Peer review

Fan, Cory. 2024. "SurfaceNet: Leveraging Aerial LIDAR Point Clouds for Post-Earthquake Building Damage Assessment." *Journal of High School Science* 8 (3): 480–500.

1.as you state "…In this particular case, the finetuning dataset (the Kumamoto Earthquake Dataset), and the dataset that the pretrained model used for self-supervised learning, were substantially different…." Would your algorithm have obtained increased accuracy over others if your datasset were similar to what the other algorithms used ? In other words, is it scientifically justifiable to claim improvement in accuracy etc. if your training/testing/validation dataset is different from your compartors? Please explain and discuss in the manuscript.

2.what would be your algorithm's accuracy…. etc. if you trained, validated and tested it using other databases (not the Kumamoto database). In other words, does the accuracy of your algorithm depend on what database it is trained on? If yes, your comparison with other models and results is fallacious. Please explain and discuss in the manuscript.

3.If your objective is saving lives, would you not need the probability of habitation in collapsed buildings ? i.e. if the buildings were residential or commercial, occupancy rates, low-high income housing, zoning and use, density, day of the week (weekdays will be less populated since children will be at school and parent(s) will be at job sites etc.). Merely a knowledge of the extent of collapse of a particular building will not maximize the chances of saving the most number of lives. This data will hence need to be integrated with your (or other) algorithms to provide the maximum possibility of saving the maximum number of lives; only LIDAR will not be sufficient; no matter how accurate the algorithm is in detecting or predicting collapse. Where do you propose to get and incorporate this information from? Please describe and explain in the manuscript.

4.You state that "…For classification and patch segmentation, we use a stratified split across each individual damage class (D0, D1… D6), as opposed to a stratified split only on damaged and undamaged classes. I am assuming this is due to the class imbalance of your dataset (only 10% approx. collapsed). Stratification basee on individual damage class should account from class imbalance. However if you have already accounted for this, then why do you need to "… train with inverse class frequency weighted cross entropy to address the class imbalance…." ? Explain and describe why you have used class balance imbalance remedial methods for data cross validation as well as to train the already stratified data.

5.Were the thresholds used for calculating the mean IOU the same for all the algorithms presented in Table 3 ? Please explain and describe in the manuscript.

6.Why are some columns left blank for some algorithms in table 3 ? Please discuss and describe in the manuscript.

7.You state that "…For practical implementation, we use K=8 for the KNN and M=3 for the number of hops…." Is your algoritm's sensitivity likely to depend more on K and M because it

incorporates an inductive bias that emphasizes intra-surface relationships ? Explain whether your model is likely to be more sensititve to changes in K and/or M because of this inductive bias.

8.You state that "….For classification and patch segmentation, we produce building patches by taking the latitude- longitude bounding box of each building footprint and apply a 1e-4 latitude-longitude padding to each building. For segmentation, we split the entire area into 100x100 meter square point clouds. " Explain the justification for your 1e-4 latitude-longitude padding and for the 100x100 meter^2 point clouds in the manuscript.

9.Does your method and/or algorithm require pre-event LIDAR data? For example, reference http://dx.doi.org/10.1117/1.JRS.11.046024 does not require pre-event LIDAR data. Please discuss in the manuscript.

10.The extent and pattern of damage, as well as the type of damage to buildings are a function of the magnitude of the quake on the Richter scale. Will your training data still be valid for different magnitude earthquakes ? Will a model trained on a Richter 7 earthquake still be valid in an environment of a Richter 5 earthquake? Explain and describe in the manuscript how the accuracy and mean IOU may be expected to change with different magnitude earchquakes.

11.Describe the methodology used to collect LIDAR data (the data that was used in this manuscript). For example, the reference cited in point 9 provides the following information "…..The lidar data used for development and testing were collected on January 21, 2010 by Kucera International Inc. and the Rochester Institute of Technology (RIT). The point clouds for the seven sites have an average point density of 4.2 pts/m2 and were captured by a Leica ALS60 at an altitude of ∼820 m with a pulse rate of 150 kHz. The vertical point measurement accuracy of the instrument is 0.15 m. Multispectral imagery simultaneously was collected on the aircraft by the Wildfire Airborne Sensing Platform (WASP) system at a resolution of 0.15 m……."
12.The reference quoted in point 9 mentions "……These sites were chosen because they contain both a wide range of building types and building damage types. Different construction types include: one- to three-story reinforced concrete buildings, masonry bearing walls, timber frames, and shanty housing made of reinforced concrete and masonry block with corrugated metal roofs.19 The damage level in buildings range from completely undamaged to fully destroyed, and everything in between……" This leads to an algorithm that is adequately generalized. In contrast, your dataset includes buildings from a first-world country that are already reinforeced due to building codes in an earthquake prone region. Explain what this homoegeneity in construction in your dataset implies for your results and accuracy of your algorithm. Describe and explain in the manuscript.
http://dx.doi.org/10.1117/1.JRS.11.046024 Building damage assessment using airborne lidar
http://dx.doi.org/10.1111/1755-6724.12781 Building Damage Extraction from Post-earthquake Airborne LiDAR Data
https://isprs-archives.copernicus.org/articles/XL-1-W5/595/2015/isprsarchives-XL-1-W5-595-2015.pdf BUILDING DAMAGE ASSESSMENT AFTER EARTHQUAKE USING POST-EVENT LiDAR DATA

Comment 1:

We included clarification that we trained all models with the same set of datasets in section 2.4 “Training Methodology”. We state that:

“All fully-supervised models are trained on the KumamotoEQ dataset and the HaitiEQ dataset for their respective tasks. For pretrained models, we use open-source checkpoints and then finetune on the KumamotoEQ and HaitiEQ datasets.”

Comment 2:

We performed a new set of experiments on a dataset created from data from the 2010 Haiti earthquake, which is formatted for the general segmentation task. SurfaceNet exceeds the next-best model by 0.6 mIOU. We edit section 2.1 “Dataset” to detail the new Haiti dataset and edit section 3 “Results” and section 4 “Discussion” to discuss the results on the new of experiments. The changes made are very extensive; hence, the authors do not list them here.

Comment 3:

We include further discussion on the implications and limitations of rapid building damage assessment on post-earthquake rescue efforts in section 1 “Introduction” and section 4 “Discussion”.

In “Introduction”, we make a variety of changes, but principally include the following paragraph:

“Building damage assessment can play a substantial role in post-earthquake rescue. In the guidelines published by INSARAG (the International Search and Rescue Advisory Group), they outline five major steps to a recovery operation: “Wide Area Assessment”, “Worksite Triage Assessment”, “Rapid Search and Rescue”, “Full Search and Rescue”, and “Total Coverage Search and Recovery”. Building damage assessment, when possible, is involved in parts of the “Wide Area Assessment” step, which involves assessing the overall damage of sectors and parts of the “Worksite Triage Assessment”, which involves determining and assessing specific rescue sites within sectors (2). Thus, building damage assessment is crucial for efficient allocation of rescue teams (3).”

In “Discussion”, we include:

“Finally, this paper only explores one aspect of the search and rescue process; we do not propose a comprehensive search and rescue methodology but a technological tool that can help improve part of the process. As stated in section 1, “Introduction”, building damage assessment can aid in parts of two of five steps of the rescue process; however, given that building damage assessment or predictions, such as the ShakeMap software, are already part of the post-earthquake rescue process (28), more effective and efficient methods of building damage assessment can be simply integrated into the existing workflow. Consequently, rescuers may rely upon both damage assessments produced by SurfaceNet as well as other existing data present in ShakeMap (such as ground movement and hospital/highway locations) and other such workflows.”

Comment 4:

We include further discussion of the necessity of inverse class frequency weighted cross entropy in section 2.4 “Training Methodology”. We state that:

“Although stratified split ensures equivalent class-distributions between training, validation, and testing, the usage of standard, non-weighted cross entropy, in the authors’ experiments, caused

the model to converge to only predict the most represented class; consequently, class frequency weighted cross entropy is necessary to improve training speed and overall trainability."

Comment 5:

We clarify the method of calculating Mean IOU in a new section, section 2.5 "Evaluation Metrics". We include the following paragraph:

"For calculating class accuracy, we consider it to predict a building as damaged if it assigns a greater than 50% probability for the damaged class, and undamaged otherwise. For calculating mean IOU, we consider a model, from the three classes (including both the two foreground and the background class), to have predicted a class if its corresponding probability is the highest amongst the three classes. This is most analogous to a threshold of 0.5 on a binary segmentation task."

Comment 6:

We include a brief extension to the clarification in the Table 3 description. We state:

"Blank spaces (denoted by "-") indicates that there is no open implementation (or checkpoint for pretrained models) for that model on that given task."

Comment 7:

We include some explanation on the effects of varying these parameters in section 2.2.2 "SurfaceConv". We state:

"Increasing M will increase the size of the receptive field; decreasing M will correspondingly decrease the size of the receptive field. We find M=3 to be a sweet spot. Increasing K will also increase the size of the receptive field but may cause the model to "jump" over gaps across surfaces. On the other hand, due to the natural stochasticity in the position of points in a point cloud, setting K to a value that is too small will result in the receptive field having a high degree of noise based off slight changes in the position of points. We find K=8 to be an approximate sweet spot."

We additionally investigate the empirical effects of varying the parameters in a new section, section 3.3 "Ablation Studies".

Comment 8:

We include further explanation of the padding and general segmentation patch sizes in section 2.1.2 "Data Preparation". We state:

"For classification and patch segmentation, we produce building patches by taking the latitude-longitude bounding box of each building footprint. Following Xiu et al., (7), we apply padding to the bounding boxes. We utilize 1e-4 padding for both longitude and latitude, which equates to around 11 meters of padding for latitude and 9 meters for longitude. This allows the model to consider the surroundings of the building, which may contain context such as rubble. For general segmentation, we split the entire area into 100x100 meter square point clouds, which is around 32,000 points per tile. This number has been chosen for being the maximum round number which will fit into the memory of an RTX 4090 (not even a single batch of 200x200 on DS-Net will fit into the memory of an RTX 4090)."

Comment 9:

We specify that our method does not need pre-earthquake LIDAR data in section 1 “Introduction” and section 2.1.1 “Data Study Area and Description”.

In section 1 “Introduction”, we state:

“Some methods (9) utilize pre- and post-earthquake LIDAR data; however, since collecting LIDAR is an active process, we choose to focus on methods that only require post-earthquake LIDAR data.”

In section 2.1.1 “KumamotoEQ Data Study Area and Description” and 2.1.2 “HaitiEQ Data Study Area and Description”, we specify again that we are using post-earthquake LIDAR data only.

Comment 10:

As mentioned in Comment 2, we introduced a new set of experiments performing general segmentation from data from the 2010 Haiti Earthquake. We discuss the generalizability of our method in section 4.1 “Results Analysis” and the limitations in section 4.2 “Limitations and Future Work”.

Comment 11:

We include the appropriate edits to 2.1 “Dataset”.

Comment 12:

Our edits with respect to this comment are similar to those with respect to comment 2 and 10. We include a new set of experiments on the 2010 Haiti Earthquake. Generalizability to other types of buildings is discussed in 4.1 “Results Analysis” and limitations are discussed in 4.2 “Limitations and Future Work”.

Thank you for addressing my comments. Accepted.