Automated Detection of Humanitarian Structures in Refugee Camps

Emma Marriott
Stanford University
Computer Science
emarriott@cs.stanford.edu

Robert Young
Stanford University
Engineering
rsyoung@stanford.edu

Daniel Sambor
Stanford University
Civil and Environmental Engineering
dsambor@stanford.edu

Abstract

Determining the number and location of structures in rapidly expanding refugee camps is a critical challenge in providing aid for millions of displaced people worldwide. In this analysis, a hybrid classification and regression model and object detection model were implemented on high-resolution satellite imagery to detect and count structures in refugee camps. The models were trained on crowdsourced data and evaluated against ground truth structure counts and locations. The object detection model out-performed the input crowdsourced dataset in localizing structures across all regions, while the hybrid model achieved superior regression results in one of the African regions. This analysis is one of the first to span over 120 refugee camps in Africa and Southwest Asia, and demonstrates the value of deep learning for humanitarian aid.

1. Introduction

There are currently 68.5 million forcibly displaced people worldwide, and 25.4 million of these people are refugees [6]. These individuals settle in refugee camps, which are often developed as temporary shelter locations but on average have lifetimes over 17 years [6]. A majority of refugees come from South Sudan, Afghanistan, and Syria, and many camps are concentrated in sub-Saharan Africa and Southwest Asia. Living conditions in these camps are quite poor, including overtaxed infrastructure, poor sanitation, malnutrition, and safety concerns [6].

In order to provide aid for refugees, relief organizations such as the United Nations High Commission for Refugees (UNHCR) need to estimate the total amount of people in a camp requiring aid as well as routes to navigate within camps. Current methods to estimate the number of people require in-situ assessment, or sending individuals to inspect thousands of structures in a camp by hand [18]. These manual site surveys are costly, time-consuming, and hazardous [14]. Moreover, camps change rapidly in size and distribution, making in-situ assessments quickly outdated [3].

Satellite imagery can offer quicker and more accurate information for documenting the overall size and geographic layout of refugee camps. Increasing numbers of satellites now gather earth-observation data, and these remotely-sensed data have increasing spatial and temporal resolution. With these improved data sources, identifying humanitarian structures at sub-meter detail is now possible [1].

Relief organizations have gradually begun to employ a variety of techniques to estimate camp size using visual inspection of satellite imagery, though many still rely on time-intensive manual counting of remotely-sensed structures. Organizations need to know the total number of people in each camp as well as their locations in order to determine the aid required and provide optimal routes through each camp for delivering this aid. Coupling machine learning techniques with high-resolution satellite imagery provides an alternative to manual counting in order to automate structure detection in refugee camps [12].

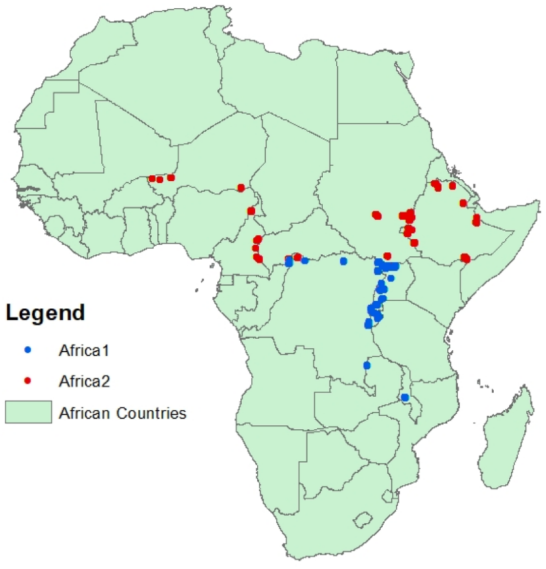

Figure 1. Refugee camps studied in Africa.

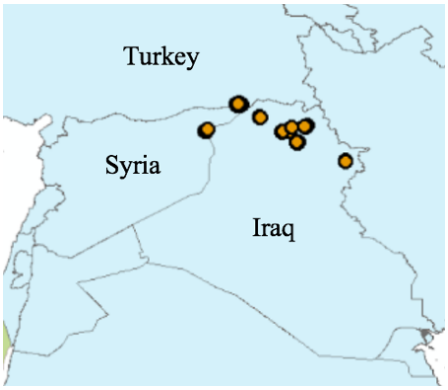

Figure 2. Refugee camps studied in Southwest Asia.

2. Background and Related Work

Given the advancement of deep learning algorithms and high-quality satellite imagery, there has been a subsequent rise in studies using these methods for deployment in humanitarian applications, including disaster response and refugee camp mapping. As remote sensing techniques have advanced, researchers have applied several methodologies to count and detect specific structures in refugee camps [17].

As an early work in this field, Bjorgo mapped the total area and extent of refugee camps given limited spatial resolution [2]. Giada investigated camps in more detail by implementing supervised and unsupervised classification, segmentation, and an object-oriented mathematical morphology approach [9]. The object-oriented morphology approach performed best, with a precision of 0.89 and recall of 0.87 in the Lukole refugee camp in Tanzania [9]. Upon correctly identifying structures, the total population of the camp could be estimated. Laneve expanded upon this study with regard to the Lukole camp and achieved a recall of 0.92 using a similar morphological approach [13]. These methods relied on utilizing a structuring element of assumed size and shape to match existing structures. These approaches and others have depended on analysts to inspect satellite imagery to determine parameters applicable only to specific regions.

A two-stage segmentation approach utilizing graph-cut optimization was introduced by Kahraman in order to address the need for more general approaches [11]. The first stage of segmentation isolated the camp area from the surrounding background, and the second stage separated structures from one another. For several camps in North Africa and Southwest Asia, the method achieved precision of 0.93 and recall of 0.84. This method assumed that structures had a higher spectral intensity than their surroundings and relied on filtering to provide necessary separation between the two categories, often performing better on grayscale imagery. Given that structures are typically constructed from local materials and thus have spectral characteristics similar to the background imagery, this segmentation approach is not strongly generalizable among different regions.

Deep learning methods to detect humanitarian structures can be grouped into several categories: segmentation, classification, and object detection [3]. In the context of remotely-sensed data of humanitarian structures, segmentation defines the pixels in the image associated with a structure versus the background topography. Classification determines whether or not a structure is located in a specific area, or what type of structure is located in the area. Object detection meanwhile identifies each structure in an image with an associated bounding box and classifies the structure within each bounding box. Within object detection, region-proposal convolutional neural network architectures, such as Faster-RCNN, have shown promise in humanitarian structure detection [15].

Recently transfer learning has been utilized to apply architectures previously trained on other structure datasets to the humanitarian detection problem [10]. These approaches allow for more generalized results across numerous areas. Quinn implemented a hybrid segmentation and object detection approach to document structure footprint areas and count the total number of structures in each camp [15]. The researchers used a Mask-RCNN model with a ResNet101 network pre-trained on ImageNet. The model was tested on high-resolution satellite imagery for 13 refugee camps. For each camp, precision ranged from 0.88 to 0.96 and recall from 0.75 to 0.89. Across all camps, the model achieved an average precision of 0.78, though the two best camp results yielded average precision values of 0.92. The approach also utilized data augmentation to rotate structure orientation in the training set, which improved average precision by approximately 3%. The authors concluded that much more work was needed before analysts' visual inspection of satellite imagery could be quickly automated with machine learning algorithms, given that current studies are not generalizable across regions [15].

The data used by Quinn included expert analysis and identification of structures, which was quality-controlled by a second expert analyst. The authors manually outlined 87,137 structures in order to train a segmentation mask for bounding box detection [15]. However, such large and well-labeled datasets are only rarely available, limiting the applicability of this work for international aid organizations.

Challenges remain to be solved in developing methods that can generalize well across different camps in various regions depending on topography, vegetation, and structure materials. Identifying discrete structures in highly clustered camps also remains an ongoing challenge for most current approaches. Integrating automated deep learning techniques into an organization's workflow is additionally

difficult in itself [15]. This work aimed to contribute to the prior body of work by addressing these challenges. In particular, we focused on creating models applicable to camps across many diverse regions using non-expert labeled input data.

3. Methods

3.1. Problem Statement

In response to key challenges in the literature, this analysis focused on addressing two main research areas: 1) determining the total number of structures in camps across multiple regions and 2) localizing structures with high precision and recall. While classifying identified structure types was an initial goal of this work, limitations in the dataset prevented deep exploration of this goal.

3.2. Data

This analysis focused on several regions characterizing a majority of current settlement areas by displaced people, including Africa and Southwest Asia (Middle East). Africa was further categorized into the equatorial region (Africa1) and the Sahel region (Africa2). The dataset spanned 125 refugee camps across the three regions. Of these camps, 112 were located in Africa, which had the highest geographic variation among camps, as shown in Figure 1 above. The remaining 13 camps were located in Southwest Asia, clustered in northern Iraq, eastern Syria, and Southeastern Turkey. These camps are shown in Figure 2.

For each of the areas surrounding these camps, we obtained high-resolution satellite imagery from DigitalGlobe [7]. The imagery consisted of panchromatic band data and was captured for one moment in time by satellites GeoEye1 and WorldView (1-3) at a spatial resolution of 40-50cm. Imagery was primarily taken between January 2017 and June 2018, though some images were acquired as far back as 2014.

Labels for the type and location of each structure were obtained through Tomnod, a crowdsourced platform, in collaboration with the Digital Globe Foundation and the UNHCR [8]. Tomnod volunteers were shown DigitalGlobe satellite imagery and asked to identify the location and types of structures in the image. Volunteers produced a raw label dataset of approximately 690,000 labels. The raw dataset was further processed to aggregate labels between volunteers, resulting in about 180,000 quality-controlled labels.

Labeled structure types consisted of UNHCR tents, round earthen structures, administrative structures, and other tents or permanent shelters. UNHCR tents and other tents or permanent structures were common in all three regions, while administrative structures accounted for less than 5% of total structures in each region. Round earthen structures were located only in the two African regions.

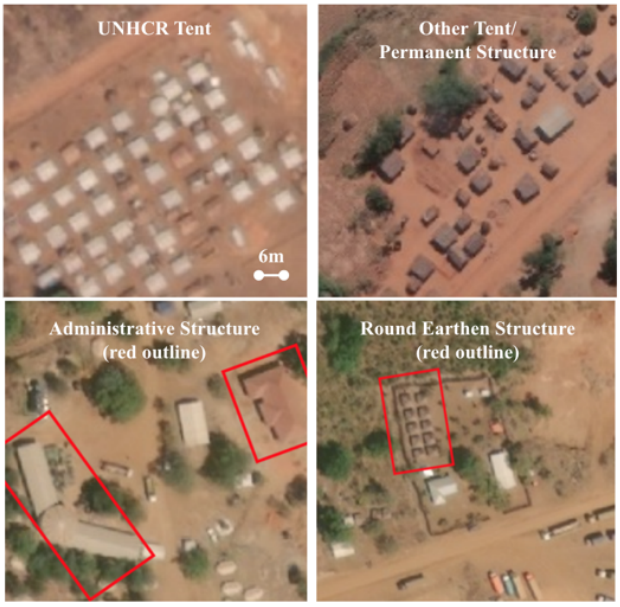

Figure 3. Types of structure labels

Detailed distributions of the structures for each region are shown in Table 1.

Example imagery of each structure type is shown in Figure 3 for reference and scale. UNHCR tents were usually white and square or pod-shaped, with a footprint of about 23 square meters, whereas other tent structures varied considerably in shape and size. Round earthen structures were typically the smallest type of structure and administrative structures were the largest. This analysis sought to count and localize structures of all types, regardless of variations in size or appearance.

3.3. Data Limitations

The use of crowdsourced data for this analysis brought substantial limitations in data format and quality. The original DigitalGlobe imagery viewed by Tomnod labelers was not included in the dataset, and was acquired separately to match the temporal and geospatial locations of the original labeled images. It is possible that some acquired imagery may not have temporally matched the images shown to Tomnod labelers, resulting in some labeling error.

Structure Type	Africa1	Africa2	SW Asia
UNHCR Tent	54%	35%	62%
Round Earthen	15%	40%	0%
Administrative	2%	3%	4%
Other Tent	29%	23%	35%
Total	95219	60670	17785

Table 1. Number of structures by type and region.

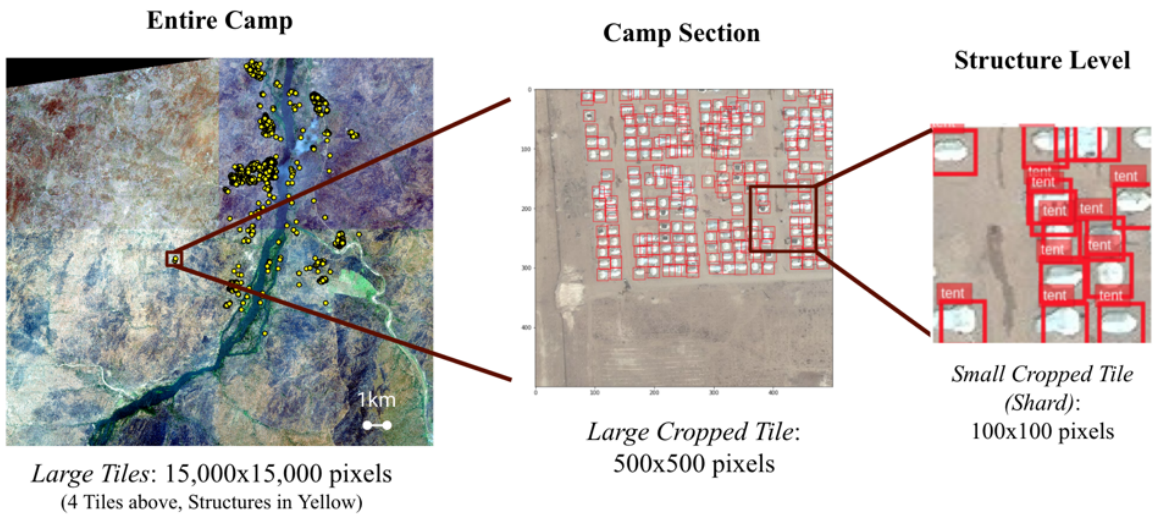

Figure 4. Data processing flow, from a large-scale camp tile to a 100 × 100 pixel crop with labels.

Throughout all three regions, many structures were incorrectly labeled or left unlabeled by Tomnod volunteers. The small size of the Southwest Asia dataset allowed for inspection of all tiles, showing that 18% of the dataset contained substantially mislabeled tiles (for an example of mislabeling, see Figure 8). Time limitations made removal of these areas across all the datasets infeasible. Additionally, no ground truth labels were available beyond the crowdsourced data from Tomnod. As a result, evaluating model performance as compared to the crowdsourced data formed a major component of this analysis.

The Tomnod platform used point-based labels to identify structures. However, individual points were unable to convey features of each structure such as the variation in size and shape between large administrative structures, dense buildings, and sparse tents. This lack of bounding boxes or area information added an additional layer of complexity to the modeling task.

While crowdsourced labels had four structure types, as described earlier, Tomnod labelers may not have interpreted these labels in the same manner. In particular, "other tents and permanent shelters" were classified as a single structure type, making the separation of non-humanitarian structures and humanitarian structures in dense urban areas unclear. Given this limitation, we chose not to attempt to classify structures based on type, and to label all structures in refugee camp areas as humanitarian structures.

3.4. Modeling Framework

In order to apply the deep learning models to the crowdsourced data, structure labels were converted from points to bounding boxes. Bounding boxes were defined as the 20 × 20 pixel or 30 × 30 pixel square centered around the crowdsourced label, depending on the model. These bounding box sizes were chosen through visual inspection of the data, but did not account for large variations in structure size, particularly for administrative structures.

We posed the goal of counting the total number of structures in an area as a regression problem, evaluated using R^2 and RMSE. We measured performance on structure localization using recall, precision, and F1 score [16].

Attempting to identify all structures in each image, even if some were non-humanitarian, could lead to high false positive rates, and thus reduced model precision. However, relief organizations indicated a preference for high recall rather than high precision, as overestimation of structures ensures that total aid is sufficient and that routes within each camp are clear of unidentified structures.

The modeling framework involved training and validating individual models for each region in order to minimize the impacts of inter-region variation. Within a region, 80% of refugee camps were selected for training and 10% were held for each of validation and testing sets. A single refugee camp consisted of several large DigitalGlobe imagery tiles, measuring approximately 15000×15000 pixels. For ease of working within a specific part of a refugee camp, these large tiles were cropped into 500×500 pixel square large cropped tiles. To apply the models at the structure level, the data was segmented further into 100×100 pixel small cropped tiles. Example data highlighting this process for a single refugee camp is shown in Figure 4.

3.5. Hybrid Classification and Regression

A hybrid classification and regression model was used to predict the total number of structures in a given 100 × 100 pixel cropped area while also giving insight into the loca-

tion of these structures within the area. These objectives were performed in tandem in order to take advantage of knowledge of structure locations in the crowdsourced data while focusing on the goal of counting the total number of structures. The hybrid model was developed and trained from scratch based on early success in binary classification of structure presence.

Separate hybrid models were developed for the Southwest Asia, Africa1, and Africa2 datasets. The inputs to each model were 100 $\times$ 100 pixel cropped areas. The outputs were (1) a binary classification for each 20 $\times$ 20 pixel shard of the cropped area representing whether any portion of a structure was present in this shard, and (2) a predicted total number of structures in the cropped area. The predicted shard classifications and total structure count were compared with crowdsourced labels. Each shard was classified as containing a structure if one or more of the four corners of a structure 20 $\times$ 20 pixel bounding box were within the shard.

The hybrid model consisted of a convolutional neural network architecture, shown in detail in Appendix A1. The input image (100 $\times$ 100 pixels) was first converted into 25 shards (20 $\times$ 20 pixels), then these shards were passed through a series of four ReLU-activated 2D 3 $\times$ 3 convolution and 2 $\times$ 2 maxpool layers that acted on each shard independently. Following these layers, a dropout layer, a dense layer, and a second dropout layer were shared for both classification and regression. Classification for each shard was performed using a final dense layer with sigmoid activation. Regression for the entire area was executed using a separate final dense layer with ReLU activation.

The model loss function consisted of a weighted sum of binary cross-entropy loss on the shard classifications and mean squared error on the regression with weights 1 and 0.1. The hybrid model was trained for 10 epochs for Southwest Asia and 20 epochs for Africa1 and Africa2.

3.6. Object Detection

Prior studies have shown positive results in counting and locating humanitarian structures through object detection methods [15]. The task of this model, improving upon the spatial resolution of the hybrid model, was to not only identify the total number of structures, but also detect their individual locations through bounding boxes. Traditionally, object detection methods include two stages - region proposal and classification of proposed regions - with loss predicated on intersection over union (IoU). IoU is calculated as the ratio of the intersection of the actual bounding box and proposed boxing area with the corresponding union. We used 30 $\times$ 30 pixel bounding boxes around each label for compatibility with standard object detection learning problems.

An advantage of object detection learning is the availability of complex models trained and evaluated on datasets

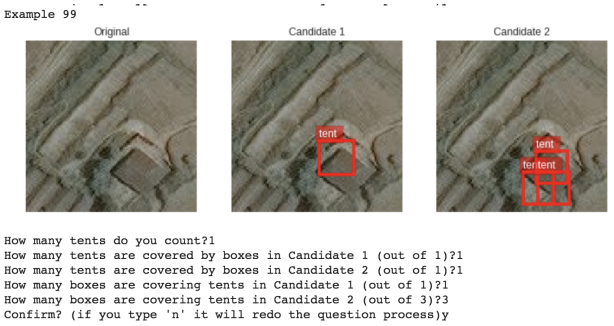

Figure 5. Example of the display and user input questions for ground truth evaluation in object detection.

orders of magnitude larger than the data available here. One such dataset is the Pascal Visual Object Classes (VOC) challenge, which contains over 27,000 ROI labels on 11,000 images [5]. For our model, we used a Single-Shot Detector (SSD) Resnet50 network, pre-trained on VOC from GluonCV [4]. This model was fine-tuned on 100 $\times$ 100 pixel labeled images from our humanitarian structure dataset. The model was trained for 2 epochs (about 80,000 examples total) for each of the Africa1 and Africa2 datasets and 4 epochs for the Southwest Asia dataset (about 5,000 examples total). For evaluation and inference, we used a score threshold of 0.5 to obtain the predicted bounding boxes.

3.7. Model Evaluation

Due to the poor quality of the crowdsourced Tomnod data, as outlined previously, a manual ground truth evaluation of label quality was performed for both the crowdsourced labels as well as the model predictions. For this evaluation, we randomly selected 300 examples from the test set for each of the three regions. For each example, we displayed the original image, the crowdsourced labels, and the predicted labels, with the order of the latter two images randomly flipped and unlabelled in order to ensure blindness. An example of the display and the questions asked are shown in Figure 5.

The questions elicited (1) the total ground truth count; (2) the number of correctly identified tents (true positives) by proxy of number of tents partially contained by boxes for each candidate; and (3) the number of correct predictions (total predicted minus false positives) by proxy of the number of boxes that partially covered tents for each candidate. From these measures, we derived ground truth precision, recall, and F1 scores for the object detection models and crowdsourced data.

Region	Labels	Precision	Recall	F1	RMSE	R^2
Africa1	Crowdsourced	**1.00**	0.448	0.618	1.32	0.551
	Detection Model	0.826	**0.858**	**0.842**	1.10	**0.769**
	Hybrid Model	-	-	-	**1.02**	0.668
Africa2	Crowdsourced	0.921	0.114	0.202	4.32	0.102
	Detection Model	**0.964**	**0.369**	**0.533**	**3.59**	**0.507**
	Hybrid Model	-	-	-	4.55	0.290
Southwest Asia	Crowdsourced	**0.994**	0.457	0.626	4.17	0.487
	Detection Model	0.966	**0.910**	**0.937**	**2.96**	**0.878**
	Hybrid Model	-	-	-	5.20	0.198

Table 2. Ground truth model evaluation results, comparing the hybrid classification and regression model and the object detection model with crowdsourced performance on the ground truth data. The best values for each metric and region are marked in bold. The detection model had the highest recall, F1 score, and R^2 for all regions. The hybrid model had the lowest RMSE for Africa1, while the detection model had the lowest RMSE for Africa2 and Southwest Asia. Ground truth precision, recall, and F1 were not calculated for the hybrid model.

4. Experiments

4.1. Hybrid Classification and Regression

Figure 6 shows an example of the hybrid model classification map and regression value compared to the crowdsourced labels and structure count for a small cropped tile in the Africa1 test dataset. The hybrid model accurately labeled the upper right structure despite no label in the crowdsourced dataset. Additionally, the hybrid model predicted a total of 6.1 structures, closer to the ground truth structure count than the crowdsourced data.

The classification and regression results of the hybrid model are summarized in Table 3. The classification performance of the hybrid model was relatively poor when using the Tomnod crowdsourced labels as the point of comparison. The model showed high levels of overfitting to the train set and significant reductions in precision, recall and F1 between the train and test sets. Overfitting was likely caused by variation in structure layouts and appearances between camps in each region. In particular, the overfitting in Southwest Asia was most severe due to the small number of camps in the region, leading to more significant variation between the train, validation, and test sets. However, the poor classification performance of the hybrid model could have been partially caused by poor crowdsourced dataset quality rather than incorrect model predictions.

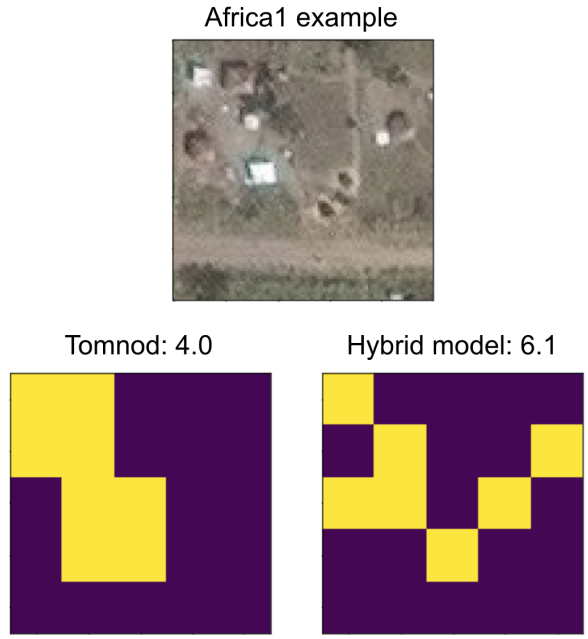

Figure 6. Examples of the predicted count and structure classification map output from the hybrid model, compared with count and structure map from Tomnod crowdsourced data.

Region	Split	Precision	Recall	F1
Africa1	Train	0.623	0.235	0.342
	Test	0.596	0.111	0.187
Africa2	Train	0.653	0.064	0.116
	Test	0.324	0.001	0.002
SW Asia	Train	0.687	0.562	0.618
	Test	0.107	0.085	0.095

Table 3. Hybrid model classification results, using Tomnod labels as comparison. Large gaps between train and test set metrics indicated model overfitting.

The ground truth evaluation methodology permitted a comparison of the hybrid model regression predictions with ground truth structure counts rather than only crowdsourced data. These results are shown in Table 2. While regression results were poor for the Southwest Asia and Africa2 datasets, the hybrid model R^2 and RMSE values improved

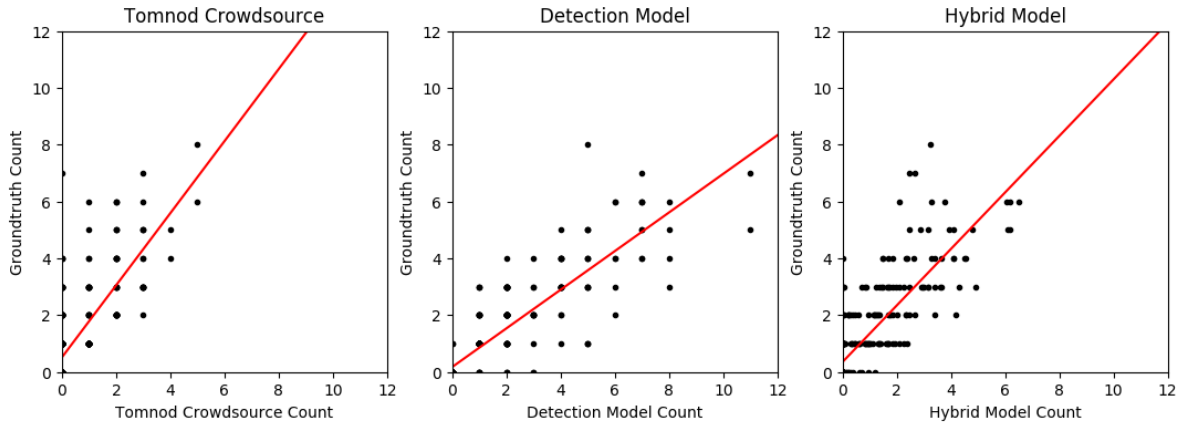

Figure 7. Regression between the number of structures predicted by each set of labels and the ground truth number of structures for Africa1. The hybrid model had the lowest RMSE (1.02) on the Africa1 dataset, followed by the detection model (1.10) and the crowdsourced data (1.32).

Figure 8. Examples of detection model results for training set (top row) and validation set (bottom row) in SW Asia. Crowdsourced labels (left column) and model predictions (right column) are shown with bounding boxes.

upon those of the Tomnod crowdsourced labels for Africa1, as shown in Figure7. The strong regression performance and poor classification performance for Africa1 could be the result of similar levels of false negatives and false positives within each image. These two factors may have balanced each other to achieve relatively accurate total counts. The difference in performance could also stem from the use of separate final model layers for classification and regression, such that the final regression layer was better-trained than the final classification layer.

4.2. Object Detection

The object detection models that were fine-tuned from SSD Resnet VOC were able to accurately place many true labels while disregarding faulty labels in the crowdsourced dataset. Examples of good generalization from the model on faulty labels are shown in Figure 8. In this example, the detection model accurately predicted all the structures, as shown in the right column, compared to missing labels in the crowdsourced data.

The results of the ground truth evaluation for the crowdsourced data and hybrid and detection models for each region are shown in Table 2. Despite slightly lower precision in Africa1 and Southwest Asia, the detection model achieved improvements in recall of more than 90% in all three regions compared to crowdsource performance. This increase in recall led to an improvement of between 36% and 164% in F1 score in the regions. The detection model achieved recall values above 0.85 in Africa1 and Southwest Asia. The lower recall value of 0.37 in Africa2 may have been obtained due to inferior training capabilities from the relatively poor labeling of Tomnod volunteers in that region, as shown by the lower crowdsourced recall and precision compared to Africa1 and Southwest Asia.

These strong localization results led to improved structure counts as well. The RMSE and R^2 for the detection model improved upon those of the crowdsourced structure counts across all three regions. Improvement over the crowdsource data was particularly large for the Southwest Asia region, where the detection model achieved 29% lower RMSE compared to the crowdsourced data. The model's ability to out-perform the crowdsourced data on structure count regression in a blind ground truth evaluation is fur-

ther evidence of strong generalization across camps. This generalization was afforded through transfer learning on a pre-trained object detection model.

5. Conclusions

Performance of the two models and the crowdsourced data varied substantially across regions, as evaluated using the ground truth results. The detection model achieved the best F1 classification performance on Southwest Asia, followed up by Africa1 and Africa2. In all cases it outperformed the hybrid model and even the crowdsourced dataset. The hybrid model also achieved its best results on the training data in Southwest Asia, perhaps due to better crowdsourced data quality and more structured refugee camps in that region. While the hybrid model achieved the best RMSE results for the Africa1 dataset, the detection model showed superior generalizability across regions and out-performed the crowdsourced data in all cases. Overall, this variation in performance between regions was expected given similar variation in prior assessments in the literature.

Detection model precision and recall results for the Africa1 and Southwest Asia regions were similar to those presented in Quinn [15]. However, the crowdsourced dataset used here had substantially lower initial recall values than did the expert-labeled dataset used in Quinn. Regardless, the object detection model's ability to perform relatively well across three regions addresses prior challenges in the literature of generalizability across regions.

The results presented here are thus best compared against the crowdsourced data itself, rather than past deep learning studies. While the detection model performance exceeded that of the crowdsourced data, this model still does not have sufficient precision and recall to replace expert labeling. Rather, the models developed in this analysis could be considered as a methodology to augment crowdsourced data, enhancing label recall while maintaining or improving label precision. In this way, deep learning tools may be more useful in tandem with existing data collection methods in relief organizations, rather than as a replacement for these methods.

6. Future Work

As a nascent field, the application of deep learning for humanitarian relief shows great promise. Work in this field is limited by the availability of well-labeled data, as highlighted in this study. Future work using crowdsourced data should consider the following enhancements to the crowdsourcing process for deep learning applications: (1) inclusion of the original images seen by labelers in the dataset for input to the deep learning models; (2) training of crowdsource volunteers and effective quality control to ensure high structure recall and precision; (3) labeling with bounding boxes or pixel segmentation to permit better area localization and more accurate object detection models; and (4) consistent and meaningful structure type categories to allow deeper research into refugee structure classification. Ultimately, future work should include the development of a more effective crowdsourced data collection methodology, along with strong ground truthing to ensure data validity.

Future work in model development could pursue paths in several areas of study. Continued development of more complex pre-trained object detection models could continue to improve performance without additional fine-tuning. Model performance could also be enhanced with a positive-unlabelled or other bootstrap learning algorithm to increase the number of labeled structures seen by the model. Alternatively, models could be made more domain-specific through training on specific camp types rather than regional datasets. For example, separate models could be trained on dense urban camps, dense tent-based camps, and sparse camps. Models could also be ensembled to enhance accuracy for all regions regardless of camp type.

While general localization and regression models may provide significant value for humanitarian relief organizations, another pathway for future work might include the development of application-specific models. For example, a model could be trained using only data from a single camp and then used to count structures in that camp at a different time step. This application would eliminate the substantial impact of cross-camp testing, which proved highly challenging. Models could also focus on specific structure types, or classifying structures in order to better understand camp development and layout.

The methodologies presented here could provide valuable options for the counting and localization of humanitarian structures in refugee camps worldwide. Beyond these initial steps, a great deal of work remains to be done in applying deep learning techniques in the field of humanitarian relief. As better-validated data and continuously improved models are developed, deep learning will undoubtedly become a useful tool for monitoring refugee camps and providing aid to those who need it most.

References

[1] P. Aravena Pelizari, K. Spröhnle, C. Geiß, E. Schoepfer, S. Plank, and H. Taubenböck. Multi-sensor feature fusion for very high spatial resolution built-up area extraction in temporary settlements. *Remote Sensing of Environment*, 209(December 2017):793–807, 2018.

[2] E. Bjorgo. Using very high spatial resolution multispectral satellite sensor imagery to monitor refugee camps. *International Journal of Remote Sensing*, 21(3):611–618, 2000.

[3] F. Checchi, B. T. Stewart, J. J. Palmer, and C. Grundy. Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations. *International Journal of Health Geographics*, 12, 2013.

[4] dmlc. Gluoncv, 2018. http://gluon-cv.mxnet.io.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[6] U. N. H. C. f. R. Filipo Grandi. Un global appeal: Final report of financial, statistical, and operational refugee camp information.

[7] D. G. Foundation. Earthwatch: Basemap satellite imageryn.

[8] D. G. Foundation and UNHCR. Tomnod: Crowdsource platform.

[9] S. Giada, T. De Groeve, D. Ehrlich, and P. Soille. Information extraction from very high resolution satellite imagery over Lukole refugee camp, Tanzania. *International Journal of Remote Sensing*, 24(22):4251–4266, 2003.

[10] F. Hu, G. Xia, J. Hu, and L. Zhang. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *International Journal of Remote Sensing*, (7):14680–14707, 2015.

[11] F. Kahraman, H. F. Ates, and S. S. Kucur Ergunay. Automated Detection Of Refugee / IDP TENTS FROM Satellite Imagery Using Two- Level Graph Cut Segmentation. (September 2015), 2013.

[12] T. Kemper, M. Pesaresi, P. Soille, and M. Jenerowicz. Enumeration of Dwellings in Darfur Camps From GeoEye-1 Satellite Images Using Mathematical Morphology. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(1):8–15, 2011.

[13] G. Laneve, G. Santilli, and I. Lingenfelder. Development of Automatic Techniques for Refugee Camps Monitoring using Very High Spatial Resolution (VHSR) Satellite Imagery. *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on*, pages 841–845, 2006.

[14] S. Lang, D. Tiede, D. Hölbling, P. Füreder, and P. Zeil. Earth observation (EO)-based ex post assessment of internally displaced person (IDP) camp evolution and population dynamics in Zam Zam, Darfur. *International Journal of Remote Sensing*, 31(21):5709–5731, 2010.

[15] J. A. Quinn, M. M. Nyhan, C. Navarro, D. Coluccia, L. Bromley, and M. Luengo-oroz. Humanitarian applications of machine learning with remote-sensing data : review and case study in refugee settlement mapping. 2018.

[16] K. Sprohnle, E. M. Fuchs, and P. Aravena Pelizari. Object-Based Analysis and Fusion of Optical and SAR Satellite Data for Dwelling Detection in Refugee Camps. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5):1780–1791, 2017.

[17] K. Spröhnle, D. Tiede, E. Schoepfer, P. Füreder, A. Svanberg, and T. Rost. Earth observation-based dwelling detection approaches in a highly complex refugee camp environment - A comparative study. *Remote Sensing*, 6(10):9277–9297, 2014.

[18] S. Wang, E. So, and P. Smith. Detecting tents to estimate the displaced populations for post-disaster relief using high resolution satellite imagery. *International Journal of Applied Earth Observation and Geoinformation*, 36:87–93, 2015.

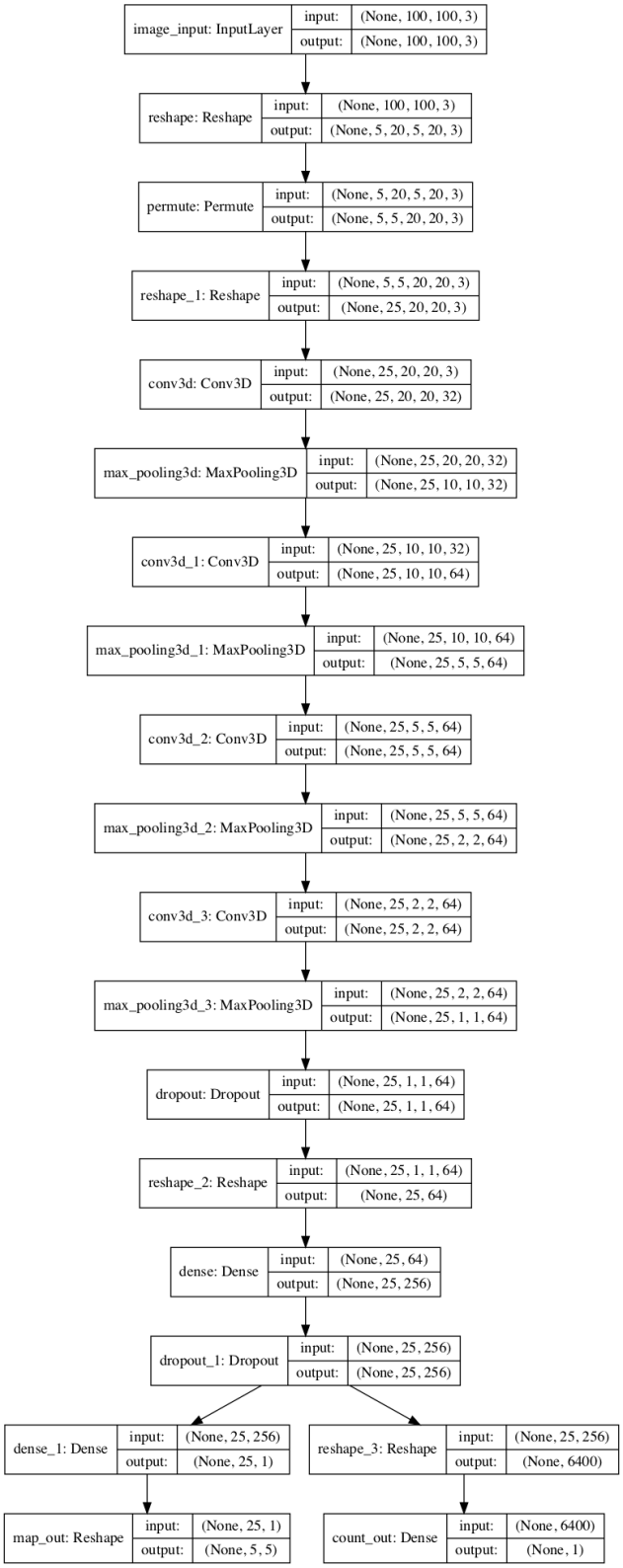

Figure A1. The hybrid classification and regression model architecture, including input and output shapes for each layer.