MSNet: A Multilevel Instance Segmentation Network for Natural Disaster Damage Assessment in Aerial Videos

Xiaoyu Zhu
Carnegie Mellon University
xiaoyuz3@cs.cmu.edu

Junwei Liang
Carnegie Mellon University
junweil@cs.cmu.edu

Alexander Hauptmann
Carnegie Mellon University
alex@cs.cmu.edu

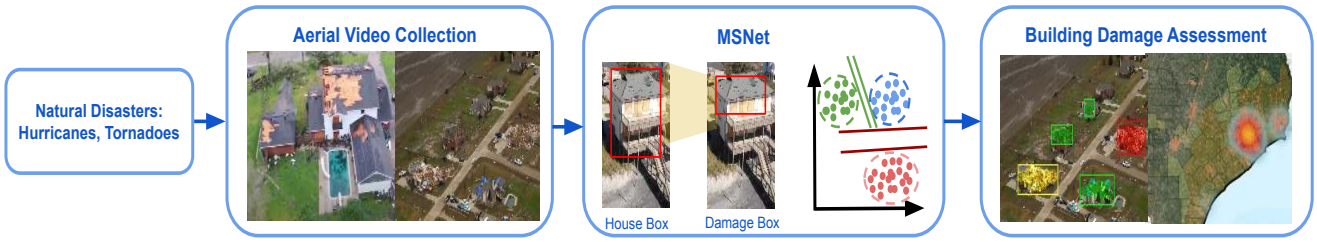

Figure 1. Illustration of the natural disaster damage assessment pipeline. Aftermaths of natural disasters are recorded by drones. Our model is able to detect damage masks and damage scales in different locations. The damage detections along with drones' GPS trajectory could generate a damage assessment location heatmap to aid timely disaster relief efforts.

Abstract

In this paper, we study the problem of efficiently assessing building damage after natural disasters like hurricanes, floods or fires, through aerial video analysis. We make two main contributions. The first contribution is a new dataset, consisting of user-generated aerial videos from social media with annotations of instance-level building damage masks. This provides the first benchmark for quantitative evaluation of models to assess building damage using aerial videos. The second contribution is a new model, namely MSNet, which contains novel region proposal network designs and an unsupervised score refinement network for confidence score calibration in both bounding box and mask branches. We show that our model achieves state-of-the-art results compared to previous methods in our dataset.[1]

1. Introduction

In recent years, natural disasters have impacted many vulnerable areas around the world. In 2019, there have been ten natural disaster events with damages of more than 1 billion dollars each across the United States [9]. Timely response to natural disasters plays a crucial role in disaster relief. However, current damage assessments are mostly based on manual damage detection and documentation, which is slow, expensive and labor-intensive work [24].

With the increasing availability of consumer-grade drones, a large number of aerial videos are recorded and shared across social media [18]. After a natural disaster, like a hurricane or a flood, people frequently share drone footage of the district, or the authorities could dispatch drones themselves to assess the damage of the area. These videos could serve as valuable resources for automatic damage assessment. Compared with satellite imagery used in previous damage assessment task works [7, 12, 26], drone videos have the advantage of capturing detailed observations of each building from different angles other than just from a top-down perspective. Valuable structural information of the buildings could be extracted from drone videos for further damage evaluation, *i.e.*, whether the buildings are going to collapse.

Consider the example in Figure 1, there are three challenges for automatic building damage assessment. The first is the diversity of buildings, the level of damages and the location of damages. Buildings could include homes, schools, coastal buildings, factories, and other facilities. Some might be slightly damaged, and others might be completely damaged. Some might only have severe damage on the roof. The second challenge is the detection of small objects and debris. The drone videos are usually recorded from a high

[1] https://github.com/zgzxy001/MSNET

altitude where many of the damaged parts are only represented by a few dozen pixels (See Section 3). The third challenge is the changes of viewpoints as the drone flies over the area. The damage of a building might only be visible from a certain viewpoint. This leads to problems like missed detection and inconsistent detections by a single image-based detector.

To overcome the aforementioned challenges, we have collected the first dataset with aerial videos for natural disaster damage assessment. Our dataset, namely ISBDA (Instance Segmentation in Building Damage Assessment), consists of fine-grained building damage bounding box and mask annotations of different damage levels. This provides the first quantitative benchmark for evaluating building damage assessment models. Our second contribution is to propose a new neural network model, *MSNet*, to address the difficulties of accurately detecting damages in buildings with aerial videos. Our model makes use of the hierarchical relationship between building and damage, and inter-frame spatial consistency of multiple viewpoints to train more robust representations. To summarize, our contribution is fourfold:

- We present the first natural disaster building damage assessment dataset, namely ISBDA, using aerial drone videos. It is annotated with fine-grained instance-level building and damage bounding boxes and masks. It provides the first quantitative benchmark for assessing damage assessment in aerial videos.
- We propose a novel neural model termed Hierarchical Region Proposal Network (HRPN), which explores the hierarchical spatial relationship among different objects, and thus significantly improving the model performance.
- We propose an unsupervised score refinement model named Score Refinement Network (SRN) based on inter-frame consistency to tackle the challenges of detections using drone videos.
- We empirically validate our model on the proposed ISBDA dataset for damage assessment, in which our model achieves the best results compared to state-of-the-art object detection models.

2. Related Work

Natural Disaster Damage Assessment Datasets. Existing damage assessment dataset can be roughly categorized into two types: ground-level images and satellite imagery. The ground-level images were mostly collected from social media [22]. Those datasets only have image-level labels available, because the scene captured by a single ground-level image is highly limited. Besides, due to the lack of geo-tags in social media, ground-level images may not be suitable for large-scale damage assessment. Another disaster data source is satellite imagery based on remote sensing [7, 12, 26, 27, 16]. However, the main limitation of satellite imagery is that it could not provide detailed damage information due to the long distance to the captured buildings and its limited vertical viewpoint. We are the first to propose a dataset from drone video viewpoints (typically about forty-five degrees) for damage assessment tasks with instance-level damage annotations.

Damage Detection Approaches. Current damage detection approaches can be put into three categories. The first category is using supervised machine learning methods which include pixel-based relevant change detection [5] and object-based local descriptors [29]. The second category includes unsupervised methods [11, 21, 23] that generally refer to outlier detection in scene changes. The third category, a recent trend on damage assessment is using semi-supervised approaches [10] aimed at using less human-labeled data and maintaining higher accuracy. Other literature also proposed deep learning frameworks such as Convolutional Neural Networks (CNN) [1, 22] to predict the damage level of each image. However, existing models only worked on building bounding box prediction tasks, which lack specific locations of damaged parts.

Anchor-based Region Proposal Networks. Existing literature on anchor-based region proposal networks mostly adopted dense anchoring scheme, where anchors are sampled densely over the spatial feature space with predefined scales and aspect ratios. The most representative work is Region Proposal Network (RPN) introduced in Faster R-CNN [25], which designed a light fully convolutional network to map sliding windows to a low-dimensional feature space. This framework has been widely adopted in later research [8, 13]. Some research [33] focused on using meta-learning to dynamically generate anchors from the arbitrary customized prior boxes. Other research works [4, 6, 34] adopted cascade architecture to regress bounding boxes iteratively for progressive anchor refinement. Some researchers [30] tried to remove the iteration process by predicting the center of objects of interest. However, there is still a lack of region proposal networks that could utilize spatial hierarchical relationships among objects which could potentially improve detection accuracy.

Figure 2. Visualization of our ISBDA dataset. The green, yellow and red polygons denote damages in Slight, Severe and Debris levels, respectively. The rectangles composed of solid lines represent damaged building bounding boxes. The polygons with dotted lines represent segmentation masks of damaged parts.

Detection Score Refinement. Current research in detection score refinement can be categorized into two streams, bounding box score refinement and mask score refinement. In bounding box score correction, most works focused on making modifications on the basis of Non-maximum Suppression (NMS) algorithm, such as Fitness NMS [28] and SoftNMS [2]. Jiang *et al*. [15] proposed IoU-Net that directly predicted box IoU, and the predicted IoU was used for the bounding boxes refinement. In terms of score refinement in mask level, Mask Scoring R-CNN [14] was proposed by adding a MaskIoU head to regress the IoU between the predicted mask and its ground truth mask. One limitation of this approach is that it can only refine the mask scores, which nearly has no impact on the bounding box branch. Our proposed score refinement algorithm based on inter-frame consistency is able to achieve consistent improvement in both bounding box and mask branches.

3. The ISBDA Dataset

3.1. Data Collection

In order to fully assess building damages in different scenarios and locations, we have collected ten videos from social media platforms, which recorded severe hurricane and tornado disaster aftermaths in recent years. Specifically, the aerial videos were recorded after Hurricane Harvey in 2017, Hurricane Micheal and Hurricane Florence in 2018 and other three tornadoes (EF-2 or EF-3) in 2017, 2018 and 2019, respectively. The affected areas recorded in the videos include Florida, Missouri, Illinois, Texas, Alabama and North Carolina in the United States. The total length of the collected videos is about 84 minutes.

To get individual frames, we first obtain video clips from the ten videos that: (1) do not have apparent camera rotations; and (2) fly with moderate and stable speed. To further improve the annotation efficiency and cover different scenarios, we extract one frame out of every ten frames from these video clips. Overall, we have collected 1,030 frames for instance-level building and damage annotation.

One important problem is to define damage scale and corresponding standards which can cover various types of damages in different scenes. Following the damage assessment practice, Joint Damage Scale [12], we divide building damages into three levels: Slight, Severe and Debris. Slight refers to visible cracks or appearance damages. Severe refers to partial wall or roof collapse, which are apparent structural damages. Debris refers to completely collapsed buildings.

3.2. Hierarchical Instance-level Annotation

To provide fine-grained localization information of individual damages, we formulate the damage assessment task as an instance segmentation problem. We annotate both the polygons of damaged buildings and the specific damaged parts of the buildings. In order to explore the hierarchical relationships between building and damaged part instances (*i.e.*, specific damaged parts are within corresponding damaged building boxes), we also include the mappings between each damaged part ID and its corresponding damaged building ID. The dataset is annotated by three experienced annotators, and one pass of verification is performed for each annotation to ensure accuracy.

3.3. Dataset Statistics

Overall, 1,030 images sampled from 10 videos are annotated with instance-level building masks and damaged part masks. The dataset has 2,961 damaged part instances which are divided into three levels: Slight, Severe, and Debris. Following Microsoft COCO's [20] size definition, we calculate the number of damaged part instances in different sizes for each damage scale, shown in Table 1.

We also analyze the distribution of the area of damage segmentation in the ISBDA dataset, shown in Figure 3. We observe that the majority of the damage segmentation are

Damage Scale	Small	Medium	Large	Total
Slight	204	1169	746	2119
Severe	-	120	440	560
Debris	-	54	228	282

Table 1. Distribution of annotation sizes. Small: area less than 32×32; Medium: area greater than 32×32 and less than 96×96; Large: area greater than 96×96. Area is measured as the number of pixels in the segmentation mask.

relatively small. Visualization of the ISBDA dataset and annotations is shown in Figure 2.

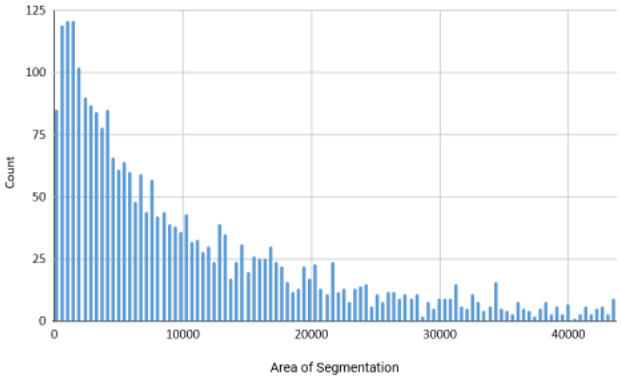

Figure 3. The distribution of the area of damage segmentation in our ISBDA dataset. We only show the distribution of areas below 90th percentile of the whole dataset for better visualization purpose. Area is measured as the number of pixels in the segmentation mask.

4. Method

4.1. Overview

To provide fine-grained localization information, similar to some of the existing works [12], we formulate the damage assessment task as an instance segmentation problem. Moreover, our model will predict damage-level instance masks instead of building-level, which is a more challenging task due to the high damage variance and small damaged area. We propose a new model named *MSNet* in order to learn more robust representations in different scenarios with different viewpoints. It includes two types of supervision: supervision of building bounding boxes for low-level damage anchor sampling and mask segmentation; and supervision of temporal and spatial relationships between adjacent video frames. In summary, it has the following key components:

Pyramid Backbone Network uses ResNet-50 based Feature Pyramid Network (FPN) [19] to extract spatial features of input images.

Hierarchical Region Proposal Network first generates high-level building proposals and then uses them to supervise low-level anchor sampling and damage proposals generation.

Score Refinement Network is proposed to calibrate the confidence scores of instances in adjacent frames which share common appearance features but have confidence score variances.

Mask R-CNN Head includes the R-CNN head for bounding box and class prediction, and the Mask head for mask prediction [13].

In the rest of this section, we will introduce the above components and the learning objectives in details.

4.2. Hierarchical Region Proposal Network

Traditional Region Proposal Network (RPN) treats all objects in the same spatial level, and uniformly generates dense anchors over the feature space. If we adopt a conventional RPN scheme and train the RPN with building and damage proposals simultaneously, the hierarchical relationship between buildings and damaged parts will not be utilized. Therefore, we propose a new model, termed Hierarchical Region Proposal Network (HRPN), to address the aforementioned problems.

In HRPN, there are two RPNs sharing the same backbone network: a high-level RPN and a low-level RPN. The high-level RPN is trained with damaged building boxes with binary labels indicating whether the proposal is a damaged building or not. The low-level RPN utilizes building proposal outputs from the high-level RPN for anchor sampling. We sample anchors based on one of the two metrics: Intersection over Union (IoU) and Inner Intersection (II) between high-level region proposals and low-level anchors. For each low-level low-level (damage) anchor $A_{\tilde{a}}$, we define its sampling score as:

$$\mathrm{S}_{IoU}(A_{\tilde{a}}, A_p) = \max_{A_p \in P} \frac{A_{\tilde{a}} \bigcap A_p}{A_{\tilde{a}} \bigcup A_p} \tag{1}$$

$$\mathrm{S}_{II}(A_{\tilde{a}}, A_p) = \max_{A_p \in P} \frac{A_{\tilde{a}} \bigcap A_p}{A_{\tilde{a}}} \tag{2}$$

where P is a set of high-level (building) region proposals. For each anchor, we compute its sampling score and only keep anchors with scores larger than a certain threshold S. Then the sampled anchors are used for damage proposals generation.

4.3. Score Refinement Network

In previous works [3], the confidence scores are determined by single-frame detection, while correspondence between two adjacent frames is not utilized. We propose a score refinement model based on inter-frame temporal and spatial correspondence termed Score Refinement Network (SRN). The input of the model is randomly generated

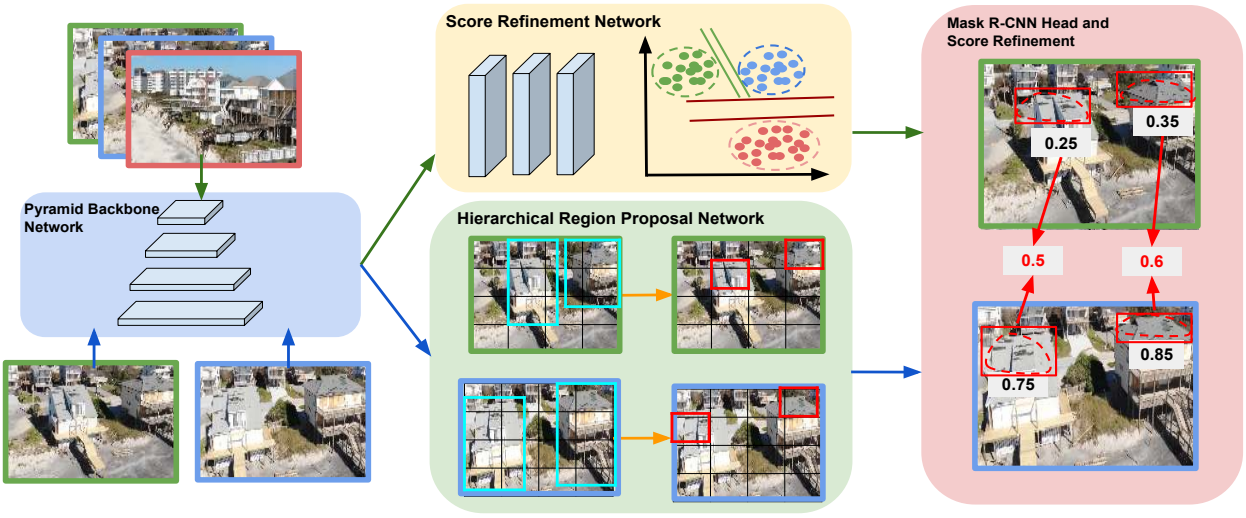

Figure 4. Network architecture of *MSNet*. The left part contains a pyramid backbone network to extract features in multi-scale levels. The backbone network is shared in the two neural network's training. The first neural network (Bottom) is for generating instance segmentation results. Specifically, for each image, Hierarchical Region Proposal Network takes the encoded features to generate proposals for damaged buildings. The building proposals are used to give supervision on damage proposals generation (Yellow Arrow). The second branch (Top) is for the training of Score Refinement Network. The adjacent frames (images with green and blue edges) along with one negative sample (image with red edges) are firstly fed into the Pyramid Backbone Network, then Score Refinement Network is trained with the proposed Multi-scale Consistency Loss to learn feature similarity. These two branches are joined at the end, where Mask R-CNN Head generates bounding box and mask predictions. Finally, the score refinement algorithm is performed to calibrate the confidence scores.

triplets and each triplet is composed of one frame and its adjacent frame as a positive frame and another random frame as a negative frame. By incorporating multi-scale features from the FPN backbone, we design a multi-scale consistency loss to force SRN to learn feature representations such that one sample's distance to its positive sample is closer than its distance to the negative one. We aim to refine the scores of instances in adjacent frames which share common appearance features but have confidence score variances.

Inspired by [31], we use patch mining to build triplets and each is composed of one sample P_i, its relative adjacent frame P_i^+ and its random sample P_i^-. The triplets are sampled based on the fact that the average drone speed is 50 mph and thus the frame variances within half seconds are small. Therefore, given a frame x_t at time t and the video frame rate r, the positive sample is defined as the frame in range $[x_t - 0.5r, x_t + 0.5r]$. The negative sample is defined as the frame in range $[0, x_t - 10r] \bigcup [x_t + 10r, T]$. T is the maximum frame number of the video.

Multi-scale features usually demonstrate significant performance improvement in object detection tasks [13, 19]. Therefore, we propose Multi-scale Consistency Loss (MCL) which makes use of multi-scale feature maps. For two image patches X_i, X_j, we firstly obtain the feature maps of each image from the last four layers of the FPN backbone, namely P_{ik}, P_{jk}, where $k \in [1, 2, 3, 4]$. These feature maps are used as input to SRN. For an input feature P, we can obtain its feature from the last SRN layer as $f(P)$, where f is a feature encoder which is composed of three fully connected layers. Then, we propose a spatial-wise similarity metric of two feature maps P_{ik}, P_{jk} in FPN level k using:

$$Sim(P_{ik}, P_{jk}) = \sum_{w=0}^{W} \sum_{h=0}^{H} \frac{f(P_{ik}^{wh}) \cdot f(P_{jk}^{wh})}{\|f(P_{ik}^{wh})\| \|f(P_{jk}^{wh})\|} \tag{3}$$

$$D(P_{ik}, P_{jk}) = 1 - Sim(P_{ik}, P_{jk}) \tag{4}$$

Given a set of triplets and each triplet is denoted as (X, X^+, X^-) , we aim to train SRN which can learn feature representations such that $D(X, X^-) > D(X, X^+)$ using the Multi-scale Consistency Loss (MCL):

$$\mathcal{L}_{mcl}(X, X^+, X^-) = \sum_{i=1}^{L} \max\{0, D(X_i, X_i^+) - D(X_i, X_i^-) + m\} \tag{5}$$

where m is a margin constraint parameter, and L is the number of multi-scale layers.

4.4. Training

In this section, we provide detailed descriptions of the training procedure. The first part of the loss function is the HRPN loss, which is defined as:

$$\mathcal{L}_{hrpn} = \mathcal{L}^h_{rpn} + \mathcal{L}^l_{rpn}. \quad (6)$$

Here, $\mathcal{L}^h_{rpn}$ and $\mathcal{L}^l_{rpn}$ represent the loss of high-level RPN and low-level RPN, respectively. The low-level RPN conducts anchor sampling and proposal generation under the supervision of high-level RPN. As described in Section 4.2, the losses of damage proposals which are filtered out under the supervision of high-level building proposals are not computed in the HRPN loss. The definition of RPN loss follows [25]. $\mathcal{L}_{cls}$, $\mathcal{L}_{box}$, and $\mathcal{L}_{mask}$ follow the definitions in [13]. $\mathcal{L}_{mcl}$ is computed using Equation 4.3.

The final multi-task loss of our proposed approach is calculated using:

$$\mathcal{L} = \mathcal{L}_{hrpn} + \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask} + \mathcal{L}_{mcl}. \quad (7)$$

The HRPN and Mask R-CNN Head can be trained end-to-end together with SRN. However, in that case, the model training and inference would be heavy due to the multi-scale feature similarity calculation. Therefore, we only calibrate confidence scores of the model which has the best instance segmentation performance.

4.5. Inference

In test time, we use HRPN to generate building region proposals. Then the building proposals are used as supervision for damage anchor sampling and proposal generation, as described in Section 4.2. In the second stage, the model extracts features using RoIAlign for each damage proposal and performs proposal classification, bounding box regression and mask prediction.

During the inference of SRN, given two adjacent frames P and Q, we firstly extract the last four layers from the Pyramid Backbone Network for each frame. The four layers are used as input for SRN described in Section 4.3 to extract similarity feature maps. Then we use RoIAlign to align the extracted features with each bounding box. For each prediction (including bounding box and mask) in frame P, we calculate its similarity score with each prediction in frame Q, using equation 3 with the aligned feature maps as input. Then we can obtain the prediction in frame Q that has the highest similarity score with it. The average of these two confidence scores is used as their final scores. Note that we only refine confidence scores that fall within the range of $[C_0, C_1]$.

5. Experiments

In this section, we compare our *MSNet* model with state-of-the-art baselines on the proposed ISBDA dataset. We randomly split the dataset into subsets with no overlapping scenes. We train our model using 80% of the dataset, and test on the rest 20% dataset. We repeat the split and experiments 3 times and report the results in Table 2. The final reported results are the average over the evaluation results of all splits.

We report the standard COCO instance segmentation metric [20] including AP (averaged over all IoU thresholds), AP@0.25, AP@0.5, and AP_S, AP_M, AP_L (AP at different scales). Unless noted, AP is evaluating using mask IoU.

5.1. Implementation Details

We compare our model with two recent state-of-the-art instance segmentation models, PolarMask [32] and Mask R-CNN [13]. All models use ResNet-50 based FPN as a backbone network. We train all the networks for 100 epochs, with a starting learning rate of 0.003 then we decrease it to 0.001 after 10 epochs. Mini-batch SGD is used as the optimizer with batch size equals 8. We initialize all the backbone networks with the weights pre-trained on COCO [20]. The input images are resized to have the shorter side being 800 and the longer side less or equal to 1333. For testing, an NMS with threshold 0.5 is used and top 100 detections are retained for each image.

For the score refinement procedure, SRN is trained using hard negative mining. We firstly generate 1,000 (X, X^+) pairs from different videos, and randomly extract 5 negative samples for each (X, X^+) pair as described in Section 4.3. We calculate the loss of 5 negative samples, and choose the top K ones with the highest losses as in [31] to optimize. For the experiments, we use $K = 1$. Adam optimizer [17] is used for network training with learning rate 0.001, and each batch is composed of one (X, X^+) pair and 5 negative samples. For testing, we choose C_0 = 0.2, and C_1 = 0.7 for the range described in Section 4.5.

5.2. Comparison to state-of-the-art

Baseline methods. We compare our method with state-of-the-art models and their variants customized for the damage instance segmentation problem. PolarMask [32] is a single shot instance segmentation model with damage masks as input only. Mask R-CNN [13] is one of the state-of-the-art instance segmentation models. Two variants of Mask R-CNN are used as baselines: (1) Mask R-CNN with damage bounding boxes and masks as input; and (2) Mask R-CNN co-trained with damaged buildings and damages. Damaged building bounding boxes are used for RPN and R-CNN head training, and damage masks are used for the training of Mask head.

Quantitative results. Table 2 lists the damage instance segmentation results. Compared with PolarMask, our model is able to obtain significant improvement, *e.g.*, an absolute increment of 14.9% mask AP. For the Mask R-CNN baselines, we observe that Mask R-CNN trained with

Method	AP	AP_{25}	AP_{50}	AP^{bb}	AP^{bb}_{25}	AP^{bb}_{50}
PolarMask+Damage	22.3	29.1	15.4	24.4	29.6	18.2
Mask R-CNN+Damage	34.4	40.6	26.9	35.9	40.9	29.4
Mask R-CNN+Building+Damage	32.2	39.5	23.3	34.0	40.3	25.7
Ours	**37.2**	**44.2**	**28.8**	**38.7**	**44.4**	**31.5**

Table 2. Cross scene evaluation results. We report detection and instance segmentation results. AP denotes instance segmentation results and AP^{bb} denotes bounding box detection results. In the results area, rows 1 and row 2 use the PolarMask and Mask R-CNN frameworks with only damage masks as input; row 3 uses Mask R-CNN co-trained with damaged buildings and damages as the baseline model. The results show that our proposed method gains significant improvements compared to state-of-the-art models.

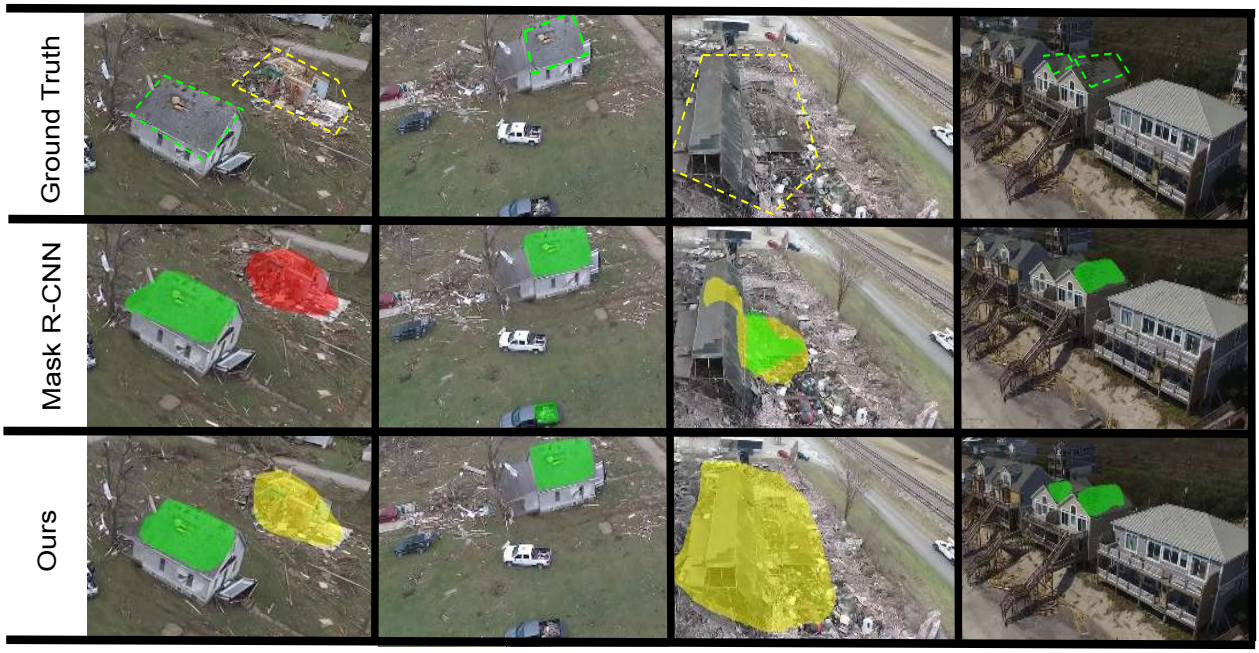

Figure 5. Visualization of the predicted damage segmentation. This figure demonstrates that our proposed model can alleviate the following errors: (1) label misclassification (first column, left to right); (2) false positive segmentation in the complex scenario with cars and buildings (second column); (3) incompleted masks in noisy video scenario (third column); and (4) missed masks (fourth column).

damage masks could be confused by the high variance of damage masks in different locations and scenarios. When the Mask R-CNN model is trained with building boxes and damage masks, the errors in building detection will impact the damage detection in the second stage. Also, the model could not precisely predict the damage masks from large building bounding boxes. Our proposed model utilizes the hierarchical nature of the damaged buildings and damaged parts, and outperforms the baseline with 5.0% AP in the segmentation branch and 4.7% AP in the bounding box branch.

Qualitative analysis. We qualitatively demonstrate the advantages of our model in Figure 5, showing that our proposed model can alleviate the following errors: (1) label misclassification (first column); (2) false positive segmentation in the complex scenario with cars and buildings (second column); (3) incompleted masks in noisy video scenario (third column); and (4) missed masks (fourth column). Thanks to the HRPN module and the inter-frame supervision, our model is able to generate accurate and robust detections even in very noisy scenarios like the third column of Figure 5.

5.3. Ablation Study

We evaluate our method on the ISBDA dataset. We use ResNet-50 FPN as a backbone network for ablation study. All experiments in this section are performed on one split.

Model	AP	AP_{25}	AP_{50}	AP^{bb}	AP^{bb}_{25}	AP^{bb}_{50}
Baseline	35.0	41.9	27.8	36.8	42.9	29.9
Baseline + HRPN	39.3 (+4.3)	46.6 (+4.7)	31.0 (+3.2)	41.4 (+4.6)	47.1 (+4.2)	33.7 (+3.8)
Baseline + HRPN + SRN	40.0 (+5.0)	47.7 (+5.8)	31.3 (+3.5)	42.1 (+5.3)	48.1 (+5.2)	33.9 (+4.0)

Table 3. Effect of HRPN and SRN. We use Mask R-CNN co-trained with building and damage instances as the baseline model. The results show that HRPN component gains significant improvement by 4.3% AP compared with the baseline model. Combined with HRPN, the SRN component also gets consistent improvement in both bounding box and mask branches.

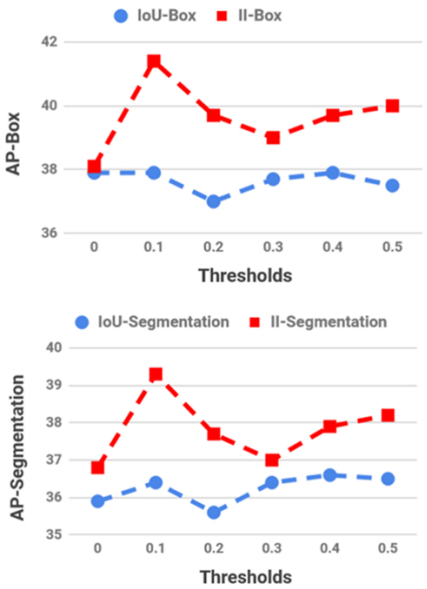

Figure 6. mAP of bounding box and segmentation using different IoU and II thresholds. The blue and red lines denote IoU and II metrics, respectively.

Different IoU and II thresholds. In Figure 6, we compare the effects of different thresholds for IoU and II on the model performance using equations in Section 4.2. We train our model with IoU and II from 0.0 to 0.5 in steps of 0.1. For the model with IoU as metrics, the model gets the best performance when IoU equals 0.4. For the model with II as metrics, the model achieves the best performance when it equals 0.1.

Choices of IoU and II metrics. In Table 4, we report the best performance model among different IoU and II thresholds, respectively, where IoU equals 0.4 and II equals 0.1. We observe that II metric gains 2.7% AP improvement compared with IoU metric. By analyzing the AP in different sizes, we find that the small objects get the most significant improvement for 7.1% absolute value. This is probably because in IoU calculation, small damage anchors only occupy a small portion of its union with a large building bounding box. Therefore, small damage instances may not be well detected. On the other hand, II could properly handle such cases as it performs anchor sampling by calculating the intersection within the damage anchors.

M	AP	AP_{25}	AP_{50}	AP_S	AP_M	AP_L
IoU	36.6	42.5	30.1	47.4	41.1	38.6
II	39.3	46.6	31.0	54.5	38.0	42.0

Table 4. Results of different anchor sampling metrics.

Effect of HRPN and SRN. In Table 3, we experiment with the effect of HRPN and SRN. We observe that the HRPN component gains significant improvement by 4.3% AP compared with the baseline model. The SRN component further improves the model performance in both bounding box and mask branches.

6. Conclusion

In this paper, we investigate the problem of conducting damage assessment using user-generated aerial video data. We provide the first benchmark, namely ISBDA, for quantitative evaluation for models to assess building damage in aerial videos. Also, our proposed *MSNet* is able to explore the hierarchical spatial relationship among different objects and calibrate confidence scores to improve the model performance in both bounding box and mask branches. We empirically validate our model on the proposed ISBDA dataset, in which our model achieves the best results compared to state-of-the-art object detection models. We believe our dataset, together with our models, will facilitate future research in remote sensing and damage assessment for better and faster natural disaster relief.

Acknowledgements This research was supported by the financial assistance award 60NANB17D156 from NIST. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST or the U.S. Government.

References

[1] Sheharyar Ahmad, Kashif Ahmad, Nasir Ahmad, and Nicola Conci. Convolutional neural networks for disaster images retrieval. In *MediaEval*, 2017.

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. In *The International Conference on Computer Vision*, 2017.

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *The International Conference on Computer Vision*, 2019.

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1822–1835, 2008.

[6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[7] Sean Andrew Chen, Andrew Escay, Christopher Haberland, Tessa Schneider, Valentina Staneva, and Youngjun Choe. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. *Arxiv*, 2018.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *The Conference and Workshop on Neural Information Processing Systems*, 2016.

[9] NOAA National Centers for Environmental Information (NCEI). U.s. billion-dollar weather and climate disasters, 2019.

[10] Lionel Gueguen and Raffay Hamid. Large-scale damage detection using satellite imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[11] L. Gueguen, P. Soille, and M. Pesaresi. Change detection based on information measure. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4503–4515, 2011.

[12] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. In *Arxiv*, 2019.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision*, 2017.

[14] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[15] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[16] Mohammad Kakooei and Yasser Baleghi. Fusion of satellite, aircraft, and uav data for automatic disaster damage assessment. *International Journal of Remote Sensing*, 38(8-10):2511–2534, 2017.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*, 2015.

[18] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Temporal localization of audio events for conflict monitoring in social media. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1597–1601. IEEE, 2017.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[21] G. Mercier, G. Moser, and S. B. Serpico. Conditional copulas for change detection in heterogeneous remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1428–1441, 2008.

[22] Dat T. Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 569–576, New York, NY, USA, 2017. Association for Computing Machinery.

[23] A. A. Nielsen. The regularized iteratively reweighted mad method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image Processing*, 16(2):463–478, 2007.

[24] Department of Homeland Security Federal Emergency Management Agency (FEMA). Damage assessment operations manual, 2016.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Conference on Neural Information Processing Systems*, 2015.

[26] Vito Romaniello, Alessandro Piscini, Christian Bignami, Roberta Anniballe, and Salvatore Stramondo. Earthquake damage mapping by using remotely sensed data: the Haiti case study. *Journal of Applied Remote Sensing*, 11(1):1 – 16, 2017.

[27] Tim G. J. Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopackova, and Piotr Bilinski. Multi^{3}net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In *Conference on Artificial Intelligence*, 2019.

[28] Lachlan Tychsen-Smith and Lars Petersson. Improving object localization with fitness nms and bounded iou loss.

In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[29] C. Vaduva, T. Costachioiu, C. Patrascu, I. Gavat, V. Lazarescu, and M. Datcu. A latent analysis of earth surface dynamic evolution using change map time series. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4):2105–2118, 2013.

[30] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[31] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision*, 2016.

[32] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[33] Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. Metaanchor: Learning to detect objects with customized anchors. In *Conference on Neural Information Processing Systems*, 2018.

[34] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.