Attention-based two-stream high-resolution networks for building damage assessment from satellite imagery

Victor Oludare, Landry Kezebou, Obafemi Jinadu, Karen Panetta, Sos Agaian

Victor Oludare, Landry Kezebou, Obafemi Jinadu, Karen Panetta, Sos Agaian, "Attention-based two-stream high-resolution networks for building damage assessment from satellite imagery," Proc. SPIE 12100, Multimodal Image Exploitation and Learning 2022, 121000L (27 May 2022); doi: 10.1117/12.2618901

Event: SPIE Defense + Commercial Sensing, 2022, Orlando, Florida, United States

Attention-Based Two-Stream High-Resolution Networks for Building Damage Assessment from Satellite Imagery

Victor Oludare[1], Landry Kezebou[1], Obafemi Jinadu[1], Karen Panetta[1], Sos Agaian[2]
[1]Tufts University, Medford, MA, 02155
[2]City University of New York, New York, NY 100161

ABSTRACT

Satellite imagery provides an efficient means of assessing and effectively planning search and rescue efforts in the aftermath of disasters such as earthquakes, flooding, tsunamis, wildfires, and conflicts. It enables timely visualization of buildings and the human population affected by these disasters and provides humanitarian organizations with crucial information needed to strategize and deliver the much need aid effectively. Recent research on remote sensing combines machine learning methodologies with satellite imagery to automate information extraction, thus reducing turn-around time and manual labor. The existing state-of-the-art approach for building damage assessment relies on an ensemble of different models to obtain independent predictions that are then aggregated to one final output. Other methods rely on a multi-stage model that involves a building localization module and a damage classification module. These methods are either not end-to-end trainable or are impractical for real-time applications. This paper proposes an Attention-based Two-Stream High-Resolution Network (**ATS-HRNet**), which unifies the building localization and classification problem in an end-to-end trainable manner. The basic residual blocks in HRNet are replaced with attention-based residual blocks to improve the model's performance. Furthermore, a modified cutmix data augmentation technique is introduced for handling class imbalance in satellite imagery. Experiments show that our approach significantly performs better than the baseline and other state-of-the-art methods for building damage classification.

Keywords - building damage assessment, convolutional neural network, deep learning, semi-supervised learning, xBD dataset, attention module, selective oversampling, pixel-aware cutmix.

1. INTRODUCTION

The significant effects of climate change are observable on the environment. Climate change effects such as rise in global temperatures, droughts, rising sea levels, intense tornados, earthquakes, and hurricanes, have been projected to worsen over the coming decades [1]. Furthermore, man-made disasters are increasing due to sustained conflicts between communities, rising acts of terrorism, and accidents. The World Health Organization-WHO estimates that over 160 million people are affected by natural disasters, with around 90,000 killed every year [2]. Therefore, there is a dire need to develop effective and efficient means of providing a rapid and accurate assessment of damages caused by such natural disasters. Intuitively, damages to buildings provide a good insight into the number of victims in the aftermath of a disaster. By estimating the extent of damages to buildings, the number of affected individuals can be projected. Such information would be critical for emergency responders to plan and deploy resources to affected areas more efficiently, hence, reducing deployment time and, in turn, minimizing casualties.

Very High Resolution (VHR) remote sensing data provides a low-cost and efficient way to visualize affected areas' structural and spatial characteristics. It also provides sufficient information needed by computer vision algorithms to extract features for automating the process of assessing damages caused by disasters [3]. Aside from VHR, Synthetic Aperture Radar (SAR) data containing information about backscatter and phase contents can also be used to detect damaged buildings [4], [5], [6]. Although SAR data are not affected by inference from clouds and shadows, they can be quite noisy and do not contain any color information [7]. This limitation makes VHR data preferred for disaster exploration.

Multimodal Image Exploitation and Learning 2022, edited by Sos S. Agaian, Vijayan K. Asari,
Stephen P. DelMarco, Sabah A. Jassim, Proc. of SPIE Vol. 12100, 121000L
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2618901

Several methods have been proposed to take advantage of remote sensing data for detecting buildings and generating building damage maps. Tong et al. proposed a way for detecting collapsed buildings due to earthquakes using pre- and post-seismic high-resolution satellite stereo imagery [8]. This is attained using geometric changes in the height of the buildings using the pre- and post-seismic stereo image pairs. Tong et al. further improved this method by proposing a hybrid shadow-analysis approach [9]. It involves establishing the 3D model of the building model using the height data and estimating the overlap between the ground shadow polygon and the casting shadow area of the building. Hua et al. adopted an online clustering algorithm for grouping the buildings' extracted motion and appearance features for detecting collapsed buildings from UAV data [10]. These methods are often ineffective because they require carefully handcrafted features and are binary in classification, i.e., they consider either damaged or undamaged buildings. Change detection techniques have also been proposed to estimate the pixel level differences between pre- and post-image pairs [11]. Although this provides a simple mathematical approximation of changes, it doesn't consider the contextual information in the changes detected.

2. RELATED WORK

Advancements in the field of machine learning, especially the use of convolutional neural networks (CNNs) for feature extraction [12], have paved the way for proffering breakthrough solutions for several computer vision-related problems. State-of-the-art results have been obtained using CNNs for object detection [13]–[15], image recognition [16], [17], semantic segmentation [18], [19] and instance segmentation [20]. Xu et al. investigated the generalizability of CNNs for automating the detection of damaged buildings in satellite imagery [21]. Fujita et al. also explored the effectiveness of CNNs on buildings affected by tsunamis from aerial images [22]. This is done by assigning a label to a post-disaster image using a Siamese-like network [23] which takes a pair of input images of an area before and after a disaster.

Fully convolutional networks (FCNs) have also been adopted for change detection tasks. The most common network has been U-Net-based [24]. This is because, by design, the U-Net architecture is capable of extracting multi-level patterns across different spatial regions of an image and combining them into high-resolution features, thus allowing for more precise localization of regions of interest. Sun et al. proposed a multitask learning framework for change detection using FCNs for detecting building changes in VHRs [25]. Bayramli et al. proposed shadow detection using U-Net for improving damage detection [26]. Although these methods achieved good results, they often lose spatial context when making predictions. To overcome the challenges of series-connected networks such as U-Net and residual networks [16], Wang et al. proposed deep high-resolution networks (HRNet) [27]. This model architecture maintains a high-resolution representation throughout the network by connecting the high-to-low resolution convolution streams in parallel and repeatedly exchanging features across resolutions, thus, making them suitable for position-sensitive (spatially precise) tasks such as semantic segmentation, pose estimation, and object detection [28].

This paper is motivated by the development of Dual-HRNet by Koo et al.[1] and extends the framework proposed in [28] but with two major changes that significantly improve the performance of our two-stream architecture for joint building localization and classification in satellite imagery. These contributions include:

1. Incorporate an attention mechanism in the HRNet blocks to improve the network's representation capacity for improved building detection and damage classification.
2. Introduce a modified Cutmix [29] strategy called Spatial (pixel) aware cutmix and a class-based oversampling technique to resolve the xBD dataset's severe class imbalance issue.

The rest of the paper is organized as follows. Section 3 presents a detailed description of the proposed methodology and the objective functions adopted. Section 4 provides the experimental setup and results. Finally, the conclusions and future work are presented in section 5.

[1] https://github.com/DIUx-xView/xView2_fifth_place

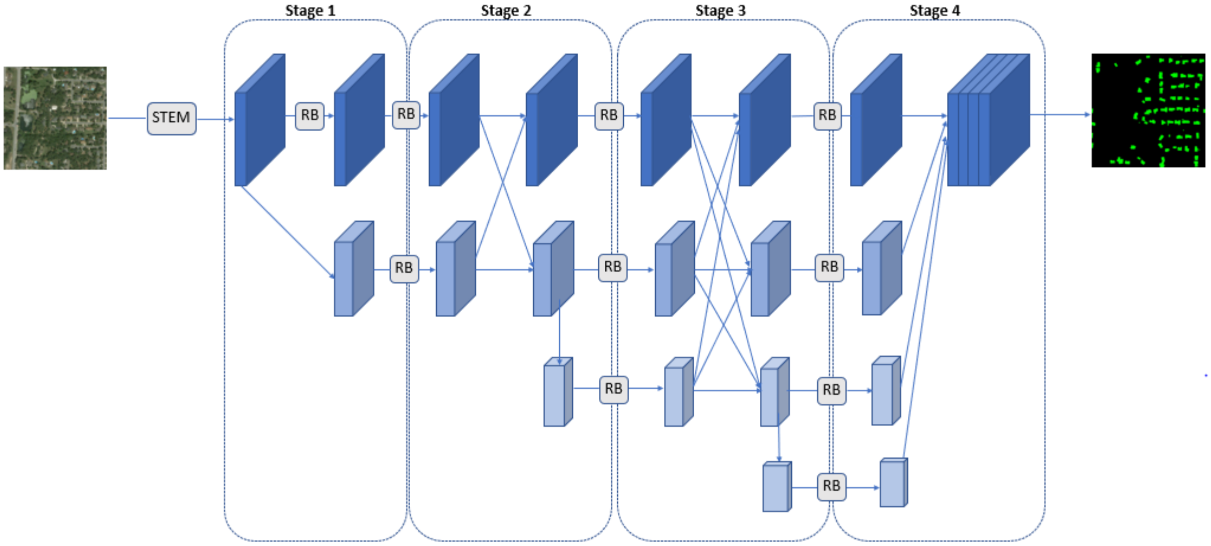
3. PROPOSED METHOD

Feature extraction using a convolutional neural network has proven to outperform conventional feature extraction strategies for image and video data. Several CNN architectures have been created for learning representations for computer vision-related problems such as object detection, semantic segmentation, pose estimation, etc. U-Net and Hourglass-style networks are designed to include an encoder that includes fully or dilated convolutional networks for learning low-resolution representations and a decoder that performs feature upsampling to gradually recover high-resolution representations from its low-level representations. The decoder is often a mirror image of the encoder, and multi-scale fusion is performed with skip connections and concatenation of the low- and high-level features [27]. These networks are often designed to be series-connected, and the subnetworks of high-to-low and low-to-high representation learning could result in information loss. High-resolution networks are designed to maintain a high-resolution representation throughout the network, making them suitable for position-sensitive tasks like semantic segmentation. Therefore, HRNet is adopted in our model architecture with some key modifications to improve the capabilities for learning better representations from satellite imagery.

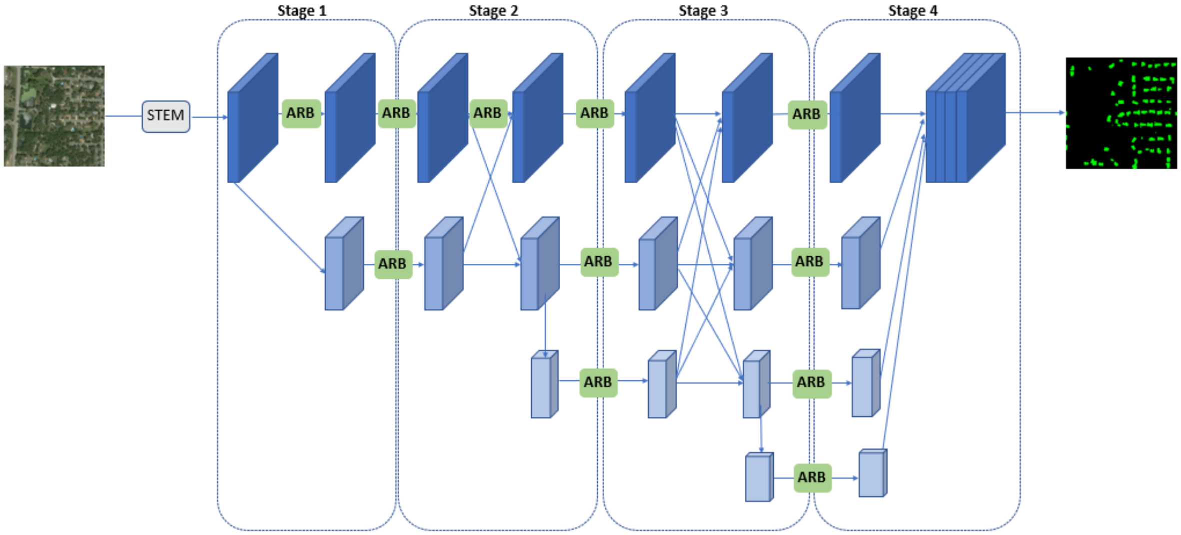
Figure 2 shows the proposed framework, which comprises two network streams. The first stream takes the pre-disaster images, and it is responsible for localizing the buildings in the images. The second stream takes the corresponding post-disaster images and classifies them into different damage levels. Weights are shared between the two streams between each stage of the HRNet via cross-feature fusion proposed by [30].

3.1. Attention-Guided HRNet

The proposed AG-HRNet model replaces the residual blocks in the original HRNet with attention-based residual blocks. The visual attention modules exploit inter-spatial and inter-channel relationships of features from each stage to extract spatial and channel maps. The channel attention maps extract features that are considered “meaningful,” while spatial attention maps extract features that are considered as “informative” [31], [32].



(a)

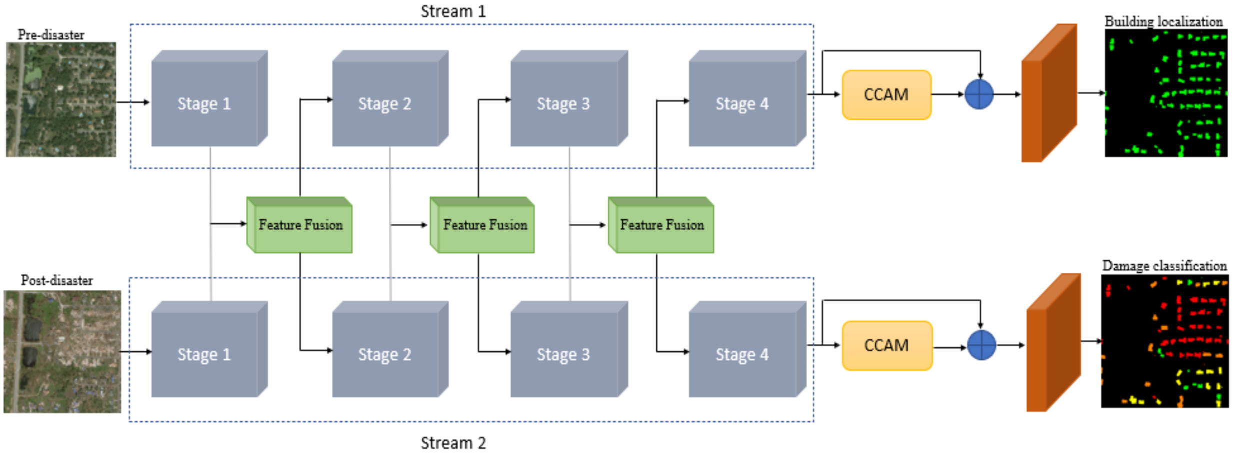


(b)

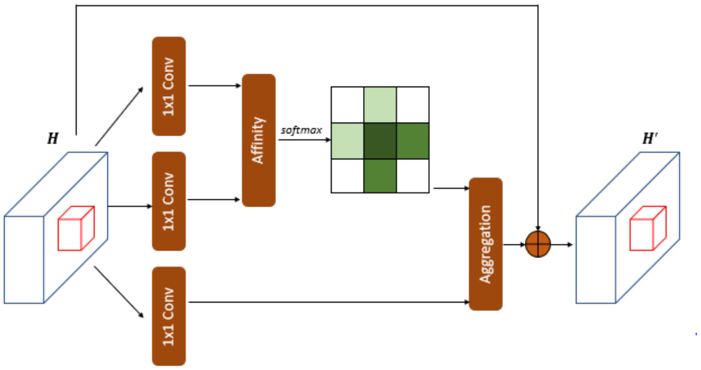
Figure 1: (a) Original high-resolution network architecture [27] (b) Attention-guided high-resolution network (AG-HRNet) inspired by the work of Liu et al. [33].

3.2 Two-Stream Attention-Guided HRNet

Our network comprises of two AG-HRNet models, the first stream takes as input a pre-disaster image and localizes all buildings in the satellite image, while the second stream takes as input a post-disaster image pair and classifies the extent of damages on the localized buildings. The criss-cross attention module (CCAM) proposed by Huang et al. [34] is incorporated in each network stream for extracting contextual information from the images.



(a)



(b)

Figure 2: (a) Attention-guided Two-Stream HRNet for joint building localization and classification. Motivated by the work of Koo et al. [2] (b) Criss-Cross Attention Module (CCAM) [34].

3.3 Pseudo-Label Generation

Similar to our prior work for building classification [28], a semi-supervised approach is explored for harnessing information from unlabeled data. The number of unlabeled samples is frequently bigger than the labeled data. Also, building types differ substantially between areas and continents due to the climate or the culture of the people who reside there. Semi-supervised learning allows for the collection of vast volumes of previously unknown data from various places, enhancing the generalizability of existing models to different building types. According to the JDS standard, these labels comprise the location of the structures and the accompanying label (see table 1). To remove noisy labels and improve the limits of the segmentation masks, label refinement is used. Label refining is done using a consistency-based method. The input images are subjected to four spatial level transformations: horizontal flip, vertical flip, transposition, and rotation, with the trained model utilized to localize and categorize buildings in each transformation. The inverse transform is applied to obtain the final output, and the most frequent class from all five outputs, including the findings from the untransformed picture, is determined. The results show that using a simple semi-supervised pipeline improves our model's performance on labeled and unlabeled data. A simple process for creating pseudo-labels is investigated for our task. The pseudo labels are generated using an ATS-HRNet model trained on the annotated datasets. Figure 3 shows the overall framework for extracting pseudo-labels.

3.4 Objective function

The objective function adopted is a weighted combination of a supervised loss $L_{S\text{-}CLS}$ and a semi-supervised loss $L_{SSL\text{-}CLS}$. For the localization step, a weighted binary cross-entropy loss (L_{loc}) is adopted for training the network. This provides a weighting function for the two-class problem, i.e., building vs. nonbuilding.

$$L_{loc} = -(\alpha * y log(\hat{y}) + (1 - y) log(1 - \hat{y})) \qquad (1)$$

α is set to 1.5 to assign higher weights for positive classes.

The classification loss is a weighted combination of the multi-label cross-entropy loss (L_{mce}) and the dice loss. The dice loss (L_{DL}) assesses the overlap between two regions and improves the ability of the model to generate finer boundaries for the segmentation task. As shown in equation 3, the numerator evaluates the overlap between the two regions at a local scale. At the same time, the denominator considers the total number of boundary pixels on a global scale[3].

[2] *https://github.com/DIUx-xView/xView2_fifth_place*

[3] *https://medium.com/ai-salon/understanding-dice-loss-for-crisp-boundary-detection-bb30c2e5f62b*

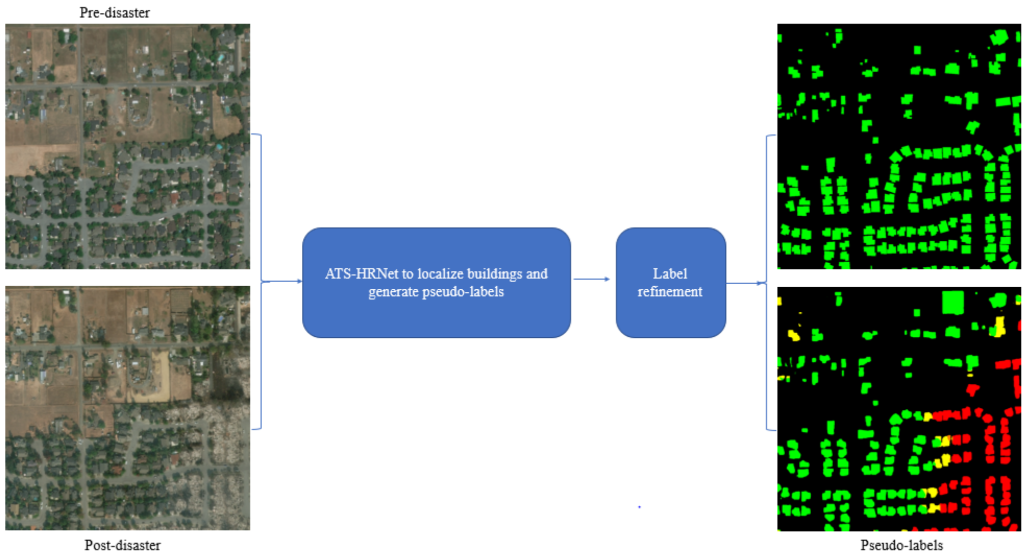


Figure 3: Pseudo-label generation pipeline takes a pair of pre- and post-disaster images and generates pseudo-labels of each pair [28].

$$L_{mce} = -\sum_{c=1}^{4} \alpha_c y(c) \log p(c) \qquad (2)$$

$$L_{DL} = \frac{2|X \cap Y|}{|X| + |Y|} \qquad (3)$$

Where,

X and Y are the ground truth and predicted label, respectively,
α_c = 0.5, 1.5, 1.0, 1.0 for class labels no damage, minor damage, major damage, destroyed.
y(c) and p(c) are the ground truth and predicted labels, respectively.

The classification loss is, therefore,

$$L_{S-CLS} = L_{mce} + \beta L_{DL} \qquad (4)$$

The total loss when training with unlabeled data

$$L_{total} = L_{S-CLS} + \gamma L_{SSL-CLS} \qquad (5)$$

$L_{SSL-CLS}$ is same as L_{S-CLS} but computed separately on the pseudo-generated labels.

4. EXPERIMENTS AND RESULTS

4.1 xBD Dataset

The xBD dataset [35] is a large-scale public dataset released alongside the xView2 Challenge [36] to advance humanitarian assistance and disaster recovery research. It provides a large number of annotated image pairs for change detection and building damage assessment. It comprises 850,736 annotated buildings covering 45,362 km^2 of imagery and covers a diverse set of disasters across multiple geographical locations. Figure 4 shows image samples from the xView2 dataset, and Table 1 presents the original paper's explanation of the Joint Damage Scale (JDS) for categorizing building damage for the xBD dataset.

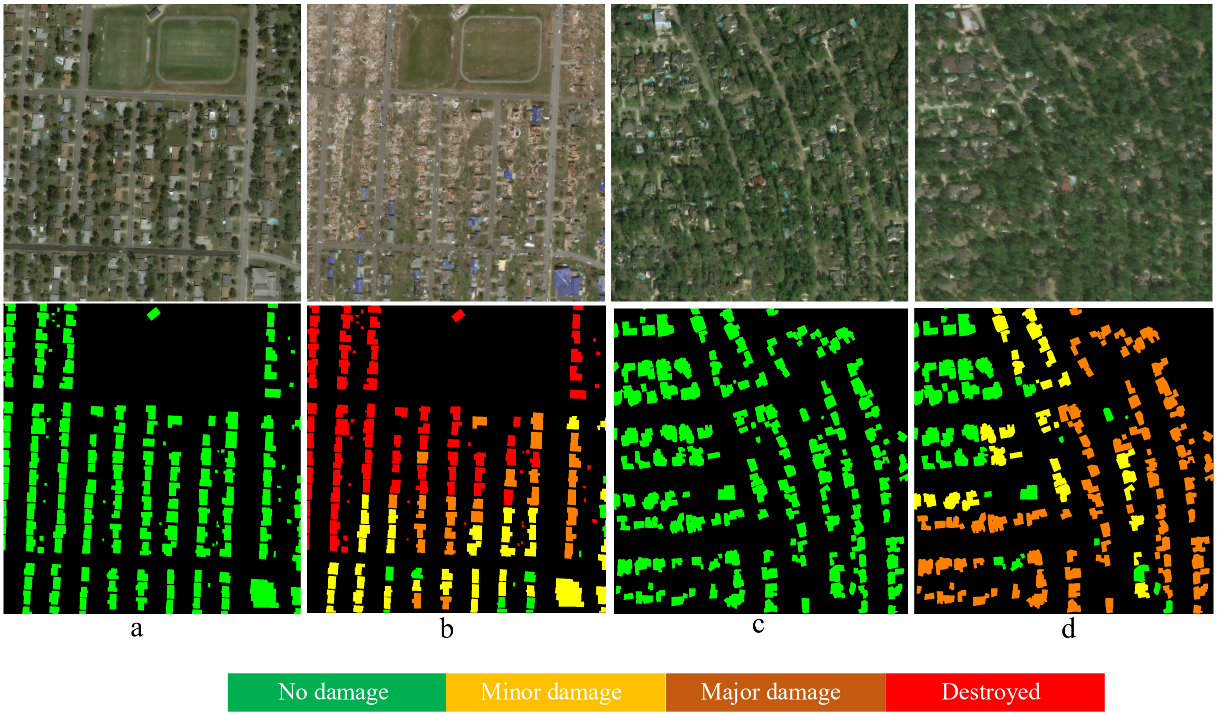


Figure 4: Sample image pairs from the xBD dataset (a) and (b) are pre- and post-disaster images from tornadoes, (c) and (d) and pre- and post-disaster images from hurricane Harvey.

Table 1: Joint Damage Scale Descriptions [35]

Damage level	**Structure Description**
0 (No damage)	Undisturbed. No sign of water, structural or shingle damage, or burn marks.
1 (Minor damage)	Building partially burnt, water surrounding structure, volcanic flow nearby, roof elements missing or visible cracks
2 (Major damage)	Partial wall or roof collapse, encroaching volcanic flow, or surrounded by water/mud
3 (Destroyed)	Scorched, completely collapsed, partially/completely covered with water/mud, or otherwise no longer present

The xBD dataset comprises three folders, as shown in Table 2. The image pairs from the Train and Tier3 folder combined to form the training and validation set. The split used was 95% training and 5% validation. The images on the held-out folder were used for our model evaluation.

Table 2: Size of xBD dataset

Folder	**Number of image pairs**
Train	2799
Tier3	6369
Held-out	933

4.2 Data Preprocessing

Two major steps are taken when preprocessing the xBD and Inria datasets:

i. Images and labels in the xBD dataset are of size 1024 x 1024, while the Inria datasets are 5000x5000. Because the input resolution of our network is 512x512, each image and label pair are cropped into non-overlapping patches of 512x512.

ii. Image pairs in which the pre-disaster contains no buildings or less than three small buildings are removed from the training dataset. This ensures that the model isn't fed a large number of images with no useful training information.

After the preprocessing stage, the number of image pairs in the train and tier3 folders was reduced to 2096 and 2833, respectively.

4.3 Data Augmentation

Figure 5 shows the distribution of the classes for the post-disaster images in the xBD dataset. The plot shows a severe class imbalance with the combined number of buildings with no damages being 10x more than any other class hence the need for a data preprocessing step to reduce the effect of this imbalance on the model's performance. This would prevent the network from being biased towards dominant classes.

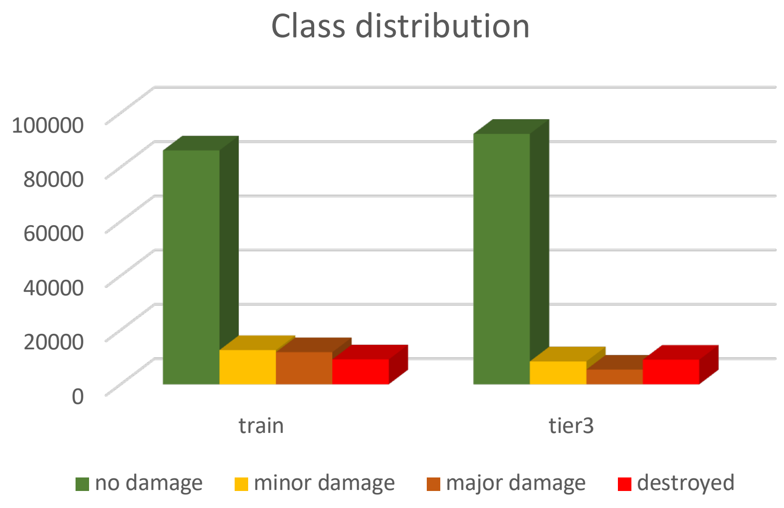


Figure 5: Chart of building damage class distribution for train and tier3 folders images. The number of buildings with no damages is significantly higher than the other three classes.

Having reduced the size of the dataset to contain images with needed training information, an additional augmentation strategy is explored for improving our model's performance. CutMix [29] strategy involves cutting and pasting patches among training images. Shen et al. [37] adopted CutMix for data augmentation by performing the cut and paste strategy on random samples from the training set. Although the same strategy is adopted, a check is performed to ensure CutMix is only performed for post-images with very few buildings. We also ensured that the cut images were from image regions with several building damage classes present. Figure 6 shows the pipeline for our modified CutMix augmentation strategy.

Images with few buildings and annotations are selected from the training set, and this includes the pre- and post-image pairs X_A^{pre}, X_A^{post} and its corresponding label, Y_A. The number of patches needed, N is computed, and N random samples of densely annotated image pairs are selected for extracting N patches which are used to create the new image pairs $\hat{X}_A^{pre}, \hat{X}_A^{post}$ and label $\hat{Y}_A$. The augmented dataset generated using this strategy will be provided here[4].

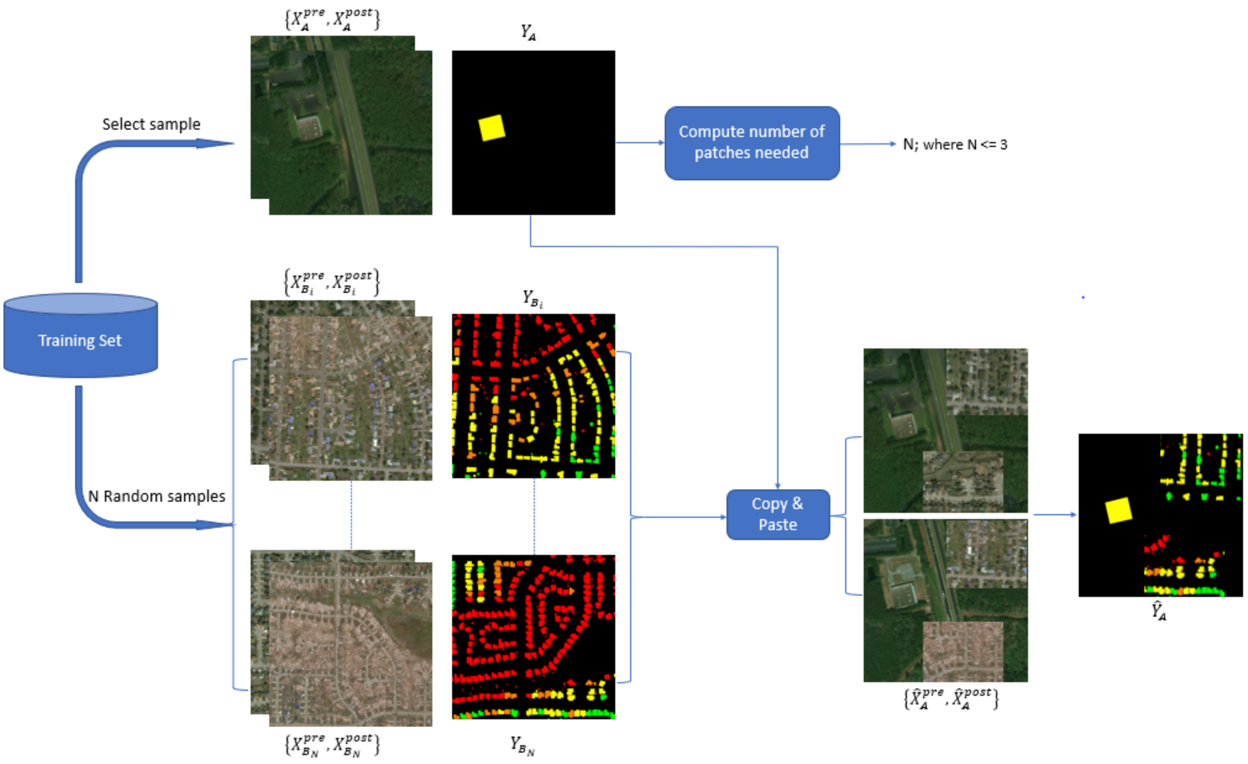

$\{X_A^{pre}, X_A^{post}\}$
Y_A
Select sample
Compute number of patches needed
N; where N <= 3
Training Set
$\{X_{B_i}^{pre}, X_{B_i}^{post}\}$
Y_{B_i}
N Random samples
Copy & Paste
$\hat{Y}_A$
$\{\hat{X}_A^{pre}, \hat{X}_A^{post}\}$
$\{X_{B_N}^{pre}, X_{B_N}^{post}\}$
Y_{B_N}

Figure 6: Data augmentation pipeline using our modified CutMix strategy

Figure 7 shows image samples from SA-CutMix-ed data. SA-CutMix is able to preserve most of the information contained in the original images while randomizing the number of image patches needed for copy-pasting.

4.4 Evaluation Metrics

Evaluation is performed on the 933 pre- and post-image pairs on the held-out data from the xBD dataset. The F1 score provides a metric for evaluating the performance of our model. The F1 score is adopted to evaluate the model's localization and classification tasks performance. It takes into account four important values; the true positive value (TP) represents the number of pixel classes that are correctly classified, true negative value (TN) represents the number of negative classes correctly classified, false-positive value (FP) is the number of negative classes misclassified as a positive and false negative value (FN) is the number of positive classes misclassified as negative. These metrics are shown in equations (1) – (4). Equation (5), $F1_{cls}$, is the harmonic mean of class-wise damage classification F1.

$$precision = \frac{TP}{TP + FP} \qquad (6)$$

$$recall = \frac{TP}{TP + FN} \qquad (7)$$

[4] *https://drive.google.com/drive/folders/1ksNiTChUy2ikFbUhCHioJk1_24_aDkbc?usp=sharing*

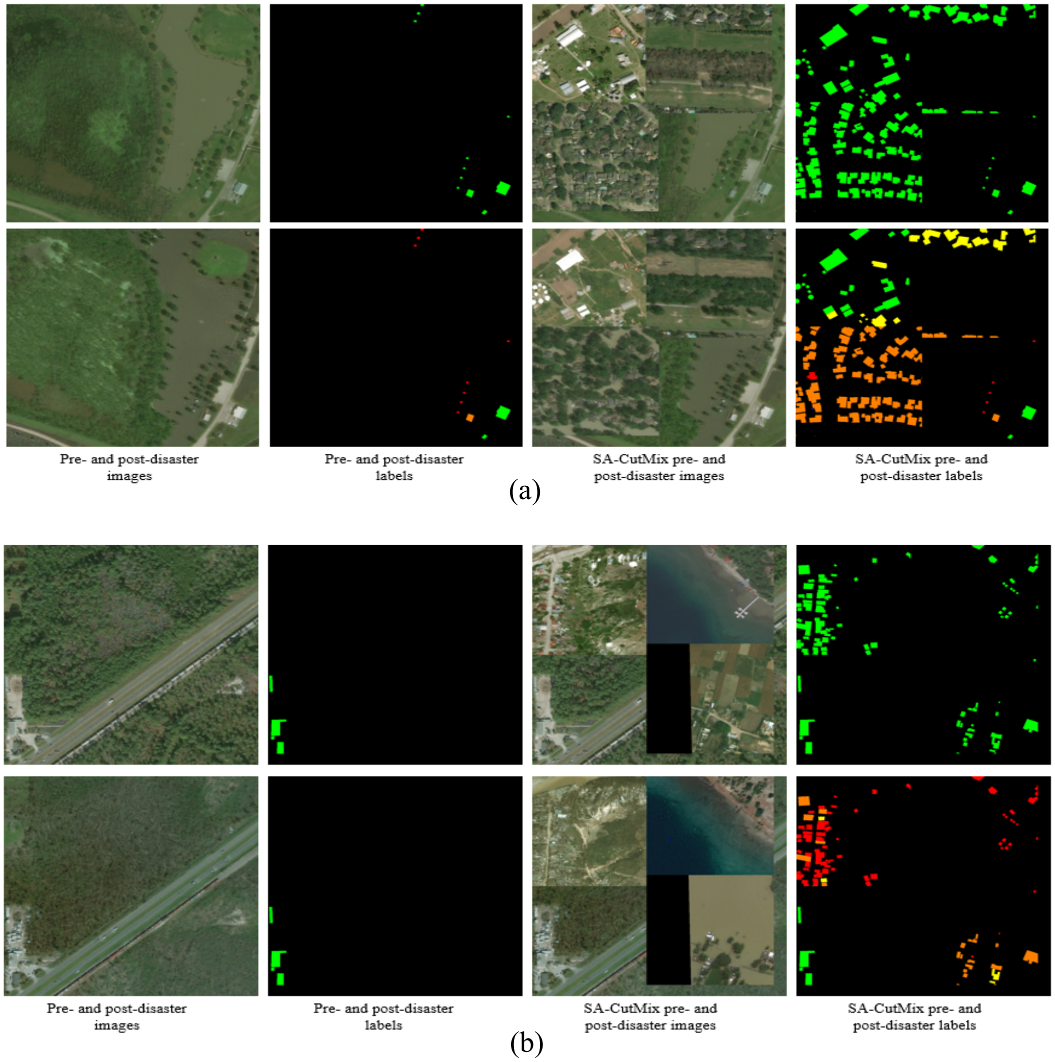

Pre- and post-disaster images
Pre- and post-disaster labels
SA-CutMix pre- and post-disaster images
SA-CutMix pre- and post-disaster labels
(a)
Pre- and post-disaster images
Pre- and post-disaster labels
SA-CutMix pre- and post-disaster images
SA-CutMix pre- and post-disaster labels
(b)

Figure 7: Visualization of the image pairing using the proposed SA-CutMix technique. Images in this figure are illustrative examples of pre- and post-image pairing. (a) and (b) are two separate examples. The first two columns for each example represent the pre-and post-disaster images and the corresponding labels, while the third and fourth columns illustrate the post-disaster images and associated labels.

$$F1_{loc} = 2\,x\,\frac{precision\ x\ recall}{precision\ +\ recall} \qquad (8)$$

$$F1_{loc} = \frac{2TP}{2TP + FP + FN} \qquad (9)$$

$$F1_{cls} = \frac{n}{\sum_{i=1}^{n}\frac{1}{F1_{C_i}}} \qquad (10)$$

$\frac{1}{F1_{C_i}}$ denotes the *F1* score of each damage level (C_i) for damage assessment. Due to the tendency of the *F1* score to heavily penalize over-represented classes, the overall score, $F1_s$ provides a more comprehensive evaluation metric for building segmentation and damage assessment [35].

$$F1_s = 0.3\,x\,F1_{loc} + 0.7\,x\,F1_{cls} \qquad (11)$$

4.5 Implementation Details

Training, validation, and testing were carried out on an Nvidia RTX 3090 GPU. A Stochastic Gradient Descent (SGD) optimizer [38] with a base learning rate of 0.01, the momentum of 0.9, and weight decay of 0.0005 is used for model optimization. For training, the batch size was set to 24, and the batch size for testing was set to 16. The weights for the two HRNet V2 (C=48) used in the localization and classification streams are initialized with ImageNet pretrained weights[5] , and the convolution layers in the feature fusion block are randomly initialized. The algorithm is implemented using the Pytorch deep learning framework [39].

4.6 Results

Table 3 shows the quantitative evaluation of the performance of ATS-HRNet to other state-of-the-art approaches. Figure 8 and Figure 9 show the results for localization and classification of building damage on a subset of the held-out data containing groundtruth annotations. Images shown in Figure 8 are selected from different regions, with different disaster types, building density, and class of building damage. ATS-HRNet + proposed SA-CutMix significantly outperforms the baseline model as well as other joint localization and classification models.

Table 3: Quantitative comparison of F1 scores with other methods. Results show that ATS-HRNet + SA-CutMix outperforms the baseline method.

	Overall score	**Localization**	**No damage**	**Minor damage**	**Major damage**	**Destroyed**
xBD baseline	0.265	-	0.663	0.144	0.009	0.466
Weber and Kane [40]	0.741	0.835	0.906	0.493	0.722	0.837
TS-HRNet	0.645	0.836	0.857	0.370	0.664	0.774
TS-HRNet + SSL	0.745	0.849	0.910	0.528	0.751	0.790
ATS-HRNet	0.756	0.880	0.920	0.629	0.782	0.864
ATS-HRNet + SA-CutMix	**0.778**	**0.892**	**0.928**	**0.640**	**0.796**	**0.875**

[5] *https://github.com/HRNet/HRNet-Image-Classification/blob/master/README.md*

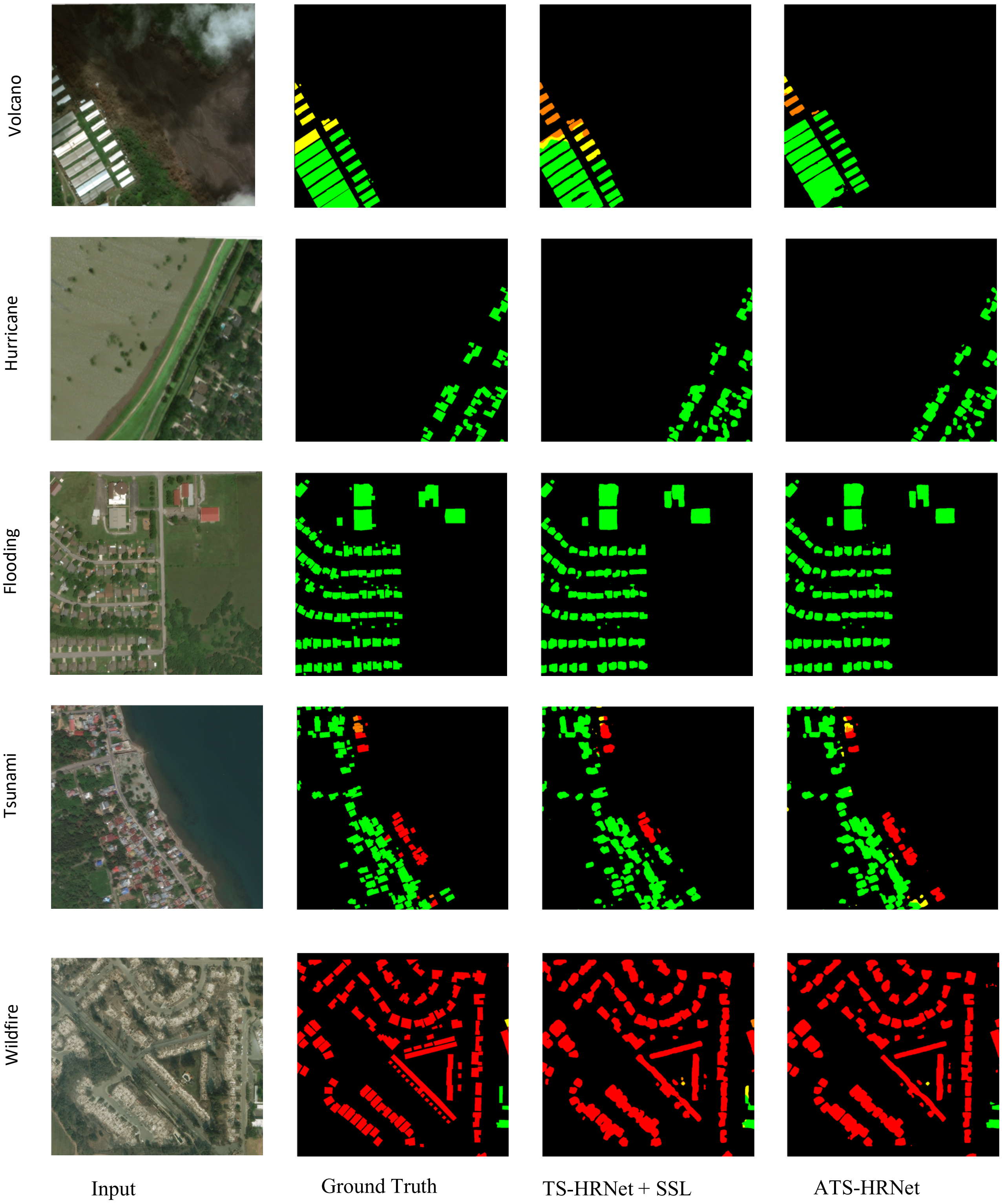


Figure 8: Localization and classification results for five disaster types. The buildings are colored based on the damage type and color mapping in figure 1. The results shows that the ATS-HRNet still outperforms TS-HRNet trained with pseudo-labels.

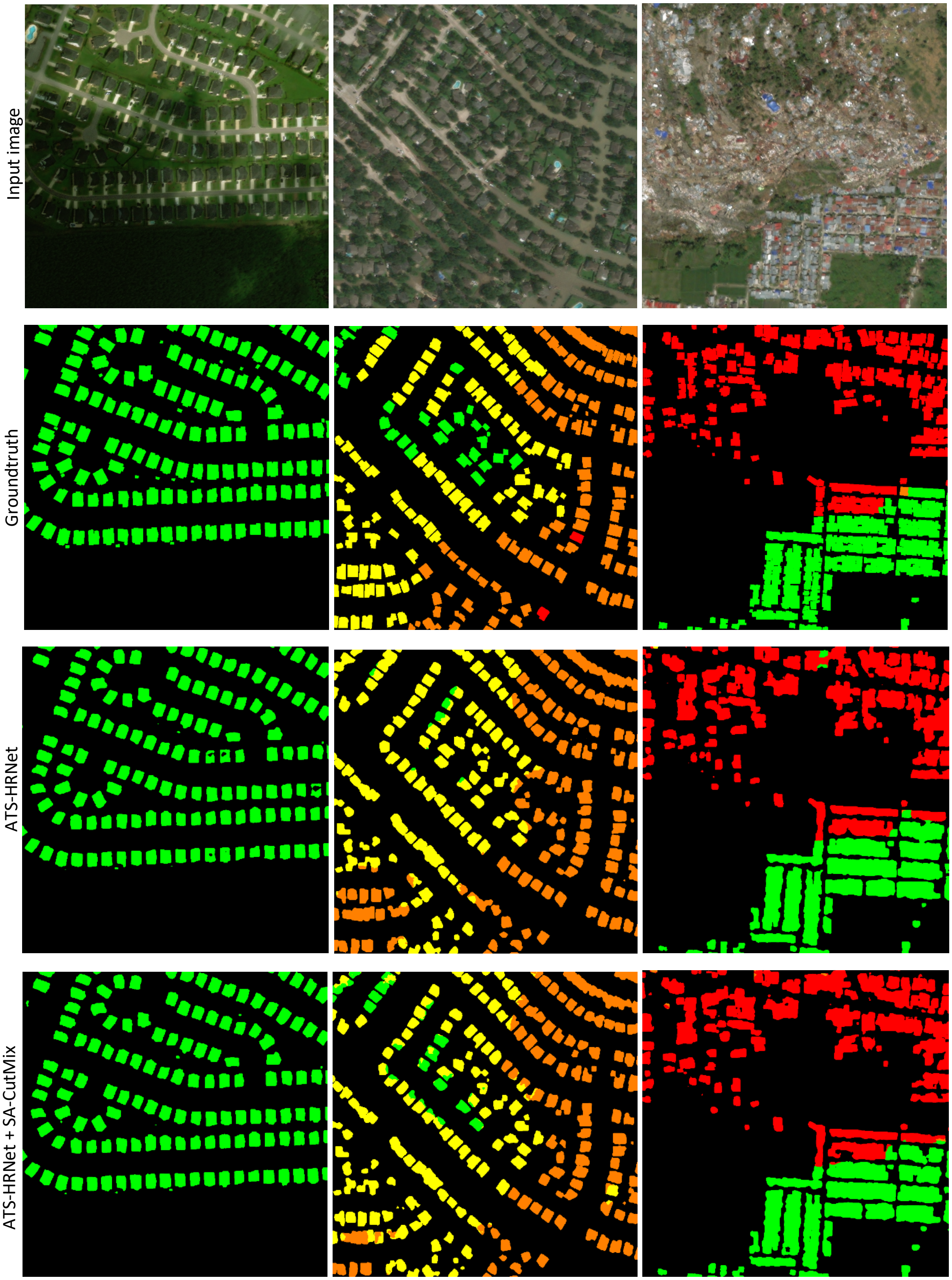


Figure 9: Localization and classification results presenting the benefits of SA-CutMix as a data augmentation strategy. These results show that SA-CutMix helps reduce the amount of misclassified building damages and force finer label assignment.

5. CONCLUSION AND FUTURE WORK

This paper introduces a modified Cutmix strategy for effective data augmentation and an attention-guided high-resolution network (ATS-HRNet) for joint building localization and damage classification. Furthermore, a semi-supervised approach for improving the model's performance for the localization and classification tasks using a set of unannotated is also presented. Our ablation experiments on the xBD demonstrate the effectiveness of the augmentation strategy and ATS-HRNet for effectively localizing buildings of various types and for damage classification. The model's performance is also compared to current state-of-the-art approaches, and the results show that ATS-HRNet + SA-cutmix significantly outperforms other methods across all building damage levels.

Model evaluation for building localization on the xBD dataset still shows some limitations of our strategy. Building styles vary significantly across regions and cultures; hence, it is impractical to account for the vast variability in building structures. As part of the future work, we will be training our framework on more datasets targeted towards building localization from satellite imagery across more regions, such as the African building dataset provided by Sirko et al. [41], to fully evaluate the potential of the proposed approach.

REFERENCES

[1] "Effects | Facts – Climate Change: Vital Signs of the Planet." [Online]. Available: https://climate.nasa.gov/effects/. [Accessed: 08-Nov-2021].

[2] "Environmental health in emergencies." [Online]. Available: https://www.who.int/teams/environment-climate-change-and-health/emergencies. [Accessed: 08-Nov-2021].

[3] C. K. Huyck, B. J. Adams, S. Cho, H. C. Chung, and R. T. Eguchi, "Towards Rapid Citywide Damage Mapping Using Neighborhood Edge Dissimilarities in Very High-Resolution Optical Satellite Imagery—Application to the 2003 Bam, Iran, Earthquake:," *https://doi.org/10.1193/1.2101907*, vol. 21, no. SUPPL. 1, Dec. 2019.

[4] G. Trianni and P. Gamba, "Damage Detection from SAR Imagery: Application to the 2003 Algeria and 2007 Peru Earthquakes," *Int. J. Navig. Obs.*, vol. 2008, pp. 1–8, Aug. 2008.

[5] H. D. Guo, X. Y. Wang, X. W. Li, G. Liu, L. Zhang, and S. Y. Yan, "Yushu earthquake synergic analysis using multimodal SAR datasets," *Chinese Sci. Bull. 2010 5531*, vol. 55, no. 31, pp. 3499–3503, Nov. 2010.

[6] Y. Dong, Q. Li, A. Dou, and X. Wang, "Extracting damages caused by the 2008 Ms 8.0 Wenchuan earthquake from SAR remote sensing data," *J. Asian Earth Sci.*, vol. 40, no. 4, pp. 907–914, Mar. 2011.

[7] M. Liao, L. Jiang, H. Lin, B. Huang, and J. Gong, "Urban Change Detection Based on Coherence and Intensity Characteristics of SAR Imagery," *undefined*, vol. 74, no. 8, pp. 999–1006, 2008.

[8] X. Tong *et al.*, "Building-damage detection using pre- and post-seismic high-resolution satellite stereo imagery: A case study of the May 2008 Wenchuan earthquake," *ISPRS J. Photogramm. Remote Sens.*, vol. 68, no. 1, pp. 13–27, Mar. 2012.

[9] X. Tong *et al.*, "Use of shadows for detection of earthquake-induced collapsed buildings in high-resolution satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 79, pp. 53–67, May 2013.

[10] C. Hua, J. Qi, H. Shang, W. Hu, and J. Han, "Detection of collapsed buildings with the aerial images captured from UAV," *Sci. China Inf. Sci. 2015 593*, vol. 59, no. 3, pp. 1–15, Dec. 2015.

[11] L. Ke, Y. Lin, Z. Zeng, L. Zhang, and L. Meng, "Adaptive Change Detection with Significance Test," *IEEE Access*, vol. 6, pp. 27442–27450, Feb. 2018.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks."

[13] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2015.

[14] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” Apr. 2020.

[15] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “CenterNet: Keypoint Triplets for Object Detection.” pp. 6569–6578, 2019.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016.

[18] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation.” pp. 3431–3440, 2015.

[19] H. Noh, S. Hong, and B. Han, “Learning Deconvolution Network for Semantic Segmentation.” pp. 1520–1528, 2015.

[20] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT: Real-Time Instance Segmentation.” pp. 9157–9166, 2019.

[21] J. Z. Xu, G. Ai, W. L. Google, Z. Li, P. Khaitan, and V. Zaytseva, “Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks,” Oct. 2019.

[22] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, and R. Nakamura, “Damage detection from aerial images via convolutional neural networks,” *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. MVA 2017*, pp. 5–8, Jul. 2017.

[23] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese Neural Networks for One-shot Image Recognition.”

[24] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, May 2015.

[25] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, “Fine-Grained Building Change Detection From Very High-Spatial-Resolution Remote Sensing Images Based on Deep Multitask Learning,” *IEEE Geosci. Remote Sens. Lett.*, pp. 1–5, Sep. 2020.

[26] I. Bayramli, E. Bondi, M. Tambe, H. John, and A. Paulson, “In the Shadow of Disaster: Finding Shadows to Improve Damage Detection.”

[27] J. Wang *et al.*, “Deep High-Resolution Representation Learning for Visual Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Aug. 2019.

[28] V. Oludare, L. Kezebou, K. Panetta, and S. S. Agaian, “Semi-supervised learning for improved post-disaster damage assessment from satellite imagery,” *https://doi.org/10.1117/12.2586232*, vol. 11734, pp. 172–182, Apr. 2021.

[29] S. Yun, D. Han, S. Chun, S. J. Oh, J. Choe, and Y. Yoo, “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-October, pp. 6022–6031, May 2019.

[30] H. Xiao, Y. Peng, H. Tan, and P. Li, “Dynamic Cross Fusion Network for Building-Based Damage Assessment,” pp. 1–6, Jun. 2021.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module.”

[32] L. Chen *et al.*, “SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning.”

[33] H. Liu, F. Liu, X. Fan, and D. Huang, “Polarized Self-Attention: Towards High-quality Pixel-wise Regression,” Jul. 2021.

[34] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-Cross Attention for Semantic Segmentation,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-October, pp. 603–612, Nov. 2018.

[35] R. Gupta *et al.*, “xBD: A Dataset for Assessing Building Damage from Satellite Imagery,” Nov. 2019.

[36] “DIU’s xVIEW2 - Assessing Building Damage | Challenge.gov.” [Online]. Available: https://www.challenge.gov/challenge/diu-xview2-assessing-building-damage/. [Accessed: 12-Nov-2021].

[37] Y. Shen *et al.*, “BDANet: Multiscale Convolutional Neural Network with Cross-directional Attention for Building Damage Assessment from Satellite Images,” *IEEE Trans. Geosci. Remote Sens.*, May 2021.

[38] S. Ruder, “An overview of gradient descent optimization algorithms *.”

[39] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” 2019.

[40] E. Weber and H. Kané, “Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion,” Apr. 2020.

[41] W. Sirko *et al.*, “Continental-Scale Building Detection from High Resolution Satellite Imagery,” 2021.