SL-YOLO-DR: YOLO-Based Object Detection for Disaster Response: Supervised Baselines and Semi-Supervised Learning Enhancements

Afsin Sultana
Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh
2022-1-60-113@std.ewubd.edu

Abstract

For disaster response, object detection is vital, but the absence of labeled data makes it difficult to develop accurate models. Using recent YOLO architectures, this paper investigates the effects of self-supervised and semi-supervised learning on disaster-focused object detection. YOLOv10n, YOLOv11s, and YOLOv12n are assessed. DINO- and BYOL-pretrained YOLOv11s are used in additional experiments on a multi-class disaster dataset containing vehicles, people, fire, and smoke. According to experimental results, self-supervised pretraining consistently improves performance and YOLOv11s performs better than other supervised baselines. While BYOL exhibits higher precision but lower recall, showing a trade-off in prediction behavior, DINO pretraining yields the best overall results among the SSL approaches, improving both detection accuracy and generalization. Performance improvements are most noticeable for vehicle classes, but visual ambiguity makes fire and smoke difficult to see. The findings support the utilization of self-supervised learning in real-world emergency response systems by highlighting it as a workable solution for disaster object detection in low-label environments.

I Introduction

Recent developments in computer vision have greatly enhanced object detection systems, making it possible to use them in practical applications like disaster relief, autonomous driving, and surveillance. The You Only Look Once (YOLO) family of models is popular for object detection tasks that require both performance and efficiency because it strikes a balance between real-time inference speed and detection accuracy. Architectural design, feature extraction, and multi-scale detection capabilities have significantly improved as a result of YOLO's development from early iterations to more recent releases (such as YOLOv12) [1]. However, there are particular difficulties in disaster situations like fires, collapsed infrastructure, and dense smoke. Large, accurately labeled datasets are expensive and time-consuming to gather, and images taken in such environments frequently contain visually complex, occluded, or confusing objects. Self-supervised learning (SSL) and semi-supervised learning (Semi-SL) have become promising methods to use unlabeled data for representation learning and detector training in order to reduce reliance on manual annotations. Recent surveys on self-supervised learning (SSL) for object detection report Alina et al. [2] includes that although SSL methods effectively learn visual representations from unlabeled data, most approaches are originally designed for image classification and do not directly optimize localization performance. As a result, their effectiveness decreases for small, occluded, or visually ambiguous objects commonly found in disaster scenarios, indicating a lack of detection-specific SSL adaptation.

Similarly, studies on semi-supervised object detection Tahira et al. [3] highlight challenges in utilizing unlabeled data due to noisy and unreliable pseudo-labels, particularly in complex environments. While these methods show promise, limited attention has been given to integrating self-supervised pretraining with semi-supervised learning in one-stage detectors such as YOLO, which are critical for real-time disaster response. Overall, existing research reveals two key gaps: the need for detection-aware self-supervised representations and robust semi-supervised strategies that preserve both accuracy and real-time efficiency in disaster object detection tasks. However, there are particular difficulties in disaster situations like fires, collapsed infrastructure, and dense smoke. Large, accurately labeled datasets are expensive and time-consuming to gather, and images taken in such environments frequently contain visually complex, occluded, or confusing objects. Self-supervised learning (SSL) and semi-supervised learning (Semi-SL) have become promising methods to use unlabeled data for representation learning and detector training in order to reduce reliance on manual annotations. **Problem Statement:** Traditional supervised object detection models require extensive labeled datasets to achieve high accuracy, a requirement that is rarely met in disaster environments due to the high cost of annotation and limited availability of domain-specific labeled data. There is a need for learning frameworks that can leverage abundant unlabeled data to improve object detection performance in disaster scenarios, while maintaining the real-time benefits of one-stage detectors.

Objectives and Hypotheses:

- Analyze how well self-supervised pretraining techniques (like DINO and BYOL) work to enhance YOLO-based object detection models in low-label disaster datasets.

- In terms of detection accuracy and generalization, compare the performance of self-supervised pretrained models with conventional supervised and semi-supervised training pipelines.
- To determine which object categories gain the most from self-supervised and semi-supervised learning techniques, examine class-wise detection behavior.

Significance of the Study: This study shows how robust object detection can be achieved with less annotation effort by utilizing self-supervised and semi-supervised learning frameworks to lessen reliance on labeled disaster imagery. Future advancements in label-efficient object detection systems and real-world disaster response automation, where quick and precise detection is essential, can benefit from the findings.

The rest of this paper is organized as follows: Section 2 reviews prior work. Section 3 presents methodology. Section 4 reports results and experiments. Section 5 discusses implications and limitations. Finally, Section 6 concludes the paper.

II Related Work

YOLO object detection models are primarily fully supervised but have been effectively adapted with semi-supervised techniques like pseudo-labeling and self-supervised pretraining like contrastive learning to reduce reliance on labeled data. This section organizes prior studies to highlight their contributions, limitations of other papers work.

Lee et al. [4] investigates post-disaster building damage assessment from satellite imagery under limited labeled data conditions by leveraging semi-supervised learning. Their methodology applies MixMatch and FixMatch frameworks, which combine pseudo-labeling and consistency regularization on paired pre- and post-disaster images. Experiments are conducted on datasets from the 2010 Haiti earthquake. In the 2017 Santa Rosa wildfire, and the 2016 Aleppo conflict, totaling approximately 74,000 building-level samples sourced from UNOSAT and DigitalGlobe. The study shows a Wide ResNet CNN as the backbone model within the SSL setup. Results show that near fully supervised performance can be achieved with fewer than 500 labeled samples, although the approach is limited by binary damage categorization and sensitivity to data augmentation noise.

Singh et al. [5] focus on multi-class preliminary damage assessment using ultra-high-resolution aerial imagery and semi-supervised transformer models. They adopt an EMA teacher–student training paradigm with weak and strong augmentations to generate pseudo-labels for large volumes of unlabeled building crops. The dataset consists of 1.7 cm GSD UHRA imagery from Hurricane Michael, including 1,072 labeled and 16,800 unlabeled buildings across five damage classes. Vision Transformers (ViT) and a Semi-ViT variant are used as the core models. Their approach achieves up to 88% classification accuracy, surpassing reported human-level performance, but its generalization is constrained by limited disaster diversity and reliance on costly UHRA data.

Almalioglu et al. [6] address SAR-based object detection for disaster response and urban monitoring through self-supervised pretraining. The proposed methodology combines masked image modeling (SSL-MIM) with a curriculum-aware adaptive sampling strategy to mitigate severe class imbalance inherent in SAR imagery. The study leverages over 25,700 km² of unlabeled X-band SAR data from Capella Space alongside a proprietary annotated vehicle dataset. Their TRANSAR model is built on a Swin-Transformer encoder augmented with auxiliary semantic segmentation and adaptive sampling. Experimental results demonstrate superior mAP and F1 scores compared to DeepLabv3, UNet, and SegFormer, though performance remains challenged in dense urban areas with strong radar reflections.

Zi et al. [7] present a comprehensive comparison between supervised and self-supervised learning approaches for property damage classification in remote sensing imagery. The methodology benchmarks conventional CNNs against contrastive self-supervised models, specifically MoCoV2 and DinoV2, with an emphasis on feature attention behavior. Experiments are conducted on the xBD dataset containing pre- and post-disaster image pairs, where severe class imbalance is addressed through cropping strategies and extensive data augmentation. The evaluated models include ResNet, VGG, and MobileNet CNNs, along with MoCoV2 and ViT-based DinoV2 without fine-tuning. Results indicate that DinoV2 achieves up to 99.9% precision on balanced datasets, although limitations persist due to edge-focused attention biases and dataset-specific artifacts.

Chen et al. [8] introduces BRIGHT, a globally distributed multimodal benchmark for building damage assessment aimed at all-weather disaster response, addressing the limitation of optical-only methods under clouds and nighttime. The methodology centers on curating pre-event very-high-resolution (0.3–1 m) optical imagery paired with post-event VHR SAR, followed by rigorous manual alignment, standardized three-class damage labeling (intact/damaged/destroyed), and extensive benchmarking under standard, zero-shot, and one-shot cross-event splits. The dataset covers 14 disaster events (5 natural + 2 human-made) across 23 regions worldwide with 384,596 building instances and 4,246 multimodal image pairs, emphasizing developing regions and open access. The authors evaluate seven deep models (UNet, DeepLabV3+, SiamAttnUNet, SiamCRNN, ChangeOS, DamageFormer, ChangeMamba) and show ChangeMamba achieves the best overall performance (approx 96.2% OA, approx 67.6% mIoU), while multimodal fusion consistently outperforms single-modality setups. Limitations include residual optical–SAR registration error (1 px), label noise from visual interpretation, class/regional imbalance, reliance on single-polarization SAR, and reduced cross-event generalization—highlighting and motivating richer modalities and domain-robust methods.

Ahn et al. [9] propose DAVI, a generalizable disaster damage assessment framework that performs building-level change detection across unseen regions and disaster types. Method-

ology: the approach uses test-time adaptation with pseudo-labeling, combining a source change-detection model with a vision foundation model (Segment Anything Model, SAM), followed by two-stage refinement at pixel and image levels to reduce noise and handle domain and decision-boundary shifts. Dataset: experiments are conducted on the xBD benchmark (multi-disaster, multi-region satellite imagery), with additional validation on the 2023 Türkiye earthquake dataset. Model: SNUNet-CD and BIT-CD backbones are used, enhanced with SAM-guided Diff-SAM pseudo labels and entropy minimization. Results: DAVI consistently outperforms unsupervised and domain-adaptation baselines, achieving higher F1-scores on wildfires, hurricanes, and tsunamis, and strong real-world performance (F1 = 0.68 on Türkiye). Limitations: performance depends on pseudo-label quality and threshold selection, and SAM may miss subtle damage patterns, requiring further robustness studies.

Barco et al. [10] propose a rapid wildfire hotspot detection framework using self-supervised learning (SSL) on multi-temporal satellite data. Methodologically, they adapt the Presto masked autoencoder transformer to Meteosat MSG15 time-series, jointly optimizing reconstruction and binary classification losses. The dataset is newly constructed from EFFIS fire events (2012–2022), combined with MSG HRSEVIRI spectral bands and ESRI LULC, yielding positive/negative pixel time-series. Results show a best test F1-score of 63.58 using a cosine scheduler, demonstrating improved near-real-time detection. Limitations include restricted dataset size, heavy preprocessing costs, and moderate accuracy, indicating scalability and generalization challenges

Rahnemoonfar et al. [11] introduces RescueNet, a high-resolution UAV dataset for semantic segmentation to assess post-disaster damage following Hurricane Michael. The methodology involves collecting 4,494 images via DJI Mavic Pro quadcopters and applying pixel-level annotations for 11 classes, including specific building damage levels. Researchers evaluated the dataset using state-of-the-art models like Attention U-Net and PSPNet, achieving a top mean Intersection over Union (mIoU) of 93.98% for semantic segmentation. Results demonstrated that high-resolution UAV imagery significantly improves the identification of complex features like debris and minor building damage compared to low-resolution satellite data. A primary limitation noted is the high computational cost and time required for manual dense pixel-level annotation of such massive datasets.

Russo et al. [12] This study introduces a novel multimodal deep learning framework that uses single-date, post-event VHR SAR imagery and auxiliary geospatial data (OSM, DSM, and GEM) for building damage assessment. By employing a CNN-based architecture, the model eliminates the need for pre-event imagery, facilitating rapid deployment in emergency scenarios like the 2023 Türkiye earthquake. Results indicate that integrating geospatial features significantly improves detection accuracy and the model's ability to generalize to unseen urban areas. However, a notable limitation is that SAR imagery is inherently noisy (speckle noise) and subject to geometric distortions, which can complicate the identification of fine structural cracks. The reliance on high-quality auxiliary geospatial data also means performance may vary in regions where such data is incomplete or unavailable.

Ho et al. [13] The researchers developed FloodDamageSense, a framework utilizing a multimodal Mamba backbone and multitask learning to assess building-level flood damage. The methodology fuses pre- and post-event SAR/InSAR scenes with optical basemaps and historical flood-risk layers to predict damage states and floodwater extent. Testing on Hurricane Harvey data showed a mean F1 score improvement of 19% over baseline models, particularly in the difficult-to-classify "minor" and "moderate" damage categories. Despite these gains, a limitation is that the model's precision remains sensitive to the native resolution of SAR data, which can be coarser than optical imagery. Additionally, while the multitask approach improves consistency, the complexity of the Mamba architecture requires substantial computational resources for training

Alisjahbana et al. [14] presented a deep-learning framework for multi-disaster building damage assessment using pre- and post-disaster satellite imagery. The researchers utilized the xBD dataset, which includes over 160,000 buildings from various disasters like hurricanes and earthquakes, and implemented a two-step methodology involving a Semantic Segmentation CNN (ResNet-50 FPN) for building footprint detection followed by a Damage Classification CNN (twin-tower ResNet-50). Their best model achieved an overall F1 score of 0.66, significantly outperforming the xView2 challenge baseline of 0.28, particularly when disaster-type labels were included as an additional feature. However, the study noted limitations such as difficulty in identifying very small buildings, poor performance on images with differing nadir angles or lighting, and a struggle with visual similarities between different damage levels. Ultimately, the research highlights that while automated segmentation is highly accurate, accurate damage classification remains challenging across diverse disaster types without incorporating prior estimates like disaster severity or region vulnerability

Wang et al. [15] curate the DisasterM3 dataset, a multi-hazard, multi-sensor, and multi-task remote sensing vision-language benchmark for global disaster assessment. The methodology involves collecting 26,988 bi-temporal optical and SAR image pairs across 36 historical events and generating 123k instruction pairs for tasks like damage assessment and report generation. They fine-tuned models like Qwen2.5-VL, InternVL3, LISA, and PSALM, achieving significant results such as a 10.4% increase in QA and a 40.8% gain in referring segmentation. Limitations include a lack of multi-resolution data, single-polarization SAR imagery, and performance biases in high-density building counting

Liu et al. [16] addresses Open-Set Semi-Supervised Object Detection (OSSOD) in remote sensing images to handle uncurated unlabeled data containing out-of-distribution (OOD) objects. The methodology introduces a Class-wise Feature Bank (CFB) to store in-distribution features, calculates OOD scores

via K-nearest neighbors, and applies an adaptive threshold to filter noisy pseudo-labels. Using a Faster-RCNN base model with a student-teacher framework, the approach was tested on DIOR and DOTA datasets. Here , Results show superior performance over existing methods, achieving an AP of 46.1 on DIOR Split 1. A key limitation is the reliance on in-distribution labeled data to define the feature space.

Zhang et al. [17] The authors propose SSOD-AT, a teacher-student framework designed to boost semi-supervised object detection in remote sensing images using active learning. The methodology introduces a RoI Comparison Module (RoICM) to identify high-confidence pseudo-labels and a Global Class Prototype to ensure sample diversity by suppressing redundant categories. Using the DOTA and DIOR datasets, the model employs a Faster-RCNN with ResNet-50 backbone. Results show it outperforms state-of-the-art methods, achieving a 1% mAP improvement in most cases and successfully mitigating the impact of noisy regions of interest. A limitation remains as uncertainty-based selection is susceptible to category imbalance without the specific diversity criterion.

Wang et al. [18] conduct research on weakly-semi-supervised object detection (WSSOD) to reduce reliance on expensive bounding box annotations in remote sensing. The methodology utilizes a "teacher-student" framework (Group R-CNN) where a teacher model, trained on limited bounding boxes, generates pseudo-labels from point annotations to train a student model. They evaluate this on the FAIR1M and a custom wind turbine dataset containing 200,000 images across 40 countries. Results show that WSSOD models trained with 2-10x fewer bounding boxes can match or exceed the performance of fully supervised models, achieving over 70 mAP at 0.5 IoU. A key limitation noted is that using less than 10% of bounding box data on FAIR1M leads to highly unstable results due to class imbalance.

Deb et al. [19] proposed an encoder-decoder-based semantic segmentation framework to identify flooded regions in high-resolution UAV aerial images, specifically leveraging self-supervised features from the DINOv2 model without fine-tuning. The methodology involves fusing DINOv2 features with traditional U-Net and DeepLabV3 architectures at the encoder bottleneck, utilizing a custom weighted focal loss combined with dice loss to address severe class imbalance. Using the FloodNet dataset (2,343 UAV images with 10 classes), the study demonstrates that the DINOv2+U-Net model achieves a 54.26% mIoU, outperforming Vanilla U-Net (53.47%) and E-Net (39.84%). Results highlight that self-supervised features learned from natural (non-aerial) images generalize effectively to top-down aerial views, significantly reducing the reliance on labor-intensive manual annotations. A noted limitation is that while DINOv2 integration significantly boosted DeepLabV3, the performance gain for the U-Net variant over its vanilla version was less pronounced for specific minority classes

Li et al. [20] addresses open-vocabulary object detection (OVD) in aerial imagery by proposing CastDet, a CLIP-activated student-teacher framework designed to detect unseen categories without extensive new labels. The methodology employs an interactive self-learning mechanism where a student model (Faster R-CNN) is guided by a localization teacher (EMA-updated) and an external teacher (RemoteCLIP) to generate high-quality pseudo-labels. Experiments were conducted on aerial datasets including VisDroneZSD, DOTA, and NWPU VHR-10, using a subset of DIOR for unlabeled training. Results show significant performance, reaching 46.5% mAP on VisDroneZSD novel categories, outperforming existing state-of-the-art OVD methods by 21.0%. Limitations include the small scale and low category vocabulary of current aerial datasets compared to natural image datasets, which hinders the direct application of standard OVD techniques

Salluri et al. [21] and other authors developed an automated object detection system using Transfer Learning to identify natural disaster aftermaths like floods and earthquakes. Using the DISASTER dataset (2,423 images), they fine-tuned pre-trained ResNet50, VGG-16, and VGG-19 models by replacing the final layers with custom predictive layers. Experimental results showed that VGG-19 achieved the highest accuracy of 94.22% while also providing faster computation compared to the other models. A limitation of this study is that it focused only on two disaster types (earthquake and flood) using a relatively small dataset. Future work aims to improve accuracy by incorporating larger, more diverse datasets with thousands of images covering various disasters.

Munawar et al. [22] used UAV-based aerial imagery and a CNN model inspired by AlexNet to detect floods and extract features for damage assessment. The methodology involved training the model on 2,150 image patches (resized to 32x32 pixels) created from three-channeled RGB pre- and post-disaster images, effectively forming a six-channel input to improve precision. The dataset included both global flood-prone regions and a specific study area in the Indus River region of Pakistan, utilizing publicly available images from sources like Google Earth. Experimental results demonstrated that the system was highly effective, achieving a flood detection accuracy of 91% and significant improvements in precision compared to supervised counterparts. However, the study's limitations include a reliance on the Caffe framework.As it is fast , It may require more modern updates for complex real-time applications across diverse global terrains.

Chang et al. [23] focuses on using the YOLOv4 (You Only Look Once) deep learning model to automate the counting of firefighters and fire trucks to enhance fireground personnel accountability. On the other hand methodology has training the object detector using a supervised learning approach to recognize specific features of emergency personnel and equipment. They utilized two datasets: one sourced from the internet (Dataset 1) and another from on-site fire departments in Taiwan (Dataset 2). The results demonstrated high performance with the on-site dataset.To reach over 96% accuracy and 97% precision/recall for identifying firefighters. A key limitation is the model's sensitivity to environmental conditions which has low light or thick smoke.It can obscure targets and reduce image resolution.

Wu et al. [24] proposes a Siamese neural network with an attention-based U-Net to simultaneously localize buildings and classify damage levels from pre- and post-disaster remote sensing imagery. However In methodology part , it utilizes a soft attention mechanism to emphasize effective spatial features while suppressing useless information across different backbones.It includes ResNet-34 and SEResNeXt. Also , Researchers used the large-scale xBD dataset for training, supplemented by Maxar Open Data for transferability testing, and applied data balancing strategies like cost-sensitive loss to address class imbalance.The Results shows that the SEResNeXt model with attention achieved the best single-model performance with an F1 score of 0.787, which improved to 0.792 after fusing multiple results. Furthermore , A primary limitation noted is that standard U-Net variants can suffer from a loss of detailed spatial characteristics, which the authors attempted to mitigate using global features and attention gates

Mei et al. [25] proposed the D2ANet, a difference-aware attention network for simultaneous building localization and multi-level change detection from dual-temporal satellite imagery.On the other hand , methodology utilizes a ResNet-101 encoder and dual decoders, incorporating a Difference-Aware Attention (D2A) block composed of a Dual-Temporal Aggregation (DTA) module for global change patterns and a Difference-Attention (DA) module to capture local dependencies between changes.However, the Experiments were conducted on the large-scale xBD dataset, which contains pre- and post-disaster imagery from 19 different natural disasters with four levels of building damage. Here, the model achieved state-of-the-art results which includes a building localization F1 score of 84.78% and a damage detection F1 score that outperformed existing methods across all damage scales. Furthermore , A potential limitation is the challenge of the heavily imbalanced distribution of damage scales in the dataset, which the evaluation metric specifically penalizes.

Sarosa et al. [26] utilized the You Only Look Once (YOLOv3) model to detect natural disaster victims via drone imagery to assist SAR teams in difficult terrain. Their methodology involved extracting single images, resizing them, and applying a Convolutional Neural Network (CNN) to predict bounding boxes and class probabilities. The research used a dataset of 200 images, split equally into 100 training and 100 testing samples. The results at 3,000 epochs achieved an accuracy of 89%, a precision of 90.82%, and a recall of 97.8%. Nevertheless , their limitations noted include the influence of object background, varied positions, and camera distance on detection accuracy.

Zhan et al.[27] developed an automated system to extract buildings and assess damage levels from post-event aerial images of the 2016 Kumamoto earthquake. Their methodology utilized a modified Mask R-CNN model featuring an improved feature pyramid network (PANet), online hard example mining (OHEM), and a multi-class non-maximum suppression (NMS) algorithm. The dataset comprised five high-resolution (10 cm/pixel) aerial images of Mashiki Town captured two weeks after the mainshock, manually labeled into four damage categories (no damage, slight, severe, and collapsed). However , results indicated high performance, achieving approximately 95% accuracy for building extraction, over 92% for detecting severely damaged buildings, and an 88% overall classification accuracy. However, its limitations included potential information loss in traditional feature proposal networks and difficulties for the model to localize buildings of extreme sizes (very large or small) in standard configurations.

Chen et al.[28] developed DamFormer, an end-to-end Siamese Transformer framework for building damage assessment that addresses the limitations of CNNs in modeling non-local relationships. The methodology utilizes a weight-sharing Siamese Transformer encoder (based on SegFormer) to extract multi-level features from pre- and post-disaster images, followed by a multitemporal adaptive fusion module and a lightweight dual-task decoder. Using the large-scale xBD dataset, the model achieved state-of-the-art results, including an overall F1 score of 79.56% and a localization F1 of 86.38%. However, a key limitation noted is the difficulty in classifying middle-tier damage levels (such as minor vs. major damage) due to highly identical visual appearances and inherent sample imbalance in disaster datasets.

Chen et al. [29] The research introduces HRTBDA, a high-resolution transformer network designed for post-disaster building damage assessment using remote sensing imagery. The methodology employs a two-phase architecture: Phase I localizes buildings from pre-disaster images using a transformer-based segmentation network, while Phase II classifies damage levels using a Siamese-based fusion network with a proposed Cross-Spatial Fusion (CSF) module to aggregate multi-temporal features.Moreover, It evaluated on the xBD and xFBD datasets, the model achieved an F1-score of 86.0% for localization and 78.4% for damage assessment. Furthermore it improved minor damage detection by 4.8%. However, a limitation is noted in the lower recall rate for the ”Destroyed” category, caused by the extreme variability in appearance, such as debris distribution and structural collapse patterns. The model utilizes a hybrid framework integrating HRNet with Swin Transformer blocks and a Depthwise Convolutional Multilayer Perceptron (DCMLP) to enhance global feature perception while reducing computational costs.

Summary: Across these works, a core gap is limited generalization across disaster types, regions, and sensors, with many models overfitting to specific events, modalities (optical-only or SAR-only), or resolutions. Label scarcity and noise remain persistent bottlenecks—SSL and pseudo-labeling help, but performance is still sensitive to thresholding, augmentation noise, and class imbalance, especially for subtle or mid-level damage. Many approaches rely on costly data sources (UHRA, UAVs, dense annotations, auxiliary GIS layers), limiting scalability and real-time deployment in low-resource regions. Architecturally, while transformers and foundation models improve representation, they introduce high computational overhead and test-time complexity, constraining operational use. Multimodal methods show strong promise, yet robust optical–SAR alignment, polarization diversity, and cross-modal fusion under

domain shift remain unresolved challenges. Future research should focus on domain-robust, label-efficient, multimodal foundation models that unify SSL, test-time adaptation, and uncertainty-aware learning, validated under realistic zero-shot, multi-disaster, and resource-constrained settings.

III Methodology

Dataset : The experiments in this study are conducted using the Disaster Response Object Detection Dataset, a publicly available dataset designed for object detection in real-world disaster scenarios [30]. The dataset consists of approximately 4,000+ images, capturing diverse emergency situations such as fires, smoke, damaged infrastructure, vehicles, and human presence in disaster-affected environments. The dataset contains six object classes, representing critical entities relevant to disaster response and emergency management. Each image is annotated with bounding box labels, enabling supervised object detection tasks. All annotations are provided in YOLO format, where each object instance is represented as a five-dimensional vector:

$$\mathbf{b} = (c,\ x_c,\ y_c,\ w,\ h) \tag{1}$$

where c denotes the class identifier, x_c and y_c represent the normalized center coordinates of the bounding box, and w and h correspond to the normalized width and height of the bounding box, respectively. All spatial values are normalized to the range $[0, 1]$ with respect to the image dimensions.This annotation format is natively supported by modern YOLO-based detectors and enables efficient training and inference. The dataset is organized into predefined training, validation, and test splits, following standard object detection practices. A data.yaml configuration file specifies the class names and directory paths, ensuring seamless integration with YOLOv10, YOLOv11, and YOLOv12 architectures. The YOLO-based annotation structure allows flexible experimentation with fully supervised, semi-supervised, and self-supervised learning paradigms. In semi-supervised learning experiments, a subset of labeled training images is retained while the remaining images are treated as unlabeled. For self-supervised learning, labels are entirely ignored during pretraining, and the full image set is used to learn robust visual representations.

In the figure 1 , describes workflow which begin with the Disaster Response Dataset, followed by data preprocessing to prepare images for training various YOLO models (YOLOv10, YOLOv11, and YOLOv12).After that got best pt file from yolov11 .Furthermore , with best-performing pretrained (.pt) model is selected for further optimization.Then ,This is used in semi-supervised training to generate pseudolabels, enabling self-supervised fine-tuning with advanced models like DINO and BYOL. Finally, a comprehensive comparative analysis evaluates the performance across all approaches.

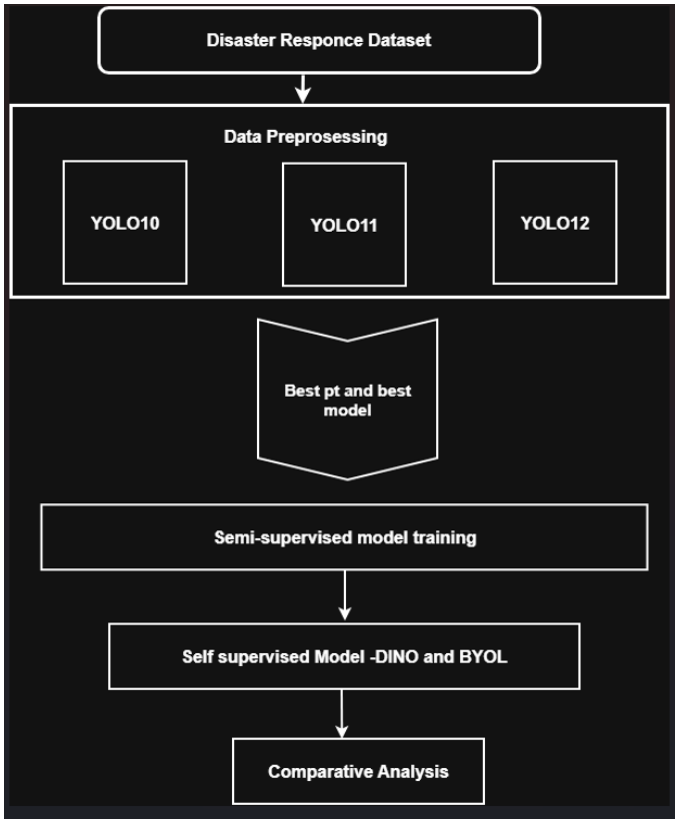

Fig. 1. Overall Methodology for disaster response object detection using YOLO variants and self-supervised learning.

Models : A standard fully supervised learning model trained using labeled data, used as a reference point to evaluate the effectiveness of new models, methods, or improvements.In baseline supervised model,helps to detect how well the task can be solved using conventional supervised learning.Also it allows to quantify performance gains attributable to new methodology, not data or scale.in this work, as a baseline model Yolov11s are used comparing analysis with Yolov10 , Yolov12 . Yolo is widely accepted object ditection framework .It has prediction ability with bounding box and classes in one forward pass with ensuring low latency.Moreover , It also provides good accuracy speed latency in maP avoiding biased baseline.Furthermore,it has scaled architecture that allows multiple varients.

In the figure 2 , it presents the detailed architecture of YOLOv11 adapted for mask segmentation in disaster response object detection. The **Backbone** uses C3K2 blocks (efficient CSPNet variant) with 3x/6x downsampling and shortcut connections (True/False variants). The **Neck** employs a PAN (Path Aggregation Network) structure with SPFF (Selective Pyramid Feature Fusion), C2PSA (Cross-stage Partial with Spatial Attention), and multi-scale feature fusion via concatenation and upsampling. The **Head** produces mask outputs at three scales through decoupled segmentation branches. Key components include Conv layers with (kernel$\times$stride, min channels,

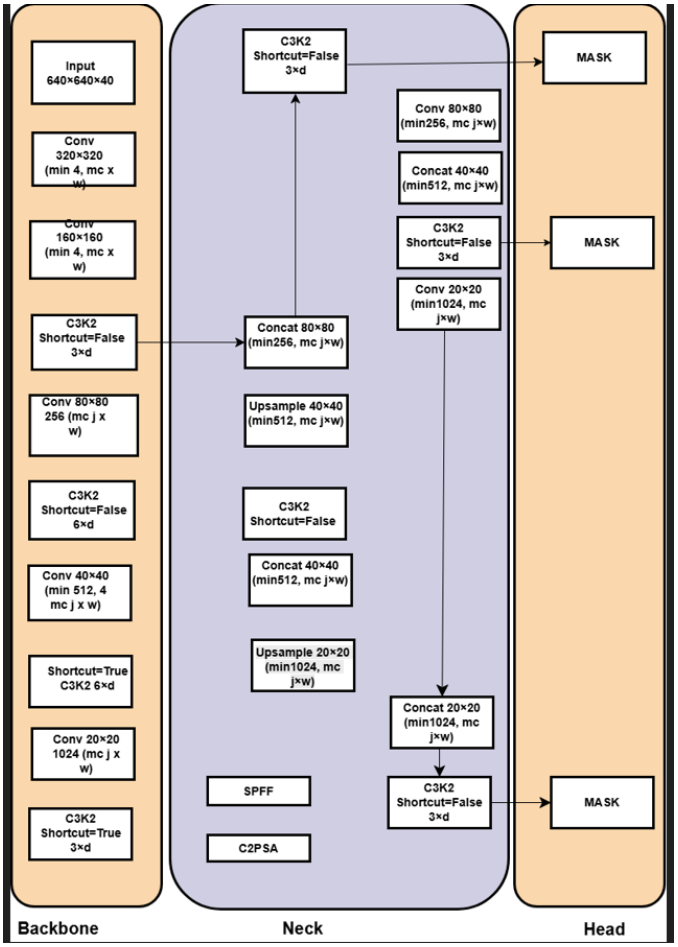

Fig. 2. Architecture of YOLOv11 showing Backbone (C3K2-based CSPNet), Neck (PAN with SPFF and C2PSA), and Head (uncoupled mask segmentation branch) [31].

activation), ensuring efficient multi-scale feature processing from 640×640 input to final mask predictions[32].

The performance of YOLOv10n, YOLOv11s, and YOLOv12n was evaluated under the same conditions to form a baseline for supervised object detection. The results showed that the YOLOv11s model performed better than the other two models concerning mAP@0.5 and mAP@0.5:0.95 as well as having a good balance between precision and recall. Therefore, YOLOv11s has been determined to be the baseline model because of its combination of representational ability, localisation accuracy and generalisation performance.

In this work used an STAC-based semi-supervised framework for object detection, where we incorporate consistency regularization and a teacher-student approach. Accordingly, our teacher model, which has been previously trained on labeled images of the disaster, produces high-quality pseudo labels, with a high degree of certainty, for all unlabeled images. Utilizing strong data augmentation strategies (geometric transformations, color distortions, etc.) as means of creating uncertainty, we then impose the application of consistency

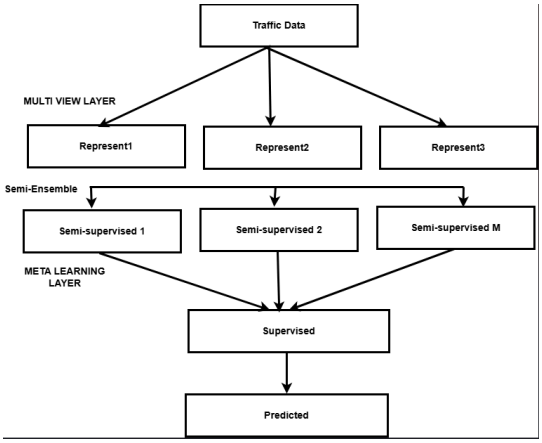

Fig. 3. Proposed multi-view semi-supervised ensemble architecture with meta-learning for traffic data prediction. The process starts with traffic data, generates multiple representations (Represent1 to Represent3) in the Multi-View Layer, applies a Semi-Ensemble to produce semi-supervised models (Semi-supervised 1 to M), integrates them via a Meta Learning Layer, and finally produces Supervised predictions leading to the final Predicted output.

regularization through training the student model so that its predictions will be consistent with the two corresponding teacher pseudo labels.A combined loss function is used to train, in which a supervised loss (calculated using the classification and localization) and an unsupervised loss (calculated from unlabeled data using pseudo-labels) are used together. To control how much unsupervised loss factors into training, a weighting factor has been created. The weight factor allows for more effective use of unlabeled data while maintaining stability and improving generalization.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{supervised}} + \lambda_u \cdot \mathcal{L}_{\text{unsupervised}} \tag{2}$$

where $\mathcal{L}_{\text{supervised}}$ denotes the detection loss computed on labeled samples, $\mathcal{L}_{\text{unsupervised}}$ represents the loss computed using pseudo-labeled unlabeled samples, and λ_u is a weighting factor that balances the contribution of supervised and unsupervised learning.

In the figure 3 ,it depicts the overall architecture of the proposed multi-view semi-supervised ensemble model for traffic data prediction. The input "Traffic Data" is processed through the **Multi-View Layer** to extract diverse representations (Represent1, Represent2, Represent3). These representations feed into the **Semi-Ensemble** stage, generating multiple semi-supervised learners (Semi-supervised 1 to Semi-supervised M). The outputs are then combined in the **Meta Learning Layer** to form a supervised model, ultimately yielding the final "Predicted" result [33].

In **Self-Supervised Learning** frameworks like **DINO** (Distillation with NO Labels) and **BYOL** (Bootstrap Your Own Latent), the pretext task revolves around contrastive learning paradigms designed to extract meaningful representations from unlabeled images without explicit supervision. For DINO, the approach employs a self-distillation mechanism where a student network learns to match the output probability distributions of a momentum-updated teacher network across multiple

augmented views of the same image, incorporating techniques such as multi-crop augmentation, centering to prevent mode collapse, and sharpening for one-hot-like outputs; this contrastive nature implicitly encourages the model to distinguish positive pairs (different views of the same image) from implicit negatives through knowledge distillation, fostering invariance to transformations. Similarly, BYOL utilizes a contrastive setup without negative samples, where an online network predicts the representations produced by a target network (updated via exponential moving average) for two augmented views of the same input, leveraging a predictor head and L2-normalized regression loss to bootstrap latent representations and avoid trivial solutions. Pretraining in these models involves optimizing the backbone architecture, such as a Vision Transformer or ResNet, on a large unlabeled dataset through these pretext objectives over numerous epochs, enabling the network to learn robust, generalizable features like edges, textures, and semantic structures. Following pretraining, fine-tuning on the object detector integrates the learned backbone into a downstream supervised task, such as attaching it to a detection head (e.g., in Faster R-CNN or YOLO frameworks) and training end-to-end or linearly probing on labeled detection datasets, where the pretrained weights initialize the feature extractor to accelerate convergence, improve generalization, and achieve higher performance metrics like mean average precision compared to random initialization, particularly in low-data regimes.

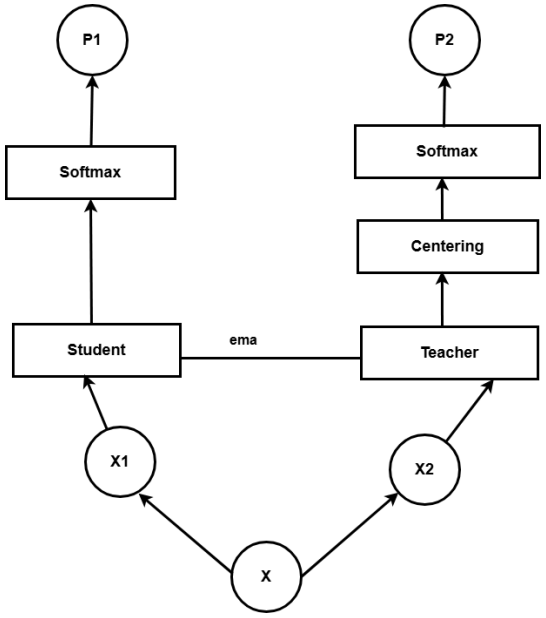

Fig. 4. Student-teacher architecture in self-supervised knowledge distillation (e.g., DINO). The input image x is augmented into two views x_1 and x_2. Both views are processed by the student network, while only x_2 (and sometimes additional crops) passes through the teacher network. The teacher is an exponential moving average (EMA) of the student weights. Predictions P_1 and P_2 from student and teacher are sharpened using softmax with temperature, and optionally centered (in DINO). The loss minimizes cross-entropy between the two distributions to align student and teacher representations.

Training SetUp :

In the **DINO pre-training** experiment, a Vision Transformer base model (ViT-B/16) was pre-trained using the DINO self-supervised learning framework with a LARS optimizer at a base learning rate of 9.0, linearly scaled according to the square root of the batch size relative to 1536 (i.e., $\sqrt{\text{bs}}/1536$). The training used a batch size of 32 over 40 epochs. Data augmentations included strong color jittering with brightness=0.8, contrast=0.8, saturation=0.4, hue=0.2, and application probability=0.8; Gaussian blurring with sigma uniformly sampled from [0.0, 0.1] and applied with probability=1.0; grayscale conversion with probability=0.2; horizontal flipping with probability=0.5; random resized cropping with scale range 0.14–1.0; and standard normalization. Key self-supervised parameters involved a teacher network based on dinov3/vitb16, utilizing the last 2 teacher blocks, a single projection layer, a projection hidden dimension of 2048, teacher momentum of 0.9, and weight decay of 1e-6.

For **DINO fine-tuning**, the pre-trained model was supervised fine-tuned with an initial learning rate of 0.005, a batch size of 16, and training for 100 epochs with patience-based early stopping after 10 epochs. Augmentations consisted of HSV adjustments (hue=0.015, saturation=0.7, value=0.4), horizontal flipping with probability=0.5, small random translation up to 0.1, scaling up to 0.5, random erasing with probability=0.4, and full RandAugment; no specific self-supervised parameters were applied during this supervised stage.

In the **BYOL pre-training** experiment, the same Vision Transformer base was pre-trained using the BYOL self-supervised method with an AdamW optimizer at a learning rate of 5e-4 and a small batch size of 8 over 40 epochs. Augmentations included random resized cropping to 160×160 with scale range (0.2, 1.0), horizontal flipping, color jittering (brightness=0.3, contrast=0.3, saturation=0.3, hue=0.1), grayscale with probability=0.2, and Gaussian blurring with kernel size 7 and sigma from (0.1, 2.0). Key self-supervised parameters featured a base predictor momentum of 0.99 scheduled cosinely toward 1.0 and weight decay of 1e-4.

Furthermore , In **BYOL fine-tuning** followed the same supervised setup as DINO fine-tuning, using an initial learning rate of 0.005, batch size of 16, 100 epochs with 10-epoch patience, and the same augmentation suite (HSV adjustments with h=0.015, s=0.7, v=0.4; horizontal flip p=0.5; translation=0.1; scale=0.5; erase=0.4; RandAugment), with no additional self-supervised components active [34] [35] .

The **STAC (Self-Training with Augmented Consistency)** experiment, as implemented in the provided Jupyter notebook, uses the following key hyperparameters.

- The learning rate follows the default Ultralytics YOLO value of 0.01 (initial lr0).
- The batch size is set to 8.
- Each student model trains for 20 epochs per iteration.
- Data augmentations rely on YOLOv11s standard settings,

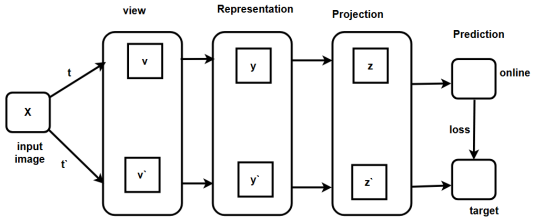

Fig. 5. Architecture of BYOL (Bootstrap Your Own Latent). The online network predicts the target network's projection from different augmented views of the same image, with loss applied between prediction and target [36].

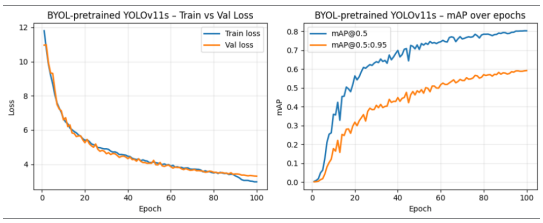

Fig. 6. Comparison of self-supervised training loss curves for BYOL (left) and DINO (right).

including HSV adjustments, flips, rotations, mosaic, and other techniques like translation, scaling, and shearing.
- The pseudo-label confidence threshold for semi-supervised learning is 0.7.
- The input image size is 640x640.
- The process runs for 5 STAC iterations.

IV Results and Experiments

For experimental setup, we have used Kaggle GPU T4 X 2.

TABLE I
PERFORMANCE COMPARISON OF YOLO VARIANTS

Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1 Score
YOLOv10n	0.792	0.627	0.787	0.779	0.783
YOLOv11s	**0.824**	**0.651**	**0.804**	0.794	0.799
YOLOv12n	0.821	0.646	**0.804**	**0.799**	**0.801**

In the table 1, the performance comparison of YOLO object detection variants reveals that YOLOv11s (small) and YOLOv12n (nano) significantly outperform YOLOv10n (nano) across key metrics like mAP@0.5 (up to 0.824 vs. 0.792), stricter mAP@0.5:0.95, precision, recall, and F1 score. These gains highlight improved localization and detection reliability in newer models. Additionally, DINO self-supervised pretraining substantially boosts final test performance over BYOL (e.g., mAP@0.5: 0.825 vs. 0.737), with comparable inference speed and parameters, underscoring DINO's superior feature transfer for downstream tasks.

TABLE II
FINAL TEST PERFORMANCE COMPARISON OF SSL-PRETRAINED MODELS

Metric	DINO-Pretrained	BYOL-Pretrained
mAP@0.5	**0.8252**	0.7969
mAP@0.5:0.95	**0.6448**	0.5858
Precision	0.8449	**0.8547**
Recall	**0.7948**	0.7100
Inference Time (ms/img)	**8.1**	8.2
Parameters (M)	9.42	9.42

In the table 2 , comparison of two pre-training approaches on the same YOLO-based object detection model shows that DINO-pretrained weights substantially outperform BYOL-pretrained ones. Using self-supervised learning on large unlabeled datasets like ImageNet, DINO (self-distillation without labels) delivers stronger transferable features than BYOL (Bootstrap Your Own Latent). The DINO variant achieves approximately 9% higher mAP@0.5, 6% higher mAP@0.5:0.95, and superior precision and recall, reducing false positives and missed detections. With nearly identical inference time and parameter count, these gains stem purely from DINO's more robust visual representations for downstream tasks.

V Discussion

TABLE III
OVERALL PERFORMANCE COMPARISON

Model Variant	mAP@0.5	mAP@0.5:0.95	Precision	Recall
Baseline YOLOv11s	0.8240	0.6510	0.8040	0.7940
DINO-pretrained YOLOv11s	0.8252	0.6448	0.8449	0.7948
BYOL-pretrained YOLOv11s	0.7969	0.5858	0.8547	0.7100

From the table 3, the results suggest that DINO pretraining transfers more effectively to downstream object detection than BYOL in this setting, likely due to its stronger spatial feature alignment. The precision–recall trade-off observed in BYOL indicates a tendency toward overconfident but sparse predictions, which is suboptimal for disaster-response scenarios where missing objects is costly. Overall, the findings highlight that SSL benefits are method-dependent, and careful selection of pretraining strategy is crucial to balance localization accuracy and detection coverage. Furthermore , all three models share the same YOLOv11s architecture and inference speed, implying that SSL pretraining does not introduce additional runtime overhead during deployment. However, SSL methods incur higher pretraining cost, with DINO generally being more computationally intensive than BYOL due to multi-crop strategies and larger batch requirements. Practically, DINO-pretrained YOLOv11s offers the best accuracy–cost trade-off for real-time disaster monitoring, while BYOL may be preferred in resource-constrained training settings where higher precision is desired but some loss in recall is acceptable.

VI Conclusion

Among the evaluated models, YOLOv11s with DINO pretraining stands out as the most effective combination. It achieves the highest mAP@0.5 while maintaining recall similar to the fully supervised baseline. This shows that DINO-learned representations transfer well to object detection. They provide a good balance between detection accuracy and robustness without raising inference costs.On the other hand , the results show that self-supervised pretraining can improve downstream detection performance in both labs, but the SSL approach has a significant impact on the improvements. While Lab 2 demonstrates that not all SSL techniques translate equally well to dense prediction tasks—BYOL increases precision but decreases recall, while DINO maintains spatial awareness crucial for detection—Lab 1 emphasizes the significance of robust visual representations learnt from unlabeled input. Moreover, future developments could investigate domain-specific pretraining on extensive disaster images, hybrid SSL techniques that combine contrastive and clustering-based aims, and larger backbones (like YOLOv11m/l) with SSL pretraining with more datasets. The framework's practical usefulness would further be strengthened by adding semi-supervised fine-tuning, uncertainty-aware detection, and evaluation under real-time deployment limitations.

References

[1] Wikipedia contributors, "You only look once," https://en.wikipedia.org/wiki/You_Only_Look_Once, 2025, accessed: September 5, 2025.

[2] A. Ciocarlan, S. Lefebvre, S. L. Hégarat-Mascle, and A. Woiselle, "Self-supervised learning for real-world object detection: a survey," 2024. [Online]. Available: https://arxiv.org/abs/2410.07442

[3] T. Shehzadi, Ifza, D. Stricker, and M. Z. Afzal, "Semi-supervised object detection: A survey on progress from cnn to transformer," 2024. [Online]. Available: https://arxiv.org/abs/2407.08460

[4] J. Lee, J. Z. Xu, K. Sohn, W. Lu, D. Berthelot, I. Gur, P. Khaitan, Ke-Wei, Huang, K. Koupparis, and B. Kowatsch, "Assessing post-disaster damage from satellite imagery using semi-supervised learning techniques," 2020. [Online]. Available: https://arxiv.org/abs/2011.14004

[5] D. K. Singh and V. Hoskere, "Post disaster damage assessment using ultra-high-resolution aerial imagery with semi-supervised transformers," *Sensors*, vol. 23, no. 19, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/19/8235

[6] Y. Almalioglu, A. Kucik, G. French, D. Antotsiou, A. Adam, and C. Archambeau, "Sar object detection with self-supervised pretraining and curriculum-aware sampling," 2025. [Online]. Available: https://arxiv.org/abs/2504.13310

[7] X. Zi, T. Zhu, Y. Shi, X. Tao, J. Li, and M. Prasad, "Assessing property damage from natural disasters: Self-supervised and supervised learning models for remote sensing imagery analysis," *SSRN Electronic Journal*, January 2024. [Online]. Available: https://ssrn.com/abstract=4699455

[8] H. Chen, J. Song, O. Dietrich, C. Broni-Bediako, W. Xuan, J. Wang, X. Shao, Y. Wei, J. Xia, C. Lan, K. Schindler, and N. Yokoya, "BRIGHT: a globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response," *Earth System Science Data*, vol. 17, no. 11, pp. 6217–6253, 2025. [Online]. Available: https://essd.copernicus.org/articles/17/6217/2025/

[9] K. Ahn, S. Han, S. Park, J. Kim, S. Park, and M. Cha, "Generalizable disaster damage assessment via change detection with vision foundation model," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 39, no. 27, 2025, pp. 27 784–27 792. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/34994

[10] L. Barco, A. Urbanelli, and C. Rossi, "Rapid wildfire hotspot detection using self-supervised learning on temporal remote sensing data," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 2061–2065.

[11] M. Rahnemoonfar, T. Chowdhury, and R. Murphy, "Rescuenet: A high resolution uav semantic segmentation dataset for natural disaster damage assessment," *Scientific Data*, vol. 10, no. 1, p. 913, December 2023. [Online]. Available: https://www.nature.com/articles/s41597-023-02799-4

[12] L. Russo, D. Tapete, S. L. Ullo, and P. Gamba, "A deep learning framework for building damage assessment using vhr sar and geospatial data: demonstration on the 2023 turkiye earthquake," 2025. [Online]. Available: https://arxiv.org/abs/2506.22338

[13] Y.-H. Ho and A. Mostafavi, "Flood-damagesense: Multimodal mamba with multitask learning for building flood damage assessment using sar remote sensing imagery," 2025. [Online]. Available: https://arxiv.org/abs/2506.06667

[14] I. Alisjahbana, J. Li, Ben, Strong, and Y. Zhang, "Deepdamagenet: A two-step deep-learning model for multi-disaster building damage segmentation and classification using satellite imagery," 2024. [Online]. Available: https://arxiv.org/abs/2405.04800

[15] J. Wang, W. Xuan, H. Qi, Z. Liu, K. Liu, Y. Wu, H. Chen, J. Song, J. Xia, Z. Zheng, and N. Yokoya, "Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response," 2025. [Online]. Available: https://arxiv.org/abs/2505.21089

[16] N. Liu, X. Xu, Y. Gao, Y. Zhao, and H.-C. Li, "Semi-supervised object detection with uncurated unlabeled data for remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 129, p. 103814, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843224001687

[17] B. Zhang, Z. Wang, and B. Du, "Boosting semi-supervised object detection in remote sensing images with active teaching," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[18] J. H. Wang, J. Irvin, B. K. Behar, H. Tran, R. Samavedam, Q. Hsu, and A. Y. Ng, "Weakly-semi-supervised object detection in remotely sensed imagery," 2023. [Online]. Available: https://arxiv.org/abs/2311.17449

[19] D. Deb and U. Verma, "Leveraging self-supervised features for efficient flooded region identification in uav aerial images," 2025. [Online]. Available: https://arxiv.org/abs/2507.04915

[20] Y. Li, W. Guo, X. Yang, N. Liao, D. He, J. Zhou, and W. Yu, "Toward open vocabulary aerial object detection with clip-activated student-teacher learning," in *Computer Vision – ECCV 2024*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Springer, Cham, 2025, vol. 15144, pp. 431–448. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-73016-0_25

[21] D. Salluri, K. Bade, and G. Madala, "Object detection using convolutional neural networks for natural disaster recovery," *International Journal of Safety and Security Engineering*, vol. 10, no. 2, pp. 285–291, 2020.

[22] H. S. Munawar, F. Ullah, S. Qayyum, S. I. Khan, and M. Mojtahedi, "Uavs in disaster management: Application of integrated aerial imagery and convolutional neural network for flood detection," *Sustainability*, vol. 13, no. 14, p. 7547, 2021.

[23] R. H. Chang, Y.-T. Peng, S. Choi, and C. Cai, "Applying artificial intelligence (ai) to improve fire response activities," *Emergency Management Science and Technology*, vol. 2, no. 1, pp. 1–6, 2022.

[24] C. Wu, F. Zhang, J. Xia, Y. Xu, G. Li, J. Xie, Z. Du, and R. Liu, "Building damage detection using u-net with attention mechanism from pre-and post-disaster remote sensing datasets," *Remote Sensing*, vol. 13, no. 5, p. 905, 2021.

[25] J. Mei, Y.-B. Zheng, and M.-M. Cheng, "D2anet: Difference-aware attention network for multi-level change detection from satellite imagery," *Computational Visual Media*, vol. 9, no. 3, pp. 563–579, 2023.

[26] M. Sarosa, N. Muna, and E. Rohadi, "Detection of natural disaster victims using you only look once (yolo)," in *IOP Conference Series: Materials Science and Engineering*, vol. 1098, no. 3. IOP Publishing, 2021, p. 032076.

[27] Y. Zhan, W. Liu, and Y. Maruyama, "Damaged building extraction using modified mask r-cnn model using post-event aerial images of the 2016 kumamoto earthquake," *Remote Sensing*, vol. 14, no. 4, p. 1002, 2022.

[28] H. Chen, E. Nemni, S. Vallecorsa, X. Li, C. Wu, and L. Bromley, "Dual-tasks siamese transformer framework for building damage assessment," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 1600–1603.

[29] F. Chen, Y. Sun, L. Wang, N. Wang, H. Zhao, and B. Yu, "Hrtbda: a network for post-disaster building damage assessment based on remote sensing images," *International Journal of Digital Earth*, vol. 17, no. 1, p. 2418880, 2024.

[30] R. Majumdar, "Disaster response object detection dataset," https://www.kaggle.com/datasets/rupankarmajumdar/disaster-response-object-detection-dataset, 2025.

[31] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv11," https://docs.ultralytics.com/models/yolo11/, 2024, accessed: 2025-12-19.

[32] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," https://arxiv.org/abs/2410.17725, Oct. 2024, the paper containing the referenced figure (Fig. 1 on ResearchGate: publication ID 386467184).

[33] S. Li, L. Huang, J. Wang, and G. Zhou, "Semi-stacking for semi-supervised sentiment classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 27–31, contains Fig. 1: General Architecture of the Semi-Stack approach (ResearchGate figure ID: 341382158). [Online]. Available: https://aclanthology.org/P15-2005/

[34] afsinnn, "Self-supervised model (dino + byol)," https://www.kaggle.com/code/afsinnn/self-supervised-model-dino-byol, 2025, accessed: 2025-12-19.

[35] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9650–9660.

[36] S.-H. Tsang. (2022, Mar.) Review — BYOL: Bootstrap your own latent a new approach to self-supervised learning. Medium. [Online]. Available: https://sh-tsang.medium.com/review-byol-bootstrap-your-own-latent-a-new-approach-to-self-supervised-learning-6f770a624441