# A Benchmark Dataset for Spatially Aligned Road Damage Assessment in Small Uncrewed Aerial Systems Disaster Imagery

**Thomas Manzini*, Priyankari Perali*, Raisa Karnik, Robin R. Murphy**

Department of Computer Science, Texas A&M University
435 Nagle St, College Station, TX, United States, 77843
{tmanzini, perali, raisak, robin.r.murphy}@tamu.edu

## Abstract

This paper presents the largest known benchmark dataset for road damage assessment and road alignment, and provides 18 baseline models trained on the CRASAR-U-DRIODs dataset's post-disaster small uncrewed aerial systems (sUAS) imagery from 10 federally declared disasters, addressing three challenges within prior post-disaster road damage assessment datasets. While prior disaster road damage assessment datasets exist, there is no current state of practice, as prior public datasets have either been small-scale or reliant on low-resolution imagery insufficient for detecting phenomena of interest to emergency managers. Further, while machine learning (ML) systems have been developed for this task previously, none are known to have been operationally validated. These limitations are overcome in this work through the labeling of 657.25km of roads according to a 10-class labeling schema, followed by training and deploying ML models during the operational response to Hurricanes Debby and Helene in 2024. Motivated by observed road line misalignment in practice, 9,184 road line adjustments were provided for spatial alignment of a priori road lines, as it was found that when the 18 baseline models are deployed against real-world misaligned road lines, model performance degraded on average by 5.596% Macro IoU. If spatial alignment is not considered, approximately 8% (11km) of adverse conditions on road lines will be labeled incorrectly, with approximately 9% (59km) of road lines misaligned off the actual road. These dynamics are gaps that should be addressed by the ML, CV, and robotics communities to enable more effective and informed decision-making during disasters.

**Code** — www.github.com/CRASAR/CRASAR-U-DROIDs-RDA
**Data & Models** — https://huggingface.co/CRASAR

## Introduction

In response to a disaster, small uncrewed aerial systems (sUAS), also known as drones, are deployed to capture aerial imagery mapping impacted areas. This imagery is then assessed to inform emergency managers of where damage is. This information collected during the early stages of the response phase can enable informed decisions, such as appropriate allocation of aid, and effective navigation through impacted areas (Hargis, Rao, and Choset 2024; Alam et al. 2025). However, efforts are often hindered by resource constraints and wireless connectivity (Manzini, Murphy, and Merrick 2023; Manzini et al. 2023), arguing for Machine Learning (ML) and Computer Vision (CV) to automate assessments on hardware deployed within the disaster scene.

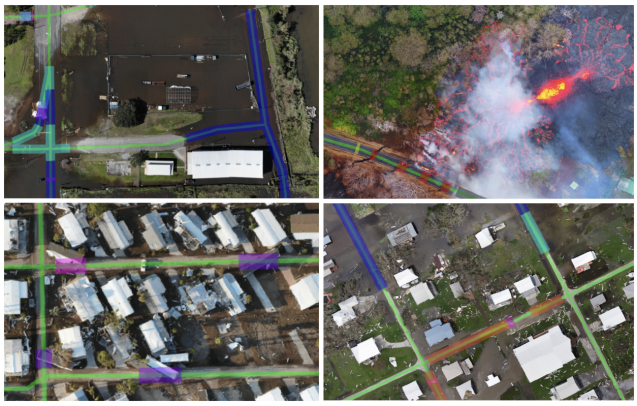

Figure 1: Figure of labeled roads within the CRASAR-U-DRIODs dataset. Road lines (Green) labeled as follows: Total Flooding (Dark Blue), Partial Flooding (Light blue), Total Obstruction (Purple), Partial Obstruction (Pink), Total Destruction (Red), Partial Road Condition (Yellow), Total Particulate (Dark Orange), Partial Particulate (Light Orange), and Not Able to Determine (Black).

This work advances the state of the art in CV/ML for sUAS post-disaster imagery by creating a labeled and spatially aligned dataset and baseline models of damaged roads, and operationally validates it with imagery collected by responders at Hurricanes Helene and Debby. Assessing road conditions is important to disaster response efforts because it enables informed decision-making concerning resource allocation, routing, and navigation of aid and evacuation. Despite its importance, prior CV/ML literature has overlooked automating post-disaster road condition assessment. Due to this, the field of CV/ML has not overcome three challenges in automating road damage assessments: limited diverse datasets ((Rahnemoonfar et al. 2021; Rahnemoonfar, Chowdhury, and Murphy 2023; Jiang et al. 2024; Pi, Nath,

*These authors contributed equally.

| Damage Label | Damage Label Description |
|---|---|
| Clear Road | Obviously clear, road lines clearly visible, and cars may be actively driving on the road surface |
| Partial Obstruction | Road partially covered by scattered debris, obstructing vegetation, piled cars, RVs, and boats |
| Partial Flooding | Standing water on road, road lines partially visible |
| Partial Road Condition | Road partially crumbling, moderate asphalt cracking |
| Partial Particulate | Road partially covered by particulates (e.g. sand, mud, lahar), road lines partially visible |
| Total Obstruction | Total coverage of the road by piles of debris, vegetation, or vehicles preventing transit |
| Total Flooding | Large quantities of standing water on the road; road lines are not visible |
| Total Destruction | Road absent or collapsed, substantial asphalt cracking |
| Total Particulate | Total coverage of the road by particulates (e.g. sand, mud, lahar); road lines are not visible |
| Not Able To Determine | Unable to determine due to obscuration |

Table 1: A simplified version of the 10 classes in the Road Damage Assessment Schema. Green corresponds to the "Road Line" class, orange corresponds to all the "Partial" classes, red corresponds to the "Total" classes, and blue corresponds to the "Not Able To Determine" class.

and Behzadan 2020; Hänsch et al. 2022)), no operational road damage assessment schema, and the lack of operational validation of road damage assessment models. Furthermore, this work reveals that road damage assessment via sUAS imagery is affected by non-uniform spatial misalignment of a priori road lines, a fundamental problem prevalent within sUAS imagery collected operationally (Manzini et al. 2024a, 2025). Such non-uniform misalignment is not seen in satellite imagery, and an evaluation of 18 baseline models reveals that if spatial alignment is not addressed, model performance degrades by 1.9 Macro IoU for the top model.

This work offers five contributions to the ML, CV, remote sensing, and emergency management communities.

1. The release of the largest known dataset, in terms of kilometers of roads, for road damage assessment in sUAS imagery, providing 657.25km of labeled roads.
2. The release of the first dataset addressing spatial alignment errors with a priori road lines, providing 9,184 adjustment annotations for spatial alignment.
3. The development of a practitioner-relevant schema for road conditions in disaster imagery, with consultation from federal and local agencies in the United States.
4. The release of baseline models that can be used within disaster response, evaluating and validating one baseline model operationally with imagery collected in response to Hurricanes Debby and Helene.
5. The identification of two critical challenges for future ML and CV efforts for road damage assessment concerning spatial alignment and the distribution shifts present in real-world road damage assessment aerial imagery data.

## Related Work

The remote sensing, civil engineering, ML, and CV communities have considered and developed datasets for road damage assessment and road condition schemas. However, these datasets are limited to specific road conditions, specific disasters, small data scales, and do not consider spatial misalignment errors with a priori road lines. Further, these datasets all use schemas that do not align with or capture the necessary road damage labels for disaster response, hindering the transfer of trained ML models to practice.

### Disaster Datasets for Road Damage Assessment

Prior literature contains eight known efforts to automatically assess road conditions using aerial imagery. These consist of three datasets of sUAS aerial imagery (Rahnemoonfar et al. 2021; Rahnemoonfar, Chowdhury, and Murphy 2023; Jiang et al. 2024), one dataset containing satellite aerial imagery (Hänsch et al. 2022), and one dataset with both sUAS and crewed aircraft imagery (Pi, Nath, and Behzadan 2020) with labels for flooded areas corresponding to roads. The remaining four efforts did not release data (Urabe and Saji 2007; Korkmaz and Poyraz 2016; Yang et al. 2020; Takyi et al. 2025) but considered sUAS and satellite imagery. Unfortunately, all lack practitioner-relevant labels, limiting use in disaster response.

The three sUAS-centric datasets, FloodNet (Rahnemoonfar et al. 2021) RescueNet (Rahnemoonfar, Chowdhury, and Murphy 2023), and EarthquakeNet (Jiang et al. 2024), are semantic segmentation datasets that provide road damage labels for whether the roads were: "flooded" or "non-flooded", "road clear" or "road blocked", and "no damage", "minor damage", or "severe damage", respectively. FloodNet provides labels for 31.58km of annotated road, consisting of 2.85km$^2$ and 28.12 gigapixels of imagery from Hurricane Harvey. RescueNet provides labels for 41.05km of road, consisting of 3.6km$^2$ and 53.93 gigapixels of imagery from Hurricane Michael. EarthquakeNet provides labels for approximately 1.4 gigapixels of imagery of roads from the 2013 Lushan Earthquake in Sichuan, China.

The remaining five efforts use similar schemas, "flooded" and "non-flooded" (Hänsch et al. 2022), "road blockage" and "no road blockage" (Yang et al. 2020; Urabe and Saji 2007), "open," "partially-open," and "undamaged," "slightly damaged," and "damaged" (Korkmaz and Poyraz 2016), and "debris," "flooded area," and "car" (Pi, Nath, and Behzadan 2020) for specific disasters and road conditions. The most diverse dataset among these is the Volan v.2018, with three classes and imagery from hurricanes Harvey, Irma, Maria, and Michael; it does not overcome the earlier limitations.

### Dedicated Road Condition Schemas

The civil engineering literature has provided a standardized schema, the "pavement condition index," for analyzing pavement condition, with four datasets (Majidifard et al. 2020; Sabouri et al. 2023; Ren et al. 2024; Yan and Zhang 2023) motivated to automate such analysis. The works in this area focus primarily on measuring the health of the pavement or estimating its lifespan; this is different from this work, which focuses on assessing which roads could be utilized immediately following a disaster.

More generally, the transportation disruption ontology provided by (Corsar et al. 2015) provides a comprehensive

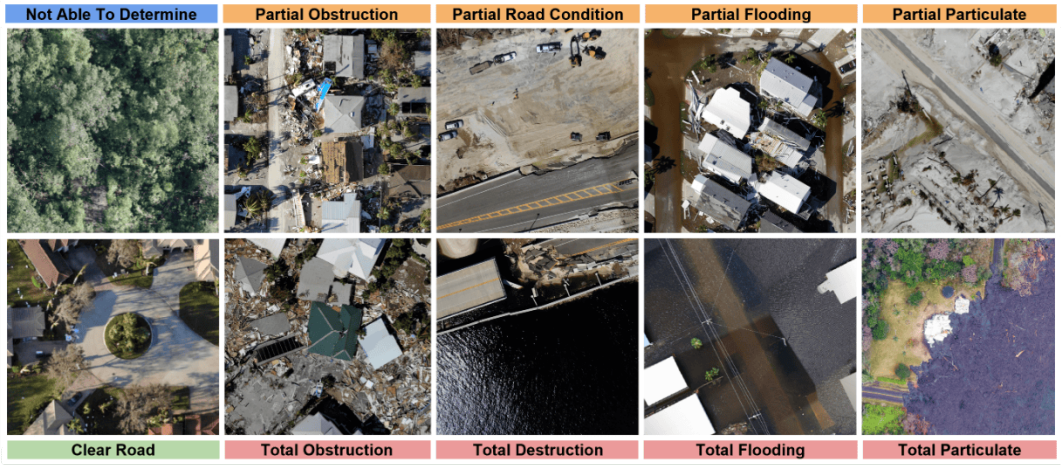


Figure 2: Visuals of the 10 Labels within the Road Damage Assessment Schema.

ontology of events that could disrupt planned travel. While this ontology does have entries that could be relevant to this effort, such as "Storm Damage," "Tornado," and "Forest Fire," these entries do not describe the conditions that are present on the road itself, nor do they describe visual features, only transportation disruptions, limiting the translation of such a schema to aerial imagery.

## Method

The method to address the three challenges, described earlier, consists of the development of a practitioner-relevant road damage schema for disaster aerial imagery, annotation of the CRASAR-U-DRIODs (Manzini et al. 2024b) imagery for road damage assessment, adjustment annotation for spatial alignment errors with the use of a priori road lines, and development of baseline models trained on this dataset.

### Road Damage Assessment

The road damage labels within this dataset were based on a schema developed with input from federal and state agencies in the United States, applied to imagery from the CRASAR-U-DRIODs dataset by a pool of 130 annotators, and reviewed through a two-stage process to reduce label noise.

**Imagery** The CRASAR-U-DROIDs was selected as the source of all imagery for this effort (Manzini et al. 2024b). The motivation for this was twofold. First, this dataset is the largest known collection of sUAS orthomosaic imagery, comprising 52 orthomosaics from 10 different federally declared disasters, collected with resolutions ranging from 12.7cm/px to 1.93cm/px. Ideally, this diverse data would enable performant and generalizable ML models. Second, this dataset contained a substantial variety of road conditions representative of real-world road conditions.

**Schema Formulation** The road damage assessment schema was developed with input from the United States Federal Emergency Management Agency (FEMA) and the Texas Department of Transportation and consists of 10 damage classes, visually shown in Figure 2. The purpose of this collaboration was to ensure that models derived from this data could provide practitioner-relevant labels.

This schema contains 10 classes, eight of which are split into two subcategories: "Partial" and "Total". The labels of "Road Line" and "Not Able To Determine" are individual labels without subcategories. Otherwise, roads can be assigned to the "Partial" class for roads that are only partially affected by an adverse condition or the "Total" class for roads that are completely affected by an adverse condition. There are four categories of adverse conditions, which can be either "Partial" or "Total," and combined, totaling the eight classes referenced above. These are "obstruction," "flooding," "particulate," and "road condition/destruction."

Obstruction describes discrete physical objects on the road's surface. Flooding describes water on the road's surface. Particulate denotes small particles that could potentially obscure conditions below them. Examples of particulates include beach sand, mud, or dirt. Lahar, while not a particulate, was intentionally folded into this class. Finally, road condition/destruction describes a condition of the road surface itself, and is either "Road Condition" or "Destruction" depending on the "Partial" or "Total" subclass. With "Partial Road Condition," the road may still be usable but has cracked or is crumbling away, whereas in "Total Destruction," the road has been destroyed and cannot be used.

**Annotation** The imagery within the CRASAR-U-DRIODs dataset was tiled, overlaid with a priori road lines, sourced from OpenStreetMap (OSM) (OpenStreetMap contributors 2024), and annotated by a pool of 130 annotators, within Labelbox(Labelbox 2024), according to the schema discussed earlier, resulting in 656.094km of road annotated. Each of the 52 orthomosaics within the dataset was tiled at a resolution of 2048x2048 (45 orthomosaics) and 8500x8500 (7 orthomosaics). While the pool of 130 annotators annotated the 2048x2048 tiles, the 8500x8500 tiles were annotated by the authors. A view of this process is shown in Figure 3. Though the annotators were non-experts, reviewers compensated for the lack of domain expertise.

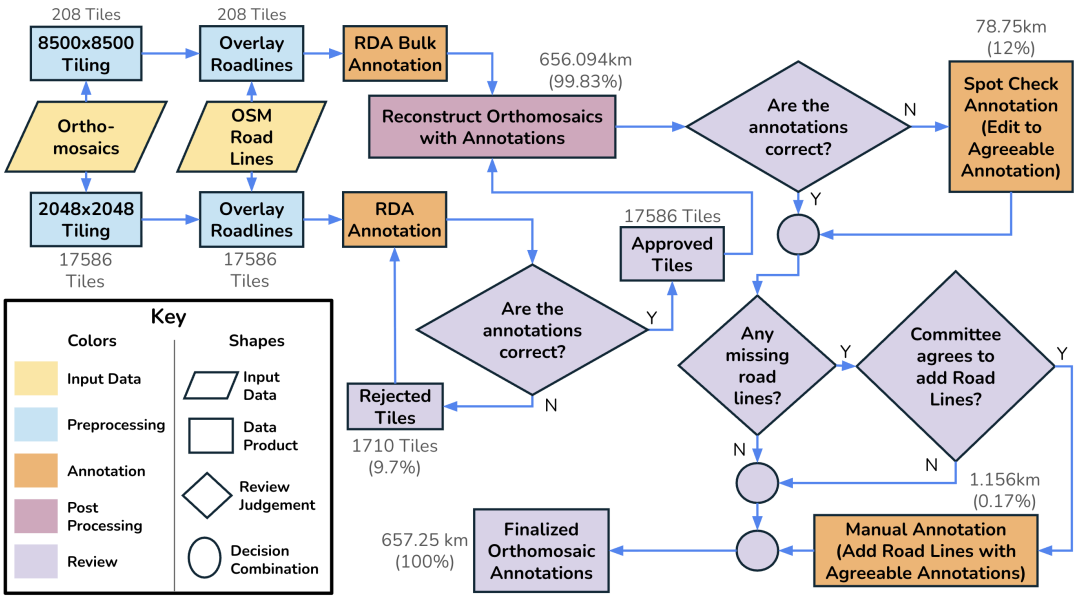


Figure 3: The visualized annotation workflow by which unlabeled imagery is labeled, reviewed, and processed.

**Annotation to Road Corridor Processing** The road condition annotation polygons provided by the annotators were often of inconsistent shape and, in some cases, did not correspond to any road lines. To make these annotation polygons more consistent and to align them with the road itself, the intersection between all road lines and all annotation polygons was computed. Then, new rectangular polygons were generated based on these intersections such that they were parallel with the road line and had a width of 7.2 meters, the reported upper limit on the width of two lanes of rural or urban road in the United States (Stein and Neuman 2007). Examples of these rectangular polygons are shown in Figure 1. This intentional decision was to first standardize the dimensions of the annotation polygon, so they corresponded to the dimensions of roads rather than the dimensions that annotators had drawn, and second, to have a dimension that could provide a consistent region to account for reasonable variations in spatial misalignment, discussed later on.

**Review** Following annotation, a two-stage review was conducted: first, a review by individual reviewers, and second, an orthomosaic review by a committee of reviewers. During the individual review, each tile was inspected by a single reviewer, and 1,710 tiles (9.7% of tiles) were rejected and re-annotated by annotators or corrected by the reviewer. Following this, the annotations were overlaid on orthomosaics and reviewed by a committee consisting of three of the authors and one external reviewer, where spot-check corrections were made. Reviewers collectively had experience working operationally at major disasters, working as an emergency vehicle operator (EVOC/CEVO), a sUAS data manager, and a sUAS pilot. Additionally, at the discretion of the reviewers, missing road lines were manually added and annotated. The spot-check corrections resulted in 78.5km (12%) of road line labels being changed.

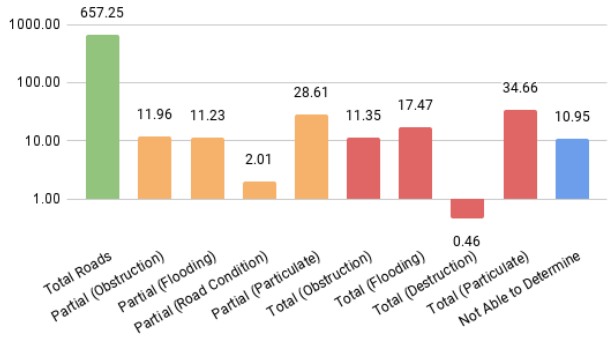


Figure 4: Kilometers of each label in the presented dataset. Note that the y-axis is on a log scale.

The manual addition of road lines, not sourced from OSM, resulted in 1.156km (0.17%) of road lines being added. Following this review, the distribution of labels shown in Figure 4 remained. This two-stage review process was employed to maximize consistency and minimize label noise, with the goal of creating a worthwhile benchmark for future models. An overview of this process can be found in Figure 3.

## Spatial Alignment

During the curation of the road damage assessment labels, spatial alignment errors with the a priori road lines were observed, an example is shown in Figure 5. Approximately 59km (9%) of road lines were misaligned off the actual road, and 11km (8%) had incorrect road damage assessment labels, presenting the need to correct the errors to avoid impeding downstream ML models trained on this data, resulting in 9,184 road line adjustment annotations.

Figure 5: Example of a road line (colored in green) that is unaligned with the source imagery. Red alignment vectors show the transform necessary to align each vertex.

This work aligns road lines using the same vector field formulation presented in (Manzini et al. 2024a), but differs by operating on individual road line vertices instead of building polygons. While the CRASAR-U-DRIODs dataset provides adjustment annotations for similar misalignment with a priori building polygons, these were generated for the building polygons, which were sourced from a different spatial resource than the road lines. As a result, additional adjustments were manually collected and curated for the road lines considered in this work, resulting in 9,184 total road line adjustments. As alignment of road lines acts on vertices, the aligned road lines differ from the misaligned by 396 meters across all 657.25km of road lines (0.06%).

### Annotation Format

The annotations are presented in the form of road lines and annotation polygons. This is an intentional choice to enable these annotations to be robust to future efforts to model and thereby vary spatial alignment. In this format, road lines can be annotated by computing the intersection of road lines with the annotation polygons. This decouples road line annotations from variations in road line alignment, meaning that ML models predicting road line labels can generate ground truth labels independently of alignment logic.

### Baseline Models

This work introduces eighteen baseline models, each of which formulates road damage assessment as a segmentation task, as benchmarks for future investigations. These eighteen are composed of nine baseline model architectures, each trained and evaluated on the "Simple" and "Full" prediction tasks, defined later. These nine architectures are as follows: 1) A Random Baseline, 2) UNet without attention (Ronneberger, Fischer, and Brox 2015), 3) UNet with attention (Oktay et al. 2018), 4) ResNet101 (He et al. 2016)+ PSPNet (Zhao et al. 2017), 5) ResNet101 (He et al. 2016) + DeepLabv3plus (Chen et al. 2018), 6) Vision Transformer (Reed et al. 2023) + Segmenter (Strudel et al. 2021), 7) Vision Transformer (Reed et al. 2023) + Segmenter (Pretrained Vision Transformer)(Strudel et al. 2021), 8) Vision Transformer (Reed et al. 2023) + UperNet (Xiao et al. 2018), 9) Vision Transformer (Reed et al. 2023) + UperNet (Pretrained Vision Transformer) (Xiao et al. 2018).

The sixteen trainable baselines provide guidepost performance for future efforts, and the random baseline provides a reasonable lower bound. Road lines were masked by buffering all road lines by 40 pixels, forming a rectangular mask, and all trained models utilized two strategies simultaneously for class imbalance mitigation: weighted sample presentation targeting uniform label propensity and CCE loss weighted by observed inverse label propensity.

## Evaluations

Two evaluations were conducted to establish a range of performances that for future model development could reference and to validate model performance on real-world data. This section introduces two labeling tasks associated with this dataset. The first task, termed "Simple," is a labeling task where the model must label road lines in one of three classes: no annotation, a "Partial" class label, or a "Total" class label, which also includes the "Not able to Determine" label. In the second task, termed "Full," the model must label the road line according to the exact ground truth label.

The first evaluation was to establish baseline performance on this dataset using the two approaches that were described earlier, concerning the labeling tasks "Simple" and "Full. The second experiment was to establish the value of adjustments in this dataset. As discussed in the literature, misalignment between imagery and preexisting spatial data can result in performance degradations (Maiti, Oude Elberink, and Vosselman 2022; Vargas-Muñoz et al. 2019). This experiment aims to determine the scale of the gap that may exist in this dataset, given the baselines. This section begins with a discussion of evaluation metrics, followed by the details and results of each of the evaluations conducted.

### Evaluation Metrics

Two metrics are used to evaluate the performance of these ML baselines: Intersection over Union (IoU) and F1. IoU was determined to most closely align with the expected use cases of the model, while F1 is presented for additional context. In this context, IoU represents the proportion of a labeled segment of road that corresponds to the label.

Computing IoU and F1 is confounded by the variable spatial dimensions within CRASAR-U-DROIDs, with imagery varying in resolution between 12.7cm/px and 1.77cm/px. Ideally, any evaluation metric would be robust to these changes so as not to be biased by imagery near the extremes of these spatial dimensions. To combat this, evaluation takes place along this spatial dimension, rather than the pixel dimension. To compute IoU, the value of the true positive length of a road line label is its length in kilometers, rather than pixels. This decision shifts the metrics into a dimension of greater value to practitioners compared to pixels. This metric will be denoted as $\text{IoU}_{km}$ and $\text{F1}_{km}$.

| Model | Simple | | | | Full | | | |
|---|---|---|---|---|---|---|---|---|
| | Adjusted | | Unadjusted | | Adjusted | | Unadjusted | |
| | $\text{IoU}_{km}$ | $\text{F1}_{km}$ | $\text{IoU}_{km}$ | $\text{F1}_{km}$ | $\text{IoU}_{km}$ | $\text{F1}_{km}$ | $\text{IoU}_{km}$ | $\text{F1}_{km}$ |
| **UNet with Attention** (Oktay et al. 2018) | **0.331** | **0.393** | **0.312** | **0.368** | **0.091** | 0.095 | **0.092** | 0.096 |
| **UNet w/out Attention** (Ronneberger, Fischer, and Brox 2015) | 0.307 | 0.376 | 0.279 | 0.342 | **0.091** | 0.095 | **0.092** | 0.096 |
| **ResNet101 + DeepLabv3Plus** [(He et al. 2016) + (Chen et al. 2018)] | 0.165 | 0.264 | 0.143 | 0.234 | 0.081 | **0.103** | 0.076 | 0.096 |
| **ResNet101 + PSPNet** [(He et al. 2016) + (Zhao et al. 2017)] | 0.283 | 0.388 | 0.269 | 0.370 | 0.084 | 0.095 | 0.083 | 0.095 |
| **ViT-L + Segmenter** [(Reed et al. 2023) + (Strudel et al. 2021)] | 0.039 | 0.074 | 0.033 | 0.064 | **0.091** | 0.095 | **0.092** | 0.096 |
| **ViT-L (Pretrained) + Segmenter** [(Reed et al. 2023) + (Strudel et al. 2021)] | 0.015 | 0.028 | 0.013 | 0.025 | 0.091 | 0.095 | 0.091 | 0.095 |
| **ViT-L + UperNet** [(Reed et al. 2023) + (Xiao et al. 2018)] | 0.211 | 0.309 | 0.200 | 0.296 | 0.087 | 0.099 | 0.087 | **0.099** |
| **ViT-L (Pretrained) + UperNet** [(Reed et al. 2023) + (Xiao et al. 2018)] | 0.209 | 0.313 | 0.185 | 0.284 | 0.087 | 0.098 | 0.087 | 0.098 |
| **Random Baseline** | 0.135 | 0.215 | 0.135 | 0.215 | 0.016 | 0.031 | 0.016 | 0.031 |

Table 2: Model Performance for Baseline Models. Model architectures with an encoder and decoder architecture follow the naming convention "Encoder + Decoder". The macro $\text{IoU}_{km}$ and macro $\text{F1}_{km}$ are reported for adjusted and unadjusted configurations. Bold values represent the maximum value per column. Underlined values represent metrics where any class in the macro average is significantly different from the next highest-performing model based on a Hoefding Bound with $p < 0.001$.

## Baseline Model Performance

All sixteen trainable baseline models were trained on the road line labels from the CRASAR-U-DROIDs training set orthomosaics and then tested, alongside the Random baseline, on the road line labels from the CRASAR-U-DROIDs test set orthomosaics. A complete breakdown of the macro $\text{IoU}_{km}$ and $\text{F1}_{km}$ metrics for each of the prediction tasks is shown in Table 2, with more detailed metrics appearing in the appendix. The model performance appears to degrade as the number of prediction classes increases. Stratified measures of performance by both disaster and class are provided in the "Data & Models" linked below the abstract.

## Evaluation of Spatial Alignment

To measure the value that adjustments add quantitatively, the same eight baseline models trained for the baseline evaluation, described earlier, were evaluated on the test set but without any alignment. Thus, road lines would be translated out of their correct position and potentially over features in the imagery that do not correspond to the road or its conditions. The results of this evaluation are shown in Table 2. Comparing each line between the $\text{IoU}_{km}$ adjusted and $\text{IoU}_{km}$ unadjusted, on average, the model performance degrades by 0.014 units of $\text{IoU}_{km}$ for the "Simple" labeling task and degrades by 0.0003 units of $\text{IoU}_{km}$ for the "Full" labeling task. As for comparing each line between the $\text{F1}_{km}$ adjusted and $\text{F1}_{km}$ unadjusted, the model performance degrades by 0.018 units of $\text{F1}_{km}$ for the "Simple" labeling task and by 0.0004 units of $\text{F1}_{km}$ for the "Full" labeling task.

## Evaluation at Hurricanes Debby & Helene

In response to Hurricanes Debby and Helene, the Attention UNet baseline, configured for the "Simple" task, was deployed operationally in Florida to assess the effectiveness of models in practice and to gather qualitative feedback from practitioners in the field. Hurricane Debby was a Category 1 Hurricane that impacted Florida in August 2024 (NOAA 2024a). Hurricane Helene was a Category 4 Hurricane that impacted Florida in September 2024 (NOAA 2024b). During the response to both Hurricanes, sUAS were deployed to collect aerial imagery of the impacted areas, and the Attention UNet model was used to label imaged roads.

Samples from the model outputs are included in Figure 6 as part of a qualitative assessment. This figure shows four images: two instances of fallen trees on roads that were correctly identified, one instance where misalignment results in mislabeling, and one instance where a clear road is correctly identified despite shadows and artifacts. This model processed an approximately $3\text{mi}^2$ orthomosaic in 5 minutes on a desktop with an NVIDIA RTX4090, hardware that could reasonably be fielded alongside sUAS teams. Model outputs were converted to KML files for dissemination in practice.

In conversations with disaster practitioners following these deployments, two important findings were made. First, false positives were far more tolerable in practice than false negatives, as practitioners immediately inspected indications of damage, but largely ignored labels for clear roads. This was because practitioners wanted to verify if labeled roads would actually be a problem for their vehicles or route, while the large prevalence of labeled clear roads led to that label being largely ignored. Second, it was found that there was no singular value for model performance that would be deemed acceptable in practice. Instead, the value of the model's predictions was time-dependent, with practitioners tolerating more errors if model outputs were delivered earlier. It was determined that the Attention UNet model for the "Simple" task was able to provide operational value in the early stages of disasters when information is especially limited, providing strong evidence of generalization, though practitioners consistently stated they would have preferred a full schema.

# Discussion

A discussion of the above evaluations must be conducted to better characterize the models' performances and to highlight spatial alignment's impact on model performance.

## Baseline Model Performance

The low absolute performance of the trainable baseline models warrants further discussion, as it currently limits applicability. While the Attention UNet model and other baselines outperform the random baseline across every pre-

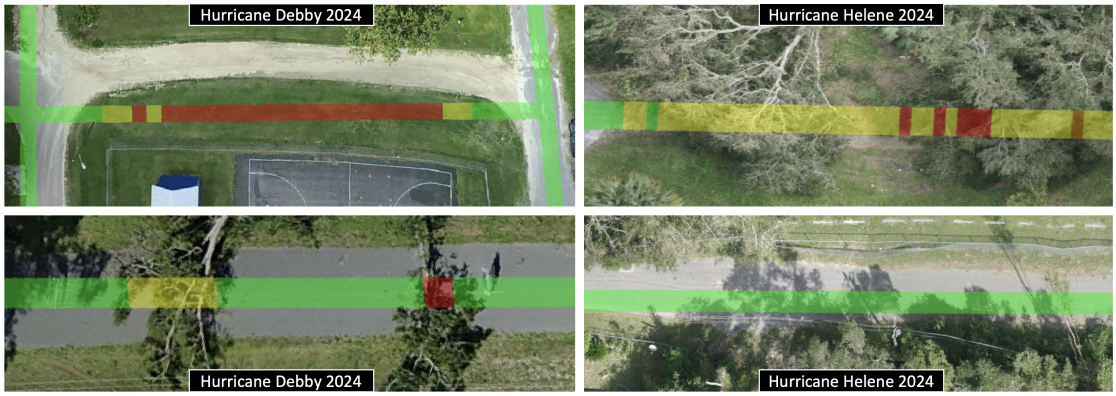


Figure 6: Sample outputs from the Attention UNet on sUAS data collected during the response to Hurricanes Debby (left) and Helene (right) in Florida. Top left: An instance where the model mislabels a road due to an alignment error. Top Right: An instance where a fallen tree blocks and obscures the road, and the model identifies obstructions. Bottom Left: An instance where the model identifies trees that have fallen onto the road. Bottom Right: a clear road that the model correctly identifies.

diction task, performance remains low compared to prior work (Rahnemoonfar, Chowdhury, and Murphy 2023; Rahnemoonfar et al. 2021). While the target function may indeed be noisy, the two-stage review process was used to mitigate this concern. Though it is reasonable to assume that some amount of label noise exists, it is believed that this contribution is minimal. Instead, the framing of this problem as a segmentation task may be a culprit, as many instances of the labels (e.g., partial obstruction, partial road condition) feature pixels that would otherwise correspond to clear roads and thus require reliance on a small number of pixels for correct classification. With high-resolution imagery, this means that models require large receptive fields to provide correct labels. As a result, it is believed that this work represents a rich space for exploring vision models' capabilities in utilizing large contexts and receptive fields.

### Relevance of Class Imbalance

One of the challenges in this dataset is class imbalance, as shown in Figure 4. Given that CRASAR-U-DROIDs sources imagery from real-world disasters and operations, it is unsurprising that this dataset has such class imbalances. With this in mind, this dataset is representative of the types and diversity of data expected in real-world operations, as imagery is sourced directly from that distribution.

### Importance of Spatial Alignment

Two of the evaluations conducted show a clear need for future efforts to focus on alignment. First and most clearly, the decrease shown between the baseline models when run on aligned and unaligned data is explicit quantitative evidence that model performance degrades when alignment is not managed. Second, however, is the qualitative evidence shown from the model deployment on imagery collected from Hurricane Debby. The top left image in Figure 6 shows unambiguously the model labeling an otherwise clear road incorrectly because the road line is not correctly aligned with the road. Additionally, an analysis of the dataset indicates that when adjustments are not applied, then 9% of all roads (59km) fall outside of the 7.2 meter nominal width of two lanes of road (Stein and Neuman 2007), and 8% (11km) of adverse road conditions would be labeled incorrectly as they fall outside of this same 7.2 meter bound. Operationally, should ML systems be deployed without oversight, spatial misalignment may reduce route efficiency or cause spurious repair dispatches due to misalignment-driven false positives. Further, practitioners distrusted model outputs when road lines were not coincident with imagery. Therefore, ML models must manage spatial misalignment to effectively assess adverse conditions on predefined road lines in practice.

## Conclusion

This paper addressed three challenges in the development of CV/ML systems for road damage assessment with post-disaster sUAS imagery, providing the infrastructure to facilitate the further development of CV/ML systems to support decision-making during disaster response. A practitioner-relevant road damage assessment schema was developed with consultation of federal and state agencies to align with operational needs. From which the largest collection of road damage assessment labels dataset was released, along with baseline models trained on this dataset. Finally, a qualitative analysis operationally validates one baseline model in response to Hurricanes Debby and Helene and finds that practitioners tolerate false positives for earlier road damage assessment, providing insight into the evaluation of such a model for operational value. Future work will focus on first, collecting additional data, potentially including non-disaster imagery, to improve the performance of the presented models; second, conducting analyses of the baseline models and experimenting to improve model performance and explore performance degradations on the "Full" formulation; third, and finally, evaluating the ethical implications of both false negatives and false positives on disaster operations. This work is expected to enable the development of operationally valid CV/ML systems for disaster response, pushing the CV/ML communities to deliver systems that will support decision-making during disaster response.

## Ethical Statement

All imagery in this work has been affirmatively released by the agencies having jurisdiction; it was collected in accordance with the appropriate FAA regulations and guidance, and it was collected at the direction of agencies having jurisdiction. In conjunction with this release, the agencies having jurisdiction screened all imagery and withheld imagery that they did not want to be released publicly.

All imagery considered in this work was collected at the direction of agencies having jurisdiction, and thus, it captures the operational distribution of data. As a result, this work was scoped such that commentary on the nature of data collection and commentary on how practitioners direct data collection, as it relates to potential over- or undersampling of specific geographies and socioeconomic statuses, is out of scope for this work. When deploying ML systems without human oversight, such analyses are crucial to ensure that models do not exacerbate biases that they have learned. Future work to measure and mitigate biases learned by models trained on this data and based on the operational distribution of data will be critical to ensure models can operate with minimal technical oversight in the future. At the time of writing, the specific ethical implications of false positives and false negatives are currently being explored in an effort to mitigate potential negative consequences created by these models and labels.

Readers are strongly discouraged from deploying these models in disaster operations without coordination with the authors of this work. As discussed in this work, these models have limitations and faults, and direct introduction of these models, without trained human oversight, risks the unmitigated introduction of model biases in operational environments. Should readers wish to utilize these models in practice, please contact the authors directly to discuss the deployment considerations in detail.

## Acknowledgements

This work is supported by the AI Research Institutes Program funded by the National Science Foundation under the AI Institute for Societal Decision Making (NSF AI-SDM), Award No. 2229881, and under "Datasets for Uncrewed Aerial System (UAS) and Remote Responder Performance from Hurricane Ian" Award No. 2306453. The authors thank the Florida State Emergency Response Team, FL-UAS1 task force, and Florida State University for their support, and the Winchester Thurston School, Ball High School, Bryan Collegiate High School, and Rudder High School for their annotation efforts.

## References

Alam, T.; Quader, S. N.; Islam, S.; and Newaz, A. A. R. 2025. Harnessing Robotic Scouts for Resilient Evacuation Policies in Disaster Scenarios. In *2025 22nd International Conference on Ubiquitous Robots (UR)*, 351–356. IEEE.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Corsar, D.; Markovic, M.; Edwards, P.; and Nelson, J. D. 2015. The transport disruption ontology. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II 14*, 329–336. Springer.

Hänsch, R.; Arndt, J.; Lunga, D.; Gibb, M.; Pedelose, T.; Boedihardjo, A.; Petrie, D.; and Bacastow, T. M. 2022. Spacenet 8-the detection of flooded roads and buildings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1472–1480.

Hargis, A.; Rao, A.; and Choset, H. 2024. Search and rescue base of operation prioritization with aerial orthomosaics. In *2024 IEEE International Symposium on Safety Security Rescue Robotics (SSRR)*, 204–209. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, S.; Bian, Y.; Wang, Y.; Li, X.; Liu, Z.; Ren, Y.; and Zhao, Y. 2024. EarthquakeNet: A High-Resolution UAV-Based Dataset for Earthquake Damage Assessment. In *2024 IEEE International Conference on Image Processing (ICIP)*, 55–61. IEEE.

Korkmaz, S. A.; and Poyraz, M. 2016. Path planning for rescue vehicles via segmented satellite disaster images and GPS road map. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 145–150. IEEE.

Labelbox. 2024. Labelbox. https://labelbox.com [Accessed: 2025-11-13].

Maiti, A.; Oude Elberink, S.; and Vosselman, G. 2022. Effect of label noise in semantic segmentation of high resolution aerial images and height data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2: 275–282.

Majidifard, H.; Jin, P.; Adu-Gyamfi, Y.; and Buttlar, W. G. 2020. Pavement image datasets: A new benchmark dataset to classify and densify pavement distresses. *Transportation Research Record*, 2674(2): 328–339.

Manzini, T.; Murphy, R.; and Merrick, D. 2023. Quantitative data analysis: Crasar small unmanned aerial systems at hurricane ian. In *2023 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 7–12. IEEE.

Manzini, T.; Murphy, R.; Merrick, D.; and Adams, J. 2023. Wireless network demands of data products from small uncrewed aerial systems at Hurricane Ian. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9941–9946. IEEE.

Manzini, T.; Perali, P.; Karnik, R.; Godbole, M.; Abdullah, H.; and Murphy, R. 2024a. Non-Uniform Spatial Alignment Errors in sUAS Imagery From Wide-Area Disasters. *arXiv preprint arXiv:2405.06593*.

Manzini, T.; Perali, P.; Karnik, R.; and Murphy, R. 2024b. CRASAR-U-DROIDs: A Large Scale Benchmark Dataset

for Building Alignment and Damage Assessment in Georectified sUAS Imagery. *arXiv preprint arXiv:2407.17673*.

Manzini, T.; Perali, P.; Murphy, R. R.; and Merrick, D. 2025. Challenges and Research Directions from the Operational Use of a Machine Learning Damage Assessment System via Small Uncrewed Aerial Systems at Hurricanes Debby and Helene. *arXiv preprint arXiv:2506.15890*.

NOAA. 2024a. Debby resources: The latest storm forecasts, maps, imagery and more. https://www.noaa.gov/debby [Accessed: 2025-11-13].

NOAA. 2024b. Helene resources: The latest storm forecasts, maps, imagery and more.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

OpenStreetMap contributors. 2024. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org [Accessed: 2025-11-13].

Pi, Y.; Nath, N. D.; and Behzadan, A. H. 2020. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, 43: 101009.

Rahnemoonfar, M.; Chowdhury, T.; and Murphy, R. 2023. RescueNet: a high resolution UAV semantic segmentation dataset for natural disaster damage assessment. *Scientific data*, 10(1): 913.

Rahnemoonfar, M.; Chowdhury, T.; Sarkar, A.; Varshney, D.; Yari, M.; and Murphy, R. R. 2021. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9: 89644–89654.

Reed, C. J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; and Darrell, T. 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4088–4099.

Ren, M.; Zhang, X.; Zhi, X.; Wei, Y.; and Feng, Z. 2024. An annotated street view image dataset for automated road damage detection. *Scientific Data*, 11(1): 407.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Sabouri, M.; Rajabi, A. B.; Hajianfar, G.; Gharibi, O.; Mohebi, M.; Avval, A. H.; Naderi, N.; and Shiri, I. 2023. Machine learning based readmission and mortality prediction in heart failure patients. *Scientific Reports*, 13(1): 18671.

Stein, W. J.; and Neuman, T. R. 2007. *Mitigation Strategies For Design Exceptions*. US Department of Transportation.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Takyi, S.; Antwi, R. B.; Ozguven, E. E.; Okine, L.; and Moses, R. 2025. Towards Sustainable and Resilient Infrastructure: Hurricane-Induced Roadway Closure and Accessibility Assessment in Florida Using Machine Learning. *Sustainability*, 17(9): 3909.

Urabe, K.; and Saji, H. 2007. Analysis of road blockage after disaster using aerial images. In *SICE Annual Conference 2007*, 1795–1798. IEEE.

Vargas-Muñoz, J. E.; Lobry, S.; Falcão, A. X.; and Tuia, D. 2019. Correcting rural building annotations in OpenStreetMap using convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, 147: 283–293.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Yan, H.; and Zhang, J. 2023. UAV-PDD2023: A benchmark dataset for pavement distress detection based on UAV images. *Data in Brief*, 51: 109692.

Yang, B.; Wang, S.; Zhou, Y.; Wang, F.; Hu, Q.; Chang, Y.; and Zhao, Q. 2020. Extraction of road blockage information for the Jiuzhaigou earthquake based on a convolution neural network and very-high-resolution satellite images. *Earth Science Informatics*, 13: 115–127.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.