AUTOMATED BUILDING SEGMENTATION AND DAMAGE ASSESSMENT FROM SATELLITE IMAGES FOR DISASTER RELIEF

X. Yuan[1,*], S. M. Azimi[1,*], C. Henry[1], V. Gstaiger[1], M. Codastefano[3], M. Manalili[3], S. Cairo[3], S. Modugno[3],
M. Wieland[2], A. Schneibel[2], N. Merkle[1]

[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
{xiangtian.yuan,seyedmajid.azimi,corentin.henry,veronika.gstaiger,nina.merkle}@dlr.de
[2] German Remote Sensing Data Center, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
{marc.wieland,anne.schneibel}@dlr.de
[3] UN World Food Programme, Rome, Italy
{marco.codastefano,michael.manalili,stefano.cairo,sirio.modugno}@wfp.org

ICWG III/IVa

KEY WORDS: Satellite Imagery, Damage Assessment, Deep Learning, Building Segmentation, Crisis Management

ABSTRACT:

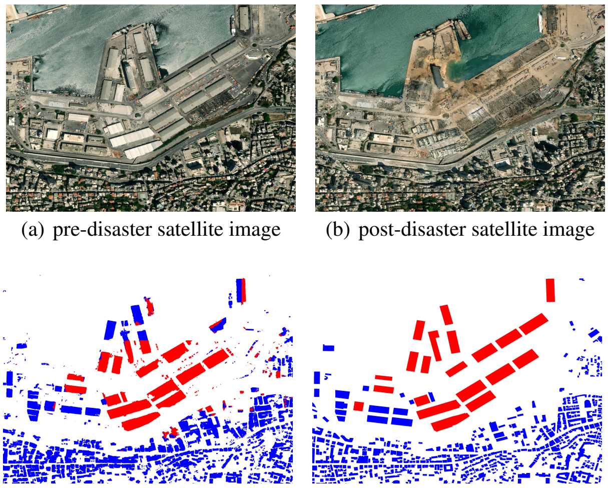
After a natural disaster or humanitarian crisis, rescue forces and relief organisations are dependent on fast, area-wide and accurate information on the damage caused to infrastructure and the situation on the ground. This study focuses on the assessment of building damage levels on optical satellite imagery with a two-step ensemble model performing building segmentation and damage classification trained on a public dataset. We provide an extensive generalization study on pre- and post-disaster data from the passage of the cyclone Idai over Beira, Mozambique, in 2019 and the explosion in Beirut, Lebanon, in 2020. Critical challenges are addressed, including the detection of clustered buildings with uncommon visual appearances, the classification of damage levels by both humans and deep learning models, and the impact of varying imagery acquisition conditions. We show promising building damage assessment results and highlight the strong performance impact of data pre-processing on the generalization capability of deep convolutional models.

1. INTRODUCTION

Every year, thousands of people lose their homes due to natural disasters and technical accidents and are dependent on external help. In order to be able to help effectively, aid organisations need information on the affected regions. Creating a fast, large-scale and reliable damage assessment map is a big challenge faced by emergency response teams after a disaster. Geo-information derived from remote sensing satellite data have been used for years to help organizing and coordinating rescue activities (Voigt et al., 2016) and as the amount of available remote sensing data is constantly increasing, methods applied to semiautomatic impact assessment experienced an exponential rise in implementation (Ma et al., 2019).

Within the framework of the "Humanitarian Technologies" initiative, the outcomes of the German Aerospace Center's (DLR) research and development are put into application, such as in the "Data4Human" project (DLR, 2020). In this project, DLR is working together with the United Nations World Food Programme (WFP), the Human Rights Watch (HRW), the German Red Cross (DRK), the Humanitarian OpenStreetMap Team (HOT) and the United Nations Development Programme (UNDP) to make geo-information and remote sensing data more usable for humanitarian relief missions.

During rapid mapping activities, damage to buildings and infrastructure has largely been assessed manually by interpreters comparing earth observation data from before and after the event. In order to obtain an area-wide damage mapping, a lot of time or a large number of interpreters is therefore necessary. The extraction of building footprints, are often the first step of the damage assessment process. In this way, the damage can not only be more precisely delineated, but also quantitatively better documented. Building footprints do not only support manual classification, but are also the first step to an automated damage classification. On the one hand it provides fast and reliable results and on the other it does not tie up personnel for time-consuming manual analyses. In recent years,



(a) pre-disaster satellite image (b) post-disaster satellite image

(c) predictions (d) building footprints (GT)

Figure 1. Building damage assessment after the explosion in the city of Beirut, Lebanon, in 2020. Color coding of the predictions and the Ground Truth (GT): 🟦 non-destroyed and 🟥 destroyed.

[*] Corresponding authors, denotes equal contribution

convolutional neural networks (CNNs) have been extensively explored for the task of automatic building segmentation. Typical CNN-based frameworks in building segmentation use an encoder-decoder structure. U-Net has been one of the most popular base architectures in optical satellite image segmentation thanks to the skip-connections that retain fine-grained information. For example in (Hamaguchi, Hikosaka, 2018) a multi-task U-Net was employed, of whom each branch is responsible for different building sizes, as well as roads. The model reached the first place in the DeepGlobe-CVPR18 challenge. In the second place solution of the same challenge, TernausNetV2 (Iglovikov et al., 2018) also adopted a U-Net-based network that integrates an additional channel to predict whether building boundaries are touching, and henceforth an instance map can be obtained.

For the assessment of building damage, reference is often made to the European Macroseismic Scale 1998 standard (EMS-98) (Grünthal et al., 1998), which includes a description of earthquake magnitudes and damage classes on buildings of different construction. However, the five damage classes (1=negligible to slight damage, 2=moderate damage, 3=substantial to heavy damage, 4=very heavy damage and 5=destruction) cannot be fully applied when assessing remote sensing data. On the one hand, minor damage, e.g. cracks in a wall, cannot be detected from a top-down viewing angle, and on the other hand the damage classes cannot be subdivided finely enough in image data with about 50 cm ground resolution. For these reasons, the organisations dealing with damage assessment based on satellite data had to reduce the number of damage classes and, for example, the International Working Group on Satellite-based Emergency Mapping (IWG-SEM) proposed a working paper for building damage assessment in their Emergency Mapping Guidelines with four damage classes (no visible damage, possibly damaged, damaged and destroyed) (IWG-SEM, 2018). The Copernicus Emergency Management Service (© European Union, 2012-2021) currently refers to this standard in its grading products. Still, this classification might be very ambitious depending on the ground resolution of the data and the structural characteristics of the buildings. Less damage classes can also be found in other studies (Ghosh et al., 2011). Here, the authors used two and three damage classes, respectively, depending on the availability of high-resolution satellite or aerial imagery.

In recent years, more and more research studies have been conducted to investigate deep learning-based approaches to the task of building damage assessment. Weber et al. (Weber, Kané, 2020) integrate the two steps for building segmentation and damage classification into one. The pre- and post-disaster images are fed into a two-branch backbone network with shared weights and the feature maps are concatenated before being fed into a damage classification network. Multi-model damage classification has also been studied in (Adriano et al., 2020). Here, optical and synthetic aperture radar (SAR) images are separately fed into the two encoder streams of an attention U-Net. The feature maps from two streams are concatenated and skip-connected to the decoder. In order to encourage the development of machine learning and computer vision solutions for building damage assessment, a challenge named xView2 (DIU, 2019) was announced in 2019. Here, the participants' solutions were expected to locate buildings and assign a damage level for each one, based on pre- and post-event satellite images.

Inspired by the proposed solutions of the xView2 challenge, we investigate the performance of a two-stage building segmentation and damage classification fully-convolutional network trained on the xBD dataset (Gupta et al., 2019). In contrast to the xView2 challenge, we study the generalization capability of our models on different pre- and post-disaster data from the MAXAR's open data program[1]. Therefore, we selected two test areas and created Ground Truth (GT) data for the building segmentation and damage assessment.

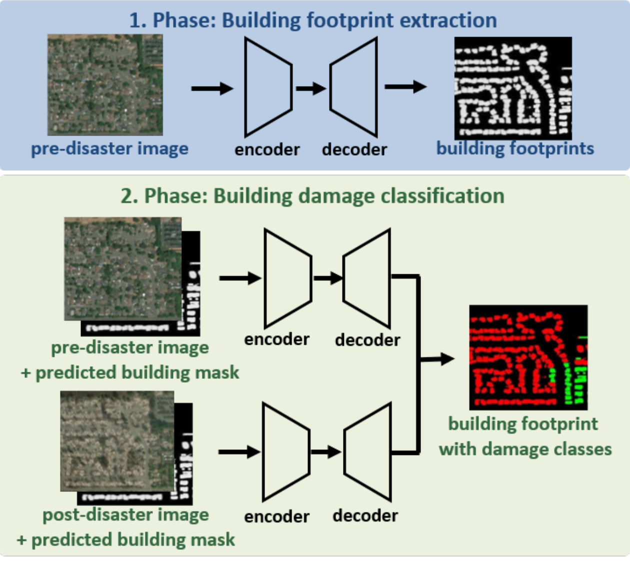


Figure 2. Network architecture of the xView2 challenge winner.

2. METHODS

In this paper, we are following the solution proposed by the xView2 challenge winning team (Durnov, 2019). Their approach consists of two phases, which are illustrated in Figure 2. In the following subsections, we are describing both phases and additionally we are comparing the proposed building segmentation network with a state-of-the-art approach and investigating the influence of image pre-processing techniques on the performance of the models, more specifically image co-registration and image fusion. The training details are described in section 2.2.

2.1 Building Segmentation

The first step towards building damage classification is to identify existing buildings from the pre-disaster imagery. We use the winning team solution of the xView2 challenge to this end. The building footprints are extracted from the pre-event satellite images using an ensemble of U-Net-based neural network architectures with encoders (ResNet34 (He et al., 2016), SE-ReNeXt50 (Hu et al., 2018, Xie et al., 2017), SENet154 (Hu et al., 2018), and DPN92 (Chen et al., 2017) (He et al., 2016) (Xie et al., 2017)) pre-trained on ImageNet (Russakovsky et al., 2015). The U-Net-based encoder-decoder networks use skip-connections to retain some of the high-resolution information. Each network is trained separately and the final predictions are derived by fusing and averaging the output of the single networks. More details can be found in (Durnov, 2019).

In order to compare the results with a state-of-the-art approach for building segmentation, we used a network called

[1] https://www.maxar.com/open-data

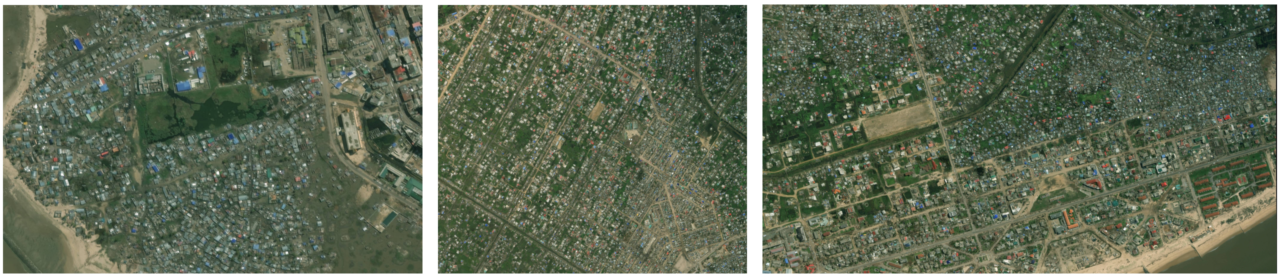

(a) Praia Nova (PN) (b) Macarungo Hospital (MH) (c) Macuti Village (MV)

Figure 3. Overview of the three regions of interest of our dataset in Beira, Mozambique.

HRNet (Sun et al., 2019), which adopts a different segmentation paradigm. More precisely, it connects the high-to-low-resolution convolution streams in parallel, maintains high-resolution through the whole process instead of recovering high-resolution from low-resolution and repeatedly exchanges the information across different resolutions (Wang et al., 2020).

2.2 Multi-resolution Damage Classification

In the second phase, damage classification is performed based on the predicted building footprints and the corresponding pre- and post-disaster satellite image pairs. Therefore, an ensemble of Siamese neural networks is used, where each Siamese branch takes as input the extracted building footprint and either the pre- or the post-event image. In more details, the pre- and post-event branches share the same weights from the localisation models and the last decoder layers are concatenated to derive pixel-wise damage classes. Keeping them separate helps ignoring co-registration differences such as shifts and camera angles. Moreover, a morphological dilation with a 5×5 kernel is applied to the classification masks to improve the accuracy on boundary regions as well as shifts and different nadir views. During training, dice and focal losses are applied to partially address the drift between pre- and post-event images. Also, data augmentation is applied during training by using flipping, rotation, color shifts, blurring and a few more techniques.

As the models extract the building footprints from the pre-disaster images only, cloud occlusion in the pre-event images as shown in Figure 6 (a) and (b), lead to incomplete building detections and therefore to an incomplete damage assessment. For this models, two effects of cloud occlusion can be observed in the damage assessment as follows: 1) if building buildings are not detected in the pre-disaster images, they are also missing in the post-disaster image and 2) if buildings are detected in the pre-disaster images, but are covered by clouds in the post-disaster images, they are commonly classified as ”no-damaged”. To resolve this issue, we investigate the possibility of merging the predictions of the building segmentation modules for various pre-disaster satellite images acquired at different dates. A detailed discussion of the results is provided in Section 4.

2.3 Image Pre-processing

A typical issue in remote sensing data pre-processing is the misalignment between images acquired at different times, which in the solutions of the xView2 challenge were handled by the models only. We instead use an open-source software called AROSICS (Scheffler et al., 2017) for the co-registration of all satellite images used for evaluation and for the co-registration between satellite and UAV images used for the generation of the ground truth data (more details in Section 3.2). This software works as follows: it applies phase correlation in a moving-window manner to a regular grid of coordinate points and estimates X/Y translations for each point. The shift vector grid is validated with a set of quality metrics. The remaining points are then used as tie points to fit an affine transformation model.

3. DATASET

3.1 Training Dataset

The xBD dataset (Gupta et al., 2019) is the largest open-source dataset for building damage assessment, covering over 5000 km^2 across 15 countries, and as such the basis of the xView2 challenge. It comprises pre- and post-event satellite imagery acquired with a Ground Sampling Distance (GSD) of 45 cm/px from five disaster types: earthquake/tsunami, flood, volcanic eruption, wildfire and wind. The dataset encapsulates four damage scales: no damage, minor damage, major damage, and destroyed (more details can be found in (Gupta et al., 2019)).

There are around 425k building instances in the dataset, annotated according to 4 levels of damage: no damage (313k instances), minor damage (36k), major damage (29k) and destroyed (31k). Additionally, around 14k building instances were left unclassified. However, while labeling images from Beira, Mozambique (see Section 3.2), we realized that annotators do not consistently agree on the minor and major damage classification. Whereas destroyed and intact buildings are easy to distinguish, minor versus major damage levels are difficult to assess from satellite imagery alone. This is mostly due to the low image resolution which does not reveal small but critical details on the structural integrity of the buildings such as crack on the walls, and the damage level definitions which do not account for regional architectural features. Therefore, and similarly to (Ghosh et al., 2011), we decide to reduce the problem complexity from 4 damage levels to 3, merging the minor and major damage to a single damaged level.

3.2 Test Dataset

To evaluate the performance and generalization capability of the approaches described above, we collect pre- and post-event WorldView optical images from MAXAR’s open data program. The first event chosen for the evaluation is the cyclone Idai from 2019, which was one of the worst cyclones to have hit Africa in this millennium. It caused a huge humanitarian crisis and economic damage to several countries. The second event chosen, is the explosion in the city of Beirut, Lebanon, in 2020. Due to the explosion, critical infrastructure at the port of Beirut got destroyed and thousands of people lost their homes.

Beira, Mozambique: WFP launched an emergency operation in Mozambique shortly after to provide timely humanitarian assistance to affected population. This operation included, among other things, the acquisition of drone imagery data over the city of Beira in cooperation with Mozambique's National Institute for Disaster Management (INGC). Based on this data, WFP and INGC created a ground truth for building footprints, a sample of which is shown in Figure 6 (c).

Out of this data, three regions of interests (RoI) with a total number of around 8000 building instances are defined (see Figure 3). Together with experts from WFP, we complement these building footprints by manually assigning a damage class label to each building. Therefore, we choose the following three classes: 1) no damage, 2) damaged and 3) destroyed. To avoid annotation bias, we divide our 6 annotators into 2 groups, and each group simultaneously annotates different areas of interest. The annotations are then merged with the following rules: if one person out of the three annotates a building as ”destroyed”, we label it as ”destroyed”. Else, if at least two people label it as ”damaged”, we label it as ”damaged”. Else, if at least two people label it as ”no damage”, we label it as at ”non-damaged”. Otherwise it is ignored during testing due to the disagreement between annotators. In total we have 2731 non-damaged, 4064 damaged and 260 destroyed building instances in the three RoI (plus 882 ignored instances).

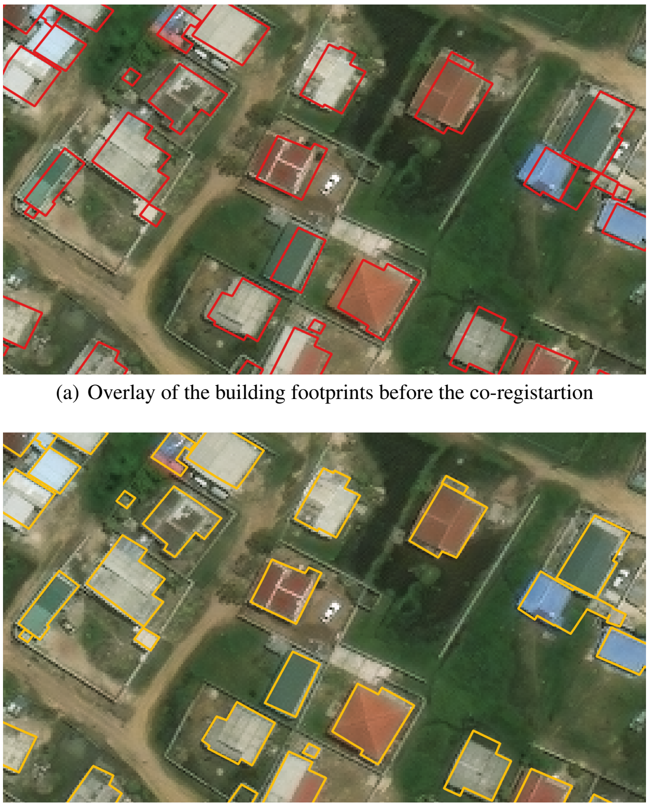
Beirut, Lebanon: For the evaluation of our results for the city of Beirut, GT data provided by the Center for Satellite Based Crisis Information (ZKI) (ZKI, 2021) is used. Shortly after the explosion, ZKI prepared a damage mapping and made it available to the public. The basis of this map is the building footprints from OpenStreetMap and very high-resolution satellite image data provided by EUSI[2]. The correctness of the building footprints is checked by experts, adjusted if necessary and the outlines of completely destroyed building are marked (see Figure 1 (d) where destroyed building are marked red). In our RoI centered around the epicenter of the explosion, the ground truth contains 1307 building footprints of which 40 are labeled as ”destroyed”.

4. EXPERIMENTS AND DISCUSSION

In the following section, we provide a detailed evaluation and discussion of the proposed methods. In addition, we provide an overview of the training procedures of the neural networks, of the conducted experiments for the building segmentation and building damage assessment, and of the influence of the proposed image pre-processing approaches.

4.1 Image Co-registration

As mentioned in Section 2.3, we used an image co-registration method to align all pre- and post-disaster satellite images used for the evaluation of our approaches. Since our GT over the city of Beira was based on UAV imagery, we also applied the image co-registration method here to reduce the misalignment between the satellite images and the GT. In all cases we used one pre-disaster image as the reference image. In Figure 4, a qualitative analysis of the process is provided. There, the misalignment between our GT and the pre-disaster satellite image is shown in Figure 4 (a). Figure 4 (b) illustrates the improved overlay between the GT and the satellite image after the co-registration.


(a) Overlay of the building footprints before the co-registartion

(b) Overlay of the building footprints after the co-registartion

Figure 4. Influence of the image co-registration process on the quality of the ground truth for Beira, Mozambique.

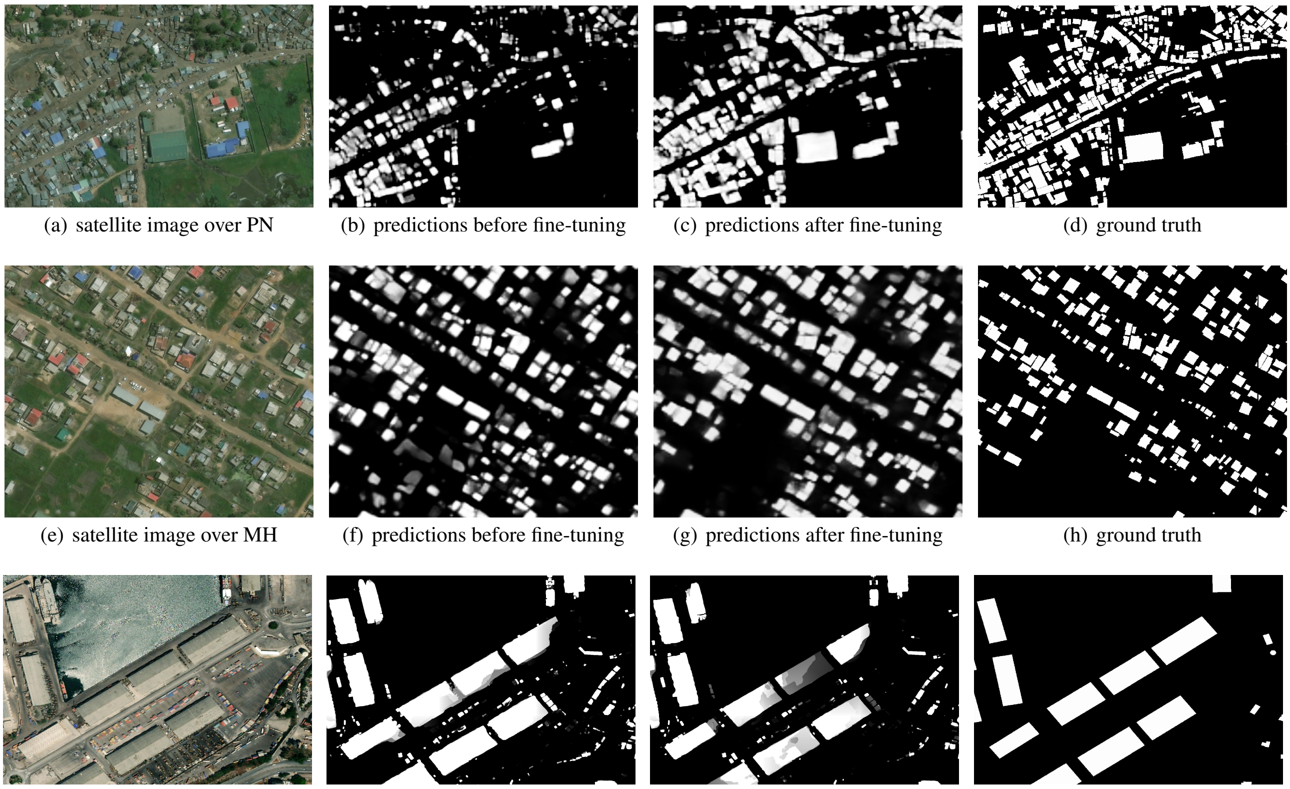
4.2 Building Segmentation

We trained both of the networks described in Subsection 2.1 using the full xBD training dataset introduced in Section 3.1. The U-Net-based ensemble model is trained on pre-disaster images with a dice and a focal losses to mitigate the class imbalance and improve the segmentation accuracy, then fine-tuned over a few epochs on post-disaster images. AdamW (Loshchilov, Hutter, 2017) is selected as the optimizer. During testing, the inputs are flipped vertically and horizontally as well as rotated by 180°. The HRNet on the other hand, is trained with an Online Hard Example Mining (OHEM) loss and optimized by a Stochastic Gradient Descent (SGD) optimizer with momentum.

The building localization results are shown in Table 1 and Figure 5. Both networks are capable of segmenting buildings despite the xBD dataset not encompassing any area in Mozambique or Lebanon. Due to the distinct appearance of the buildings in the Beira dataset compared to the xBD dataset, we investigate the influence of fine-tuning both models on additional data on their performance. Therefore, we use the RoI ”Macuti village” from the Beira dataset as our fine-tuning data. To make full use of the relatively small dataset as well as to avoid overfitting, we empirically set the numbers of epoch during fine-tuning to 40. All other hyper-parameters stayed the same as during training.

For the remaining RoIs from the Beirut dataset, fine-tuning contributes to an increased building segmentation performance. However, the F1 score and Intersection over Union (IoU) drop slightly in the Beirut test area for both models (see Table 1). This is not surprising as the overfitting is expected when fine-tuning on a small dataset. Overall, the fine-tuned HRNet outperforms the ensemble model in the two RoIs in Mozambique

[2] https://www.euspaceimaging.com


(a) satellite image over PN (b) predictions before fine-tuning (c) predictions after fine-tuning (d) ground truth

(e) satellite image over MH (f) predictions before fine-tuning (g) predictions after fine-tuning (h) ground truth

(i) satellite image over Beirut (j) predictions before fine-tuning (k) predictions after fine-tuning (l) ground truth

Figure 5. Examples of the building detection results for the two RoIs Praia Nova (PN) and Macarungo Hospital (MH) in the city of Beira, Mozambique and our test area in the city of Beirut, Lebanon. The results for Beira are generated using the HRNet model while the results for Beirut are derived from applying the U-Net-based ensemble model.

Method	RoI	IoU [%]	F1 Score [%]
xView2-1st	Praia Nova	50.44	67.06
	Hospital	45.24	62.30
	Beirut	**47.06**	**64.00**
xView2-1st (fine-tuned)	Praia Nova	51.33	67.83
	Hospital	47.78	64.66
	Beirut	45.84	62.87
HRNet	PraiaNova	49.07	65.84
	Hospital	48.72	65.52
	Beirut	44.99	62.06
HRNet (fine-tuned)	PraiaNova	**53.95**	**70.09**
	Hospital	**56.97**	**72.59**
	Beirut	42.38	59.53

Table 1. Quantitative evaluation of the building segmentation results for the test areas Praia Nova (PN) and Macarungo Hospital (MH) in the city of Beira, Mozambique, and the test area in Beirut, Lebanon. In this table only the building class is evaluated.

and yields predictions with sharper boundaries, but demonstrates inferior performance in the Beirut test area. The U-Net-based ensemble model, on the other hand, provides more stable results between the various test areas and thus a higher generalization capability.

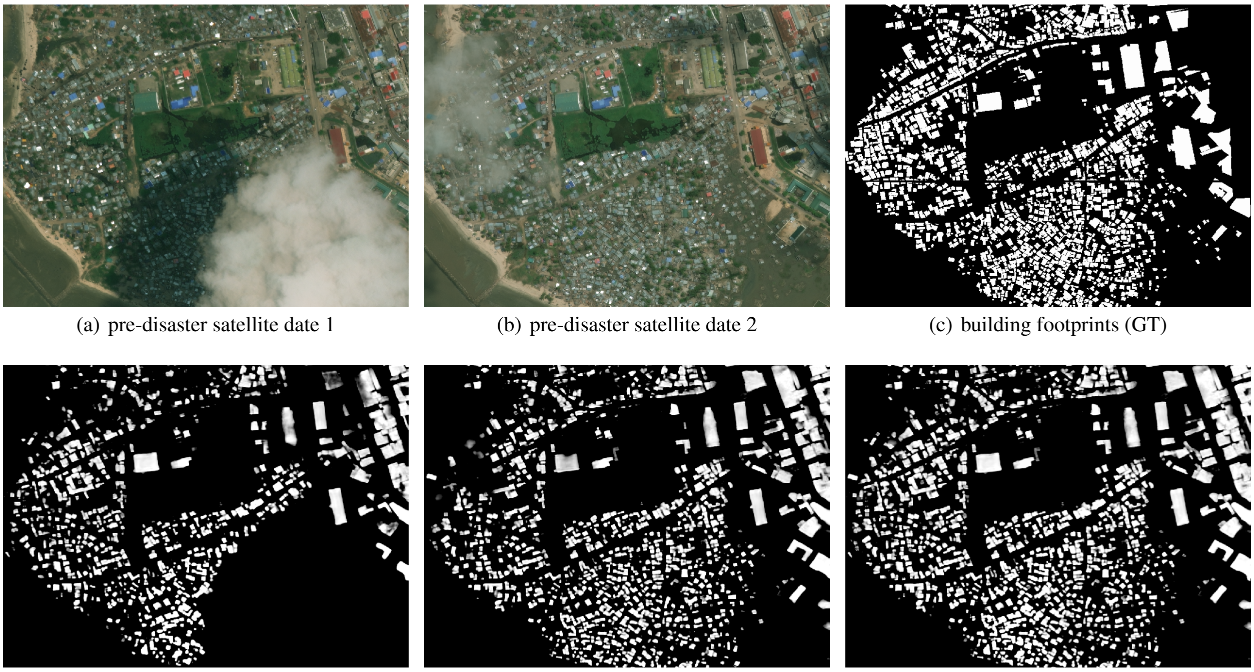
Method	RoI	IoU [%]			F1 score [%]		
		pre1	pre2	fused	pre1	pre2	fused
HRNet (fine-tuned)	PN	38.87	49.94	**53.19**	55.98	66.61	**69.44**

Table 2. Quantitative evaluation of the prediction fusion from two pre-disaster satellite images on the building segmentation for the RoI Praia Nova (PN) in Beira, Mozambique. Here, ”pre1” and ”pre2” relate to the results obtained using two pre-disaster satellite image acquired on date 1 and date 2 respectively.

In addition to building segmentation from a single satellite image, we investigate the use of multiple pre-disaster satellite images. As clouds are often present in satellite imagery, the aim of this experiment is to reduce the impact of their occlusion, which causes some buildings to not be detected by the networks. Therefore, we fuse the building detection results of several pre-disaster satellite images acquired at different dates with each other. Figure 6 provides a qualitative evaluation of the results. Here, the benefits of the images fusion are clearly visible in the complete building mask of Figure 6 (f) compared to (d) and (e). Also the quantitative evaluation provided in Table 2 shows that the fusion improves the performance of the building segmentation.

4.3 Damage Classification

In order to find the best model for the damage classification task, we trained the ensemble network described in Section 2.2 on the xBD dataset with different configurations. For the classification networks, dice, focal and cross-entropy loss functions are used. Similarly to the building segmentation, an Adam optimizer is used.



(a) pre-disaster satellite date 1

(b) pre-disaster satellite date 2

(c) building footprints (GT)

(d) predictions pre1

(e) predictions pre2

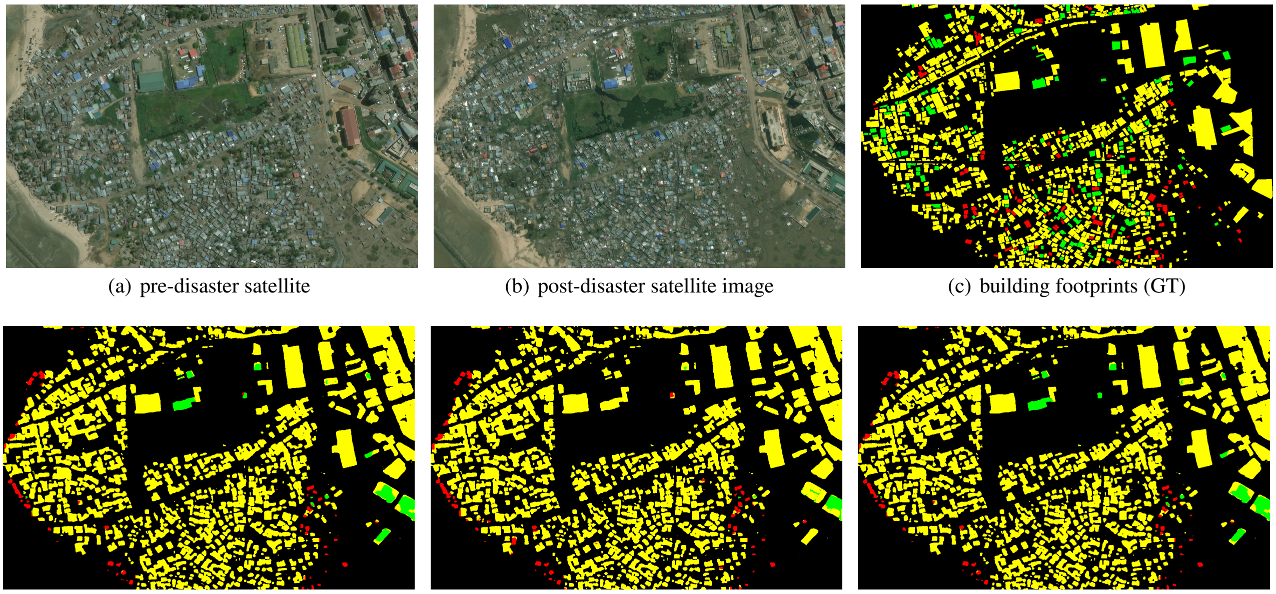
(f) fused predictions

Figure 6. Building segmentation for the RoI Praia Nova in Beira, Mozambique. The results of two pre-disaster satellite images were fused to enhance the results by removing the effects of cloud occlusions on the predictions.

Method	Training dataset	Fine-tuning	RoI	IoU Beira [%]				IoU Beirut [%]		
				mean	no-damage	damaged	destroyed	mean	non-destroyed	destroyed
xView2-1st	xBD	-	PN	36.84	4.37	50.12	5.42			
			MH	38.18	23.41	34.91	4.82			
			Beirut					46.14	28.99	28.50
SE-ResNeXt-50	xBD	-	PN	35.96	0.18	**50.93**	5.32			
			MH	38.43	22.75	**35.46**	**5.91**			
			Beirut					53.33	35.34	43.00
SE-ResNeXt-50	xBD low res 90cm GSD	-	PN	35.55	2.93	50.49	1.28			
			MH	36.40	**27.63**	26.91	1.63			
			Beirut					56.33	36.39	50.24
SE-ResNeXt-50	xBD mix res 45+90cm GSD	-	PN	35.44	3.99	50.23	0.17			
			MH	36.88	25.45	32.41	0.02			
			Beirut					**58.69**	**37.89**	**56.62**
SE-ResNeXt-50	xBD 3 classes 4 epochs	✓	PN	**37.41**	**6.55**	50.09	5.57			
			MH	**37.89**	25.78	32.94	3.21			
			Beirut					45.89	28.92	27.84
SE-ResNeXt-50	xBD 3 classes 8 epochs	✓	PN	36.51	0.81	49.73	**8.06**			
			MH	33.18	6.94	33.93	2.28			
			Beirut					51.67	32.5	32.96

Table 3. Quantitative evaluation of the building damage assessment results on the RoIs Praia Nova (PN) and Macarungo Hospital (MH) in Beira, Mozambique and on our test area in Beirut, Lebanon. The fine-tuning is performed on the Macuti Village in Beira.

Table 3 gives an overview of the results obtained from the different configurations for the two test areas in Beira and the test area in Beirut. First, we apply the trained network of the xView2 winning team without modifications on the three test regions. As training and testing the ensemble model is time-consuming, we choose the best sub-model (SE-ResNeXt-50) from the ensemble and investigate its performance. As the next experiment, we reduce the resolution of the original dataset by half to about 90cm GSD and train the same model on the down-sampled training data and a mixture of low and original resolutions. Here, we pad the down-sampled images with mirroring boundary regions in order to keep the size of the input images the same. The idea behind down-sampling is to match the extent of a house in the training data to that from the test data (as an average house in Beira is much smaller than an average house from the xBD dataset). In parallel, we fine-tune the networks on two RoIs of the Beira dataset, for which we extract patches with a size of 1024 × 1024px and an overlap of 40% from the the pre- and post-disaster images. Note that for this experiments training and test are performed using 3 damage classes only (the classes minor and majored damage of the xBD dataset are combined into the class ”damaged”). We fine-tune the damage classification network with a learning rate of 0.00001 and 0.00005 for 2, 4 and 8 epochs with scheduled learning rate de-



(a) pre-disaster satellite

(b) post-disaster satellite image

(c) building footprints (GT)

(d) predictions using xView2-1st

(e) predictions using SE-ResNeXt-50

(f) best predictions using fine-tuning

Figure 7. Building damage assessment for the RoI Praia Nova in Beira, Mozambique. In (d)-(f) the predictions of three different models are provided. Color coding: 🟩 no damage, 🟨 damaged and 🟥 destroyed.

cay. Note, that the IoU for the class ”background” is not listed in the Table 3.

The results in Table 3 show that the best results for the two test areas Praia Nova (PN) and Macarungo Hospita (MH) in Beira could be achieved by fine-tuning the SE-ResNeXt-50 network with a learning rate of 0.00005 over 4 epochs. Fine-tuning with 8 epochs do not help to improve the results further, which indicates overfitting of the model with a higher number of epochs (see the last row of Table 3). For the RoI Praia Nova, most of the models yield a good IoU of around 50% for the class ”damaged”, whereas all models fail in providing good prediction for the other two damage classes. For the area Macarungo Hospital, the classes ”no-damage” and ”damaged” obtain a IoU of around 27% and 35% at maximum. The low performance for the classes ”no-damage” and ”destroyed” could be explained by 1) the low number of building instances of these classes in our GT data and 2) by the large differences of the training data compared to out test set in terms of the size and shape of buildings and the appearance of their damage.

As the GT of our test area in Beirut contains two classes only, we merged the classes ”no-damage” and ”damaged” of our predictions into the class ”non-destroyed”. Overall, the model trained on the mixture of resolutions xBD dataset yields the best results for all classes in the Beirut test area. More precisely, the usage of low resolution data (90cm GSD) seems to help improving the performance for the damage classification, especially for the destroyed buildings class. On the other hand, fine-tuning the networks on the Beira datasets decreases the performance on the Beirut region. A possible reason for this could be the significantly different size and appearance of buildings between those areas.

A visual illustration of the damage classification results for the best models of Beira and Beirut are shown in Figure 7 and Figure 1 respectively. Regarding Figure 7 (d), (e), and (f), the algorithm is categorizing the majority of buildings correctly as damaged, but it under-performs on the classification of non-damaged and destroyed buildings inside the regions. The fine-tuning on Figure 7 (f) shows improvement on the undamaged and destroyed building classification, but still similar performance on the damaged building category indicating the low effect of data from the Beira region, which we assume could be due to the low amount of training data. By comparing the predictions of Beirut in Figure 1 (c) to the GT in (d), it can be seen that some of the larger buildings are missing in the building mask, but that the model is capable of classifying most of the destroyed buildings correctly. As mentioned, training the models with the mixture of resolutions yields less false positives on the damaged buildings while categorizing the majority of the true destroyed buildings correctly.

5. CONCLUSION AND FUTURE WORK

Building segmentation and damage level assessment using remote sensing data remains challenging. On the one hand, damaged buildings are often difficult to identify due to the limited geometric resolution of remote sensing data, especially satellite data. On the other hand, the assessment of the extent of the damage always depends on the observers and their experience. We therefore proposed to evaluate models trained on the xView2 challenge data on imagery from two other disasters for which we created a reliable reference building damage annotation following the same damage levels definitions, but reducing the number of classes from 4 to 3. Our goal is to assess the generalization capability of such models and annotation guidelines in the context of post-disaster relief missions, so this study benefited considerably from the extensive experience of the WFP with regard to damage classification. We reported an important difficulty for annotators to agree on damage levels, not only because of the low-level of detail in satellite imagery, but also due to the specificity of regional building features, which are not encompassed in the current standard damage definitions. The latter issue is also affecting the performance of models trained on the xBD data set, which need to be fine-tuned on the target event imagery to perform satisfyingly. In the future, the focus should be put on creating a consistent and reliable reference data set in order to improve the training of the damage assessment networks and to increase their generalization capability. In this regard, drone imagery could provide a quicker and more detailed view of the damage extent in future studies, at is would allow to identify key damage features on buildings that would not be distinguishable on satellite imagery.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Mozambique's National Institute for Disaster Management (INGC) and thank them for sharing UAV data with us. These data form the basis for our ground truth and allowed a quantitative evaluation of our models on the city of Beria, Mozambique.

REFERENCES

Adriano, B., Xia, J., Yokoya, N., Miura, H., Matsuoka, M., Koshimura, S., 2020. Damage characterization in urban environments from multitemporal remote sensing datasets built from previous events. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 3751–3754.

Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J., 2017. Dual path networks. *arXiv preprint arXiv:1707.01629*.

DIU, 2019. DIU's xVIEW2 - Assessing Building Damage - Computer Vision for Building Damage Assessment - Automate damage assessment to accelerate recovery from natural disasters: `https://www.challenge.gov/challenge/diu-xview2-assessing-building-damage/`.

DLR, 2020. Dossiers - Aerospace technologies for humanitarian aid: `https://www.dlr.de/content/en/dossiers/2020/humanitarian-technologies.html`.

Durnov, V., 2019. 1st place solution for "xView2: Assess Building Damage" challenge: `https://github.com/DIUx-xView/xView2_first_place`.

Ghosh, S., Huyck, C. K., Greene, M., Gill, S. P., Bevington, J., Svekla, W., DesRoches, R., Eguchi, R. T., 2011. Crowdsourcing for Rapid Damage Assessment: The Global Earth Observation Catastrophe Assessment Network (GEO-CAN). *Earthquake Spectra*, 27, 179-198.

Grünthal, G. et al., 1998. European Macroseismic Scale 1998: EMS-98. *GFZ German Research Centre for Geosciences*.

Gupta, R. et al., 2019. Creating xbd: A dataset for assessing building damage from satellite imagery. *CVPR Workshops*.

Hamaguchi, R., Hikosaka, S., 2018. Building detection from satellite imagery using ensemble of size-specific detectors. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 223–2234.

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

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Iglovikov, V., Seferbekov, S., Buslaev, A., Shvets, A., 2018. Ternausnetv2: Fully convolutional network for instance segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 228–2284.

IWG-SEM, 2018. Emergency mapping guidelines - building damage assessment chapter: `https://un-spider.org/sites/default/files/IWG_SEM_Guidelines_Building%20Damage%20Assessment_v1.0.pdf`. Accessed: 09-04-2021.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B. A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152, 166-177.

Russakovsky, O. et al., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 211-252.

Scheffler, D. et al., 2017. AROSICS: An automated and robust open-source image co-registration software for multi-sensor satellite data. *Remote sensing*, 9(7), 676.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. *CVPR*.

Voigt, S. et al., 2016. Global trends in satellite-based emergency mapping. *Science*, 353, 247-252.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X. et al., 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Weber, E., Kané, H., 2020. Building disaster damage assessment in satellite imagery with multi-temporal fusion.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

ZKI, 2021. Center for Satellite Based Crisis Information (ZKI): `https://zki.dlr.de`.