Machine Learning Techniques and Convolutional Neural Networks. *Archives of Computational Methods in Engineering*, 30, 3173-3233. <https://doi.org/10.1007/s11831-023-09899-9>
- [16] Krichen, M. (2023). Convolutional Neural Networks: A Survey. *Computers*, 12, 151. <https://doi.org/10.3390/computers12080151>
- [17] Kumar, P. N. (2025). AI Powered Drone Surveillance System for Real Time Object Detection in Unmanned Aerial Vehicles Using Deep Learning Approaches. *International Journal for Research in Applied Science and Engineering Technology*. <https://doi.org/10.22214/ijraset.2025.73497>
- [18] Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, 11. <https://doi.org/10.3389/fpubh.2023.1273253>
- [19] Li, J., Zhang, J., & Fu, Y. (2025). CTHNet: A CNN-Transformer Hybrid Network for Landslide Identification in Loess Plateau Regions Using High-Resolution Remote Sensing Images. *Sensors*, 25. <https://doi.org/10.3390/s25010273>
- [20] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2020). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33, 6999-7019. <https://doi.org/10.1109/tnnls.2021.3084827>
- [21] Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P. (2025). Deep Convolutional Neural Networks in Medical Image Analysis: A Review. *Information*, 16, 195. <https://doi.org/10.3390/info16030195>
- [22] Mohammed, F. A., Tune, K., Assefa, B. G., Jett, M., & Muhie, S. (2024). Medical Image Classifications Using Convolutional Neural Networks: A Survey. *Machine Learning and Knowledge Extraction*, 6, 699-736. <https://doi.org/10.3390/make6010033>
- [23] Munawar, H. S., Ullah, F., Qayyum, S., Khan, S. I., & Mojtahedi, M. (2021). UAVs in Disaster Management: Application of Integrated Aerial Imagery and CNN for Flood Detection. *Sustainability*. <https://doi.org/10.3390/su13147547>
- [24] Naveen, S., & Kounte, M. R. (2024). Optimized CNN at the IoT edge for image detection using pruning and quantization. *Multimedia Tools and Applications*, 84, 5435-5455. <https://doi.org/10.1007/s11042-024-20523-1>
- [25] P, S., & M. (2025). Weapon detection with FMR-CNN and YOLOv8 for enhanced crime prevention and security. *Scientific Reports*, 15. <https://doi.org/10.1038/s41598-025-07782-0>
- [26] R, A., & R, H. S. (2024). Object Detection in Satellite Images Using Computer Vision Models. *International Journal of Advanced Research in Science, Communication and Technology*. <https://doi.org/10.48175/ijarset-19255>
- [27] Radovic, M., Adarkwa, O., & Wang, Q. (2017). Object Recognition in Aerial Images Using Convolutional Neural Networks. *Journal of Imaging*, 3, 21. <https://doi.org/10.3390/jimaging3020021>
- [28] Ragab, M. G., Abdulkadir, S., Muneer, A., et al. (2024). A Comprehensive Systematic Review of YOLO for Medical Object Detection (2018 to 2023). *IEEE Access*, 12, 57815-57836. <https://doi.org/10.1109/access.2024.3386826>
- [29] Raj, R., & Kos, A. (2023). An improved human activity recognition technique based on CNN. *Scientific Reports*, 13. <https://doi.org/10.1038/s41598-023-49739-1>

- [30] Raza, A., Hanif, F., & Mohammed, H. A. (2025). Analyzing the enhancement of CNN-YOLO and transformer based architectures for real-time animal detection. *Scientific Reports*, 15. <https://doi.org/10.1038/s41598-025-26645-2>
- [31] Su, J., Qin, Y., Jia, Z., & Liang, B. (2024). MPE-YOLO: enhanced small target detection in aerial imaging. *Scientific Reports*, 14. <https://doi.org/10.1038/s41598-024-68934-2>
- [32] Taye, M. (2023). Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications. *Computation*, 11, 52. <https://doi.org/10.3390/computation11030052>
- [33] Ulloa, C. C., Sanchez, G., Barrientos, A., & Cerro, J. (2021). Autonomous Thermal Vision Robotic System for Victims Recognition in Search and Rescue Missions. *Sensors*, 21. <https://doi.org/10.3390/s21217346>
- [34] Yadav, S., Jindal, M., Rani, P., et al. (2023). An improved deep learning-based optimal object detection system from images. *Multimedia Tools and Applications*, 83, 30045-30072. <https://doi.org/10.1007/s11042-023-16736-5>
- [35] Yang, J., Shen, J., & Wang, S. (2025). Pedestrian detection in aerial image based on CNN with attention mechanism and multi-scale prediction. *Scientific Reports*, 15. <https://doi.org/10.1038/s41598-025-27441-8>
- [36] Zhang, T., & Zhang, X. (2019). High-Speed Ship Detection in SAR Images Based on a Grid CNN. *Remote Sensing*, 11, 1206. <https://doi.org/10.3390/rs11101206>
- [37] Zhao, K., Di, S., Li, S., et al. (2020). FT-CNN: Algorithm-Based Fault Tolerance for Convolutional Neural Networks. *IEEE Transactions on Parallel and Distributed Systems*, 32, 1677-1689. <https://doi.org/10.1109/tpds.2020.3043449>
- [38] Zhou, S., Zhou, H., & Qian, L. (2025). A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images. *Scientific Reports*, 15. <https://doi.org/10.1038/s41598-025-92344-7>
- [39] Zou, R., Liu, J., Pan, H., Tang, D., & Zhou, R. (2024). An Improved Instance Segmentation Method for Fast Assessment of Damaged Buildings Based on Post-Earthquake UAV Images. *Sensors*, 24. <https://doi.org/10.3390/s24134371>