remote sensing

MDPI

Article

Mapping Human Settlements with Higher Accuracy and Less Volunteer Efforts by Combining Crowdsourcing and Deep Learning

Benjamin Herfort *, Hao Li, Sascha Fendrich, Sven Lautenbach and Alexander Zipf

GIScience Chair, Institute of Geography, Heidelberg University, 69120 Heidelberg, Germany
* Correspondence: herfort@uni-heidelberg.de; Tel.: +49-6221-54-5534

Received: 26 June 2019; Accepted: 26 July 2019; Published: 31 July 2019

Abstract: Reliable techniques to generate accurate data sets of human built-up areas at national, regional, and global scales are a key factor to monitor the implementation progress of the Sustainable Development Goals as defined by the United Nations. However, the scarce availability of accurate and up-to-date human settlement data remains a major challenge, e.g., for humanitarian organizations. In this paper, we investigated the complementary value of crowdsourcing and deep learning to fill the data gaps of existing earth observation-based (EO) products. To this end, we propose a novel workflow to combine deep learning (DeepVGI) and crowdsourcing (MapSwipe). Our strategy for allocating classification tasks to deep learning or crowdsourcing is based on confidence of the derived binary classification. We conducted case studies in three different sites located in Guatemala, Laos, and Malawi to evaluate the proposed workflow. Our study reveals that crowdsourcing and deep learning outperform existing EO-based approaches and products such as the Global Urban Footprint. Compared to a crowdsourcing-only approach, the combination increased the quality (measured by Matthew's correlation coefficient) of the generated human settlement maps by 3 to 5 percentage points. At the same time, it reduced the volunteer efforts needed by at least 80 percentage points for all study sites. The study suggests that for the efficient creation of human settlement maps, we should rely on human skills when needed and rely on automated approaches when possible.

Keywords: volunteered geographic information; human settlements; deep learning; humanitarian mapping; building detection; crowdsourcing

1. Introduction

Currently, 55% of the world's population reside in urban areas, and especially in low-income and lower-middle-income countries rapid urbanization is expected between now and 2050 [1]. The Sustainable Development Goals (SDGs) [2] and the Sendai Framework for Disaster Risk Reduction (SFDRR) [3] both highlight the relevance and increasing need for up-to-date information on the spatial distribution of human settlements. For instance, humanitarian organizations cannot help people if they cannot find them. Consequently, reliable techniques to generate accurate data sets of human settlements at national, regional, and global scales are crucial in manifold domains such as disaster management, habitat and ecological system conservation, and public health monitoring.

Earth observation (EO) using satellites already provides data for a broad range of purposes such as disaster assessment, forestry or crop land monitoring, and land-use/land-cover classification. Recently, remote sensing technologies have been successfully employed to derive information on human settlements at regional to global scales. Accuracy and completeness of EO derived human settlement data sets have improved a lot in the last 15 years. Current data sets, which have been made available recently, include the Global Human Settlement Layer (GHSL) [4], the Global Urban Footprint (GUF) data set [5], and the High-Resolution Settlement Layer (HRSL) [6]. However, these data sets still show great variations for different regions and geographic settings [7]. Especially rural areas and non-solid building structures are still disregarded or under-represented in these data sets.

Several researchers highlight the potential of crowdsourcing to collect information on human settlements and to complement the data that is produced using satellite imagery [8–10]. Additionally, humanitarian organizations start using new methods from Citizen Science and Volunteered Geographic Information (VGI), to gather information on the spatial distribution of human settlements [11]. However, the quality and the reliability of those methods and resulting data sets remain major concerns, which are extensively discussed in current research [12]. Spatial varying data quality and the lack of reference data with sufficient quality still constitute barriers in using VGI data in general and for humanitarian purposes or in disaster management in particular [13].

The mapping of human settlements tends to be done either from an earth observation perspective or from a citizen science position. However, a tighter integration of both approaches has the potential to derive improved data sets that presumably outperform existing one [8].

In this article, we present a novel workflow to overcome the scarce availability of accurate and up-to-date human settlement data sets (see Figure 1). Our proposed workflow combines two methods: (1) object detection and classification using deep learning algorithms (DeepVGI [14]) and (2) crowdsourced mapping of human settlements by volunteers (MapSwipe [15]). To combine both methods we propose a task allocation strategy (3) that choses classification labels either from DeepVGI or MapSwipe. Moreover, we investigate whether our proposed methodology helps to produce better maps faster with respect to the following research questions.

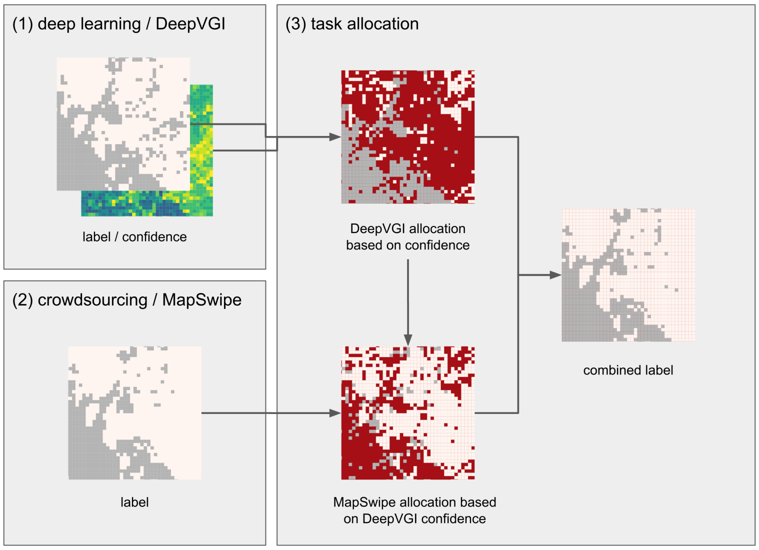
(1) deep learning / DeepVGI
label / confidence
(2) crowdsourcing / MapSwipe
label
(3) task allocation
DeepVGI allocation based on confidence
MapSwipe allocation based on DeepVGI confidence
combined label

Figure 1. Proposed workflow to combine deep learning and crowdsourcing methods: Combined labels are obtained, by choosing labels either from DeepVGI or MapSwipe based on the confidence of the DeepVGI labels.

- RQ1: How good are crowdsourcing (MapSwipe) and deep learning (DeepVGI) with respect to generating human settlement maps in comparison to existing EO-based approaches?
- RQ2: Which spatial and non-spatial characteristics of misclassifications are accompanied by applying the DeepVGI approach?
- RQ3: What is the added value of the proposed task allocation strategy with respect to performance and effort?

The remainder of this paper is organized as follows: Section 2 provides background information on techniques for deriving information on human settlements using either deep learning or crowdsourcing. The study areas and data sets are described in Section 3. We present our overall methodology in Section 4 and show results in Section 5. Section 6 discusses our findings and makes suggestions for future work, whereas Section 7 draws conclusions.

2. Background: Mapping Human Settlements Using Crowdsourcing or Deep Neural Networks

Previous research shows the strengths and shortcomings of existing global settlement data products such as the Global Human Settlement Layer or the Global Urban Footprint data set [7,16]. In this section we review to what degree VGI and citizen science approaches and techniques based on deep learning have been applied to tackle the known challenges.

Citizen science projects and VGI tools are widely used to collect information on human settlements. The OpenStreetMap (OSM) platform played a central role in generating map data during several international disasters, e.g., after the 2010 Haiti earthquake [17] and after the 2015 Nepal earthquake [18]. Additionally, OpenStreetMap is used also in disaster preparedness and disaster risk reduction activities, for instance, by the organizations of the Missing Maps project [11]. Regarding urban planning, Crooks et al. (2015) [19] show how various user-generated data sets (GPS trajectories, social media data) enrich our understanding of urban form and function from a bottom up perspective.

The increased usage of VGI has been accompanied by discussions on the quality of data sets, which have not been produced by experts [20,21]. The research on VGI data quality reveals that spatial heterogeneity, e.g., regarding completeness of building footprints, remains a major challenge on different geographic scales [13]. Fan et al. (2014) [22] confirm a high completeness for building footprint features in an urban area in Munich, Germany. For crowdsourced classification of human settlements in Madagascar and South Sudan Herfort et al. (2017) [15] conclude that disagreement between users is not randomly distributed in space but rather clustered, indicating that reliability of information varies spatially. Additionally, Comber et al. (2016) [23] show for land cover mapping that the quality of VGI data sets is influenced by differences between user groups, which are a potential source of error and uncertainty.

In addition to the above VGI-related work, several authors have investigated the potential of deep learning technology in various satellite image processing tasks including human settlement mapping (see [24] for a comprehensive overview). In Jean et al. (2016) [25] a convolutional neural network is used to distinguish urban areas, non-urban area, roads, and water in optical satellite imagery for predicting poverty in Nigeria, Tanzania, Uganda, Malawi, and Rwanda. In Li et al. (2019) [26], a pre-trained neural network is employed to estimate large scale OSM missing built-up areas in Tanzania. Regarding land cover mapping, several authors propose workflows based on deep neural networks with a focus on urban areas [27,28]. Furthermore, building footprint extraction based on deep learning has been a central research topic in recent years [29,30].

Although deep learning shows promising results with respect to object detection in images in general, analyzing data quality of geographic approaches (e.g., using geographic data sets such as satellite imagery and building footprints) remains a huge challenge. The transferability of deep learning models constitutes a key challenge towards global scale data products, e.g., for human settlements. For instance, Yuan et al. (2018) [30] highlight difficulties in extracting footprints for buildings in rural areas which differed significantly from the footprints presented in the training set. Missing benchmark data sets tailored to remote sensing tasks make it difficult to compare the growing number of deep learning algorithms [24]. Similar to the factors contributing to spatial heterogeneity of VGI data sets, deep learning approaches are vulnerable to changes in input factors such as atmospheric scattering conditions, intraclass variability, culture-dependent characteristics and a limited number of training samples [24].

In addition to studies focusing either on deep learning or on VGI approaches in isolation to generate information on human settlements, some studies have successfully combined both approaches. A study by Gueguen et al. (2017) [8] produces regional- and country-scale population distribution maps from very high-resolution satellite imagery based on the detection of village boundaries by a deep neural network and a crowdsourced validation of the results. The study reports benefits from combining the high recall of automated methods with the high precision of human validators. By combining data from multiple crowdsourcing projects in an active learning framework for convolutional neural networks Chen et al. (2018) [31] address incompleteness and spatial heterogeneity of input training samples regarding road and building mapping. Their results show a promising avenue how deep learning can be used to improve VGI data. However, the small sample size used for validation hinders conclusions on the transferability of their findings. Vargas-Munoz et al. (2019) [32] investigate the quality of OSM data in study sites in Zimbabwe and Tanzania using a deep learning approach. Their approach can detect missing building footprints and misalignment in the OSM data, but the validation data sets contain only 1000 buildings per site, which again casts doubts on the transferability of the proposed approach.

Previous work has shown how citizen science, VGI, and deep learning can contribute to improve large scale geographic data sets on human settlements. Those methods help to produce data sets desperately needed for monitoring urban growth, sustainable development, disaster risk reduction, and many other applications. Nevertheless, researchers have also revealed that spatial heterogeneity is a key issue, which needs to be addressed to understand and enhance data quality. Regarding VGI data sets, spatial heterogeneity is expressed by regional difference in data completeness and varying data quality due to diverging user experience. Considering deep learning approaches, spatial heterogeneity can be interpreted as the difficulty to transfer models from one region to another and to provide training samples which incorporate the geographic properties for all object structures or characteristics and regions.

We propose a workflow to combine the strengths of recent machine learning algorithms and crowdsourced data production by means of a confidence-based task allocation strategy. Our work is guided by the common hypothesis that humans rarely identify something as a building which is not a building, but tend to miss some objects. Furthermore, deep learning approaches will miss fewer buildings at the cost of detecting also several objects which are not buildings. By bringing together those two research streams we aim at producing human settlement data sets which are both more complete and precise.

3. Description of the Study Areas and Data Sets

3.1. Study Areas

We investigated our combined mapping workflow at three study sites: (a) Guatemala, (b) Laos and (c) Malawi. The study sites covered an area between 675 to 2700 square kilometers (see Table 1). For Guatemala and Laos training regions were slightly bigger than testing regions, whereas for Malawi the opposite was the case. Regarding the testing sites, all case studies showed an imbalanced proportion of “no building” and “building” tiles. This imbalance was very strong towards “no building” tiles for Guatemala and Laos (87%), and less marked for Malawi (56%).

All study regions have been mapped by OpenStreetMap users in response to requests by the Humanitarian OpenStreetMap Team (HOT), The Netherlands Red Cross, and the Clinton Health Initiative. Our analysis was based on the tile level. In this study a tile corresponded to the definition applied by tile map services (TMS) at zoom level 18 [33]. Hence, each study region was made of several thousand tiles. A tile covered around 0.02 square kilometers (0.15 × 0.15 kilometers) with slight variations depending on geographic latitude. Table 1 summarizes the details about the training and testing areas for each study site.

Table 1. Study Sites Characteristics.

		Guatemala	**Laos**	**Malawi**
Training	Area	929.0 km^2	1556.3 km^2	265.7 km^2
	Tiles	42,833	72,360	12,408
Testing	Area	745.5 km^2	1136.6 km^2	410.3 km^2
	Tiles	34,181	52,796	19,272
	No Building Tiles	87%	87%	56%

The Guatemala region was made up of several different rather contrary regions. The northern and western areas were slopes of volcanic mountains covered with dense forests, whereas urbanized areas and agricultural land covered the valleys. Most settlements were part of a compact city with only a few buildings lying within farmland. The Laos region was characterized by dense forested areas covering more than 90% of the total area and by a few smaller settlements along the road network. Besides those villages, many buildings were distributed sparsely over the entire area accompanied by smaller patches of agricultural land. For Malawi larger cities were missing. The region was characterized by many smaller villages and intensive agricultural land use. Only a very small fraction of the area was covered by forests.

3.2. Data Sets

This section describes the four datasets we used to evaluate the performance of our workflow. The reference data set was derived from OpenStreetMap and depicts the presence of human settlements in a satellite imagery tile. Building footprint geometries were directly obtained from the OpenStreetMap database using the Overpass API by filtering for all OSM ways tagged with the key "building" for each study site. To ensure the validity of the OSM data, we intentionally selected study areas for which the mapping was organized through the HOT Tasking Manager tool and a validation had taken place. This validation procedure ensured the precision of the OSM data set. More important, humanitarian mapping experts carefully validated all tiles for which no human settlements have been mapped in OSM, but a positive result was given either in MapSwipe or predicted by the deep neural network. Following this additional manual validation approach ensured the completeness of the reference data set.

MapSwipe yielded results on the tile level. In MapSwipe a tile was also called a task. For each individual tile at least three different users contributed a binary label ("building", "no building"). These results have been aggregated using majority voting. Furthermore, for each task the proportion of building labels (MapSwipe score) on the total number of labels has been generated. For example, if two out of three volunteers classified a tile as "building", the aggregated label would be considered to be "building" (MapSwipe score: 0.66). Further details on the MapSwipe data model can be found in [15].

The Global Urban Footprint has been derived fully automatically from TanDEM-X and TerraSAR-X radar images with 3 m ground resolution by the German Aerospace Center (DLR) [5]. The imagery has been collected between 2011 and 2012. In this paper, we used the binary GUF settlement mask with a spatial resolution of 0.4″, which corresponds to a ground resolution of around 12 m at the equator.

The High-Resolution Settlement Layer maps human settlements derived from high-resolution satellite imagery (0.5 m) by the Connectivity Lab at Facebook. The data has been produced for 18 countries using deep neural networks [6].

4. Methodology

The workflow we propose in this paper addresses the challenge of combining deep learning and crowdsourcing to generate high-quality human settlement maps. Section 4.1 explains the DeepVGI method to automatically classify satellite imagery tiles into "building" and "no building". Section 4.2 explains the data quality evaluation procedure. In Section 4.3 we present the procedure to analyze spatial and non-spatial characteristics of misclassifications of the DeepVGI method. Finally, Section 4.4 investigates the proposed task allocation strategy and how the combined use of MapSwipe and DeepVGI affects performance and volunteer efforts.

4.1. Data Preparation

We employed the DeepVGI method presented in Figure 2 for classifying satellite imagery tiles into "no building" and "building" classes. The DeepVGI building detection model consisted of three parts: feature extraction, object detection and binary classification.

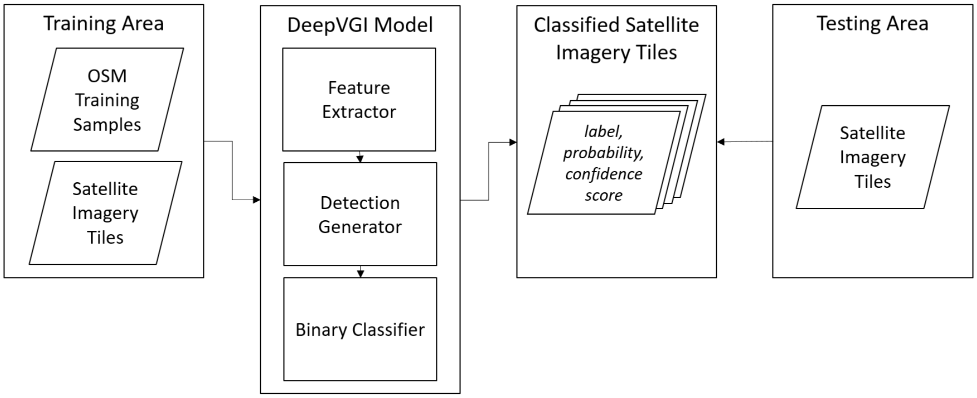

Figure 2. DeepVGI Workflow: The DeepVGI model is trained using building footprint sample from OpenStreetMap and satellite imagery tiles from Bing. For each tile in the testing area the model generates label, probability and confidence score.

Based on Single Shot Detection (SSD) networks [34] the model extracted heterogeneous features from either the base network or from extra layers. This enables SSD networks to better handle complex objects (such as buildings) of diverse scale and shape. For doing so, SSD networks apply the concept of tiling into default boxes so that specific feature maps learn to be responsive to particular scales of the objects [34].

The object detection generated a set of predicted building bounding boxes together with the corresponding probability scores. The implementation of the SSD network was based on the programming language Python 3.6 and the deep learning library Tensorflow [35]. For the initialization of the SSD parameters, a pre-trained network based on the Microsoft COCO data set [36] has been employed, which reported a mAP^{-1} (mean average precision) of 24. The maximum training epochs has been set to 60,000, and the initial learning rate is set to 0.0004 with a momentum of 0.9. The pre-trained network is available at the Tensorflow detection model zoo [37].

The object detection generated up to 50 bounding boxes per tile; however, the majority of those bounding boxes did not represent buildings and were associated with very low probabilities. Since we were interested in a binary classification of a tile t into "no building" and "building" we only selected the highest bounding box score $P_{\max}(t)$ for each tile. From the training data set derived a classification threshold θ. Tiles with $P_{\max}(t)$ below θ were most probable to belong to the "no building" class. Vice-versa tiles with $P_{\max}(t)$ above θ were most probable to belong to the "building" class". Based on θ we generated the label ("no building" and "building") for each tile in the testing data set.

Finally, we derived a confidence score δ_t for each tile by computing the absolute difference between the binary threshold θ and (deep learning) probability $P_{\max}(t)$. This was used as a proxy for how confident we can be in binary classification into either "building" or "no building". Taking the absolute value eased the visual interpretation. However, situations where the highest bounding box score was lower than the threshold (potential false negatives) could not be distinguished from those where the bounding box score was higher than the threshold (potential false positives) from δ_t alone. Hence, the analysis of the confidence score provided insights into accuracy of the DeepVGI approach, but not towards its specificity or sensitivity.

To train the DeepVGI model, building footprint samples from OpenStreetMap and satellite imagery tiles from Microsoft Bing were employed. In our experiment, satellite imagery tiles were collected by requesting a tile map service (TMS) from Microsoft Bing at zoom level 18. This corresponded to a spatial resolution of the displayed image of roughly 0.6 m per pixel, as measured at the equator. The size of all image tiles was 256 × 256 pixels.

4.2. Overall Performance Evaluation

We evaluated our method by investigating the quality of the produced results against the reference data (see Section 3.2). Initially we derived the proportions of false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP). To address the imbalance of building and no building labels in our study areas, we used the following metrics: specificity (TNR), sensitivity (TPR), and Matthews correlation coefficient (MCC). We further derive the accuracy (ACC). The statistics were computed as shown in Equation (1) through Equation (4). TNR, TPR and ACC are restricted between 0 and 1. MCC is in essence a correlation coefficient between the observed and predicted binary classification. It is bound between -1 and 1. For all statistics higher values indicate a better model fit [38].

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{1}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{3}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{4}$$

Whereas ACC was mainly used to compare the results with works of other authors, MCC gives a more reliable indicator on quality for imbalanced data sets [38]. TNR and TPR were used to investigate how well the analyzed methods identify positives ("building" class) and negatives ("no building" class). This provided insights on the strengths and weaknesses of each method. Other metrics commonly applied for machine learning performance assessment such as F1 score or precision were not considered since they are highly biased for imbalanced data sets [38].

Additionally, we generated a map representation of the confusion matrix for each method and study site to spot spatial pattern in the false negatives and false positives.

4.3. Spatial and Non-Spatial Characteristics of Misclassifications

After investigating the overall performance of all methods, we conducted a detailed analysis of spatial and non-spatial characteristics of misclassifications of the DeepVGI approach. For doing so, we generated kernel density distribution plots of the confidence scores δ_t for (a) all tiles, (b) "building" tiles and (c) "no building" tiles. The implementation of the kernel density functions was based on the programming language Python, version 3.6 and the scipy library, version 1.1.0 [39]. The kernel bandwidth was set by applying Scott's rule [40]. We generated a map representation of the distribution of δ_t to analyze the spatial characteristics of the confidence in the presence of buildings. This map should be interpreted together with the maps showing the distribution of false negatives and false positives.

Furthermore, we generated conditional density plots to visualize the conditional distribution of accuracy ACC in respect to δ_t and to compare the performance of the DeepVGI and MapSwipe approaches.

We tested if tasks with lower δ_t had a higher probability of being wrong by using a logistic regression model for the DeepVGI and MapSwipe approaches, using δ_t as the predictor and Y as the response. Y was defined as "0" for wrong classifications (e.g., DeepVGI label and reference label were not the same) and "1" for correct classifications (e.g., DeepVGI label and reference label were the same). For the logistic regression we report on regression coefficient, standard error, significance, and McFadden's pseudo-r-squared values [41]. Those were computed based on the programming language Python 3.6 and the statsmodels library, version 0.9.0 [42].

4.4. Performance of Task Allocation Strategy

Finally, we proposed a task allocation strategy based on δ_t and compared the approach to a random allocation of tiles between DeepVGI and MapSwipe. The task allocation strategy defines for which tasks it is better to rely on results being produced by crowdsourcing and for which it is preferable to use the DeepVGI workflow. First, tiles were sorted ascending by δ_t. We generated the labels based on the fusion of both approaches by choosing a proportion α of tiles which should be labeled by the crowd (we will refer to this as crowd proportion). For the remaining tiles we assigned the label of the DeepVGI method. Due to this design tiles with the lowest confidence were allocated to the crowd first.

We generated 100 set of the combined labels for each study site by choosing a crowd proportion α between [0,1] and adopting a step size of 0.01. For each α we further derived 250 random combinations of MapSwipe and DeepVGI. For each α we investigated the performance of both task allocation strategies in terms of accuracy ACC, Matthews correlation coefficient MCC, specificity TNR and sensitivity TPR (described in Section 4.2). The results were visualized in a graph depicting performance in relation to the crowd proportion α.

5. Results

5.1. Overall Performance Evaluation

This part of the results section describes the performance of different methods to create human settlement data sets. For all study sites we analyzed data from crowdsourcing (MapSwipe), deep neural networks (DeepVGI) and existing EO products (GUF). The HRSL data set was only available for the Guatemala and Malawi study sites.

For all three case studies MapSwipe performed best with respect to accuracy ACC (91–96%) and Matthews correlation coefficient MCC (80–83%) (see Table 2). The MapSwipe approach was characterized by high specificity TNR (99% for all study sites) and intermediate sensitivity TPR (75–82%). The spatial representation of the confusion matrices indicated that MapSwipe was able to correctly depict most "no building" tiles. Nevertheless, for the MapSwipe approach clusters of false negatives were observed, e.g., in the north-western part of the Guatemala study site (Figure 3) or in the south-eastern part of the Malawi study site (Figure 5). Thus, MapSwipe users were more likely to miss buildings (false negatives), than to map something as a building, which actually is not a building (false positives).

Regarding accuracy ACC the DeepVGI approach reached similar results to MapSwipe (91–96%); however the differences between those approaches became slightly more distinct when considering the less biased MCC (74–84%). The DeepVGI approach achieved lower specificity TNR compared to MapSwipe and GUF (95–97%), but higher sensitivity TPR (81–89%). Hence, false positives were expectedly the major concern for the DeepVGI approach. The visual interpretation of the confusion matrix maps confirmed higher spatial concentrations of false positives. For instance, clusters of false positives were present in the central and southern part of the Guatemala study site (Figure 3) or in the north-western part of the Malawi study site (Figure 5).

GUF was characterized by lower ACC (58–92%) and MCC (18–60%) compared to MapSwipe or DeepVGI. Regardless the very high specificity TNR (>99% for all study sites), the data set mapped only parts of the buildings as described by low sensitivity TPR (6–44%). The huge number of false negatives was the major drawback of the GUF approach. The map representation of the results for Guatemala (Figure 3) indicates that only major settlements were captured. For the rural study sites in Laos (Figure 4) and Malawi (Figure 5) the GUF approach generated many false negatives and thus missed most buildings.

The HRSL approach ranked below MapSwipe and DeepVGI in terms of accuracy ACC (86–90%) and MCC (69–73%). This approach reached the best results regarding sensitivity TPR (94–96%), but achieved only moderate specificity TNR (80–89%). Similar to the DeepVGI approach, the HRSL suffered from a high number of false positives. This is also depicted in the maps for Guatemala (Figure 3) and Malawi (Figure 5). It seems that false positives were located at the edges of correctly identified building tiles and along the road network.

Table 2. Performance of different methods to generate human settlement data sets.

		Guatemala	**Laos**	**Malawi**
TNR	MapSwipe	0.99	0.99	0.99
	DeepVGI	0.96	0.97	0.95
	GUF	0.99	0.99	0.99
	HRSL	0.89	-	0.80
TPR	MapSwipe	0.74	0.79	0.82
	DeepVGI	0.81	0.89	0.85
	GUF	0.44	0.06	0.07
	HRSL	0.96	-	0.94
ACC	MapSwipe	0.96	0.97	0.91
	DeepVGI	0.94	0.96	0.91
	GUF	0.92	0.87	0.58
	HRSL	0.90	-	0.86
MCC	MapSwipe	0.80	0.85	0.83
	DeepVGI	0.74	0.84	0.81
	GUF	0.60	0.22	0.18
	HRSL	0.69	-	0.73

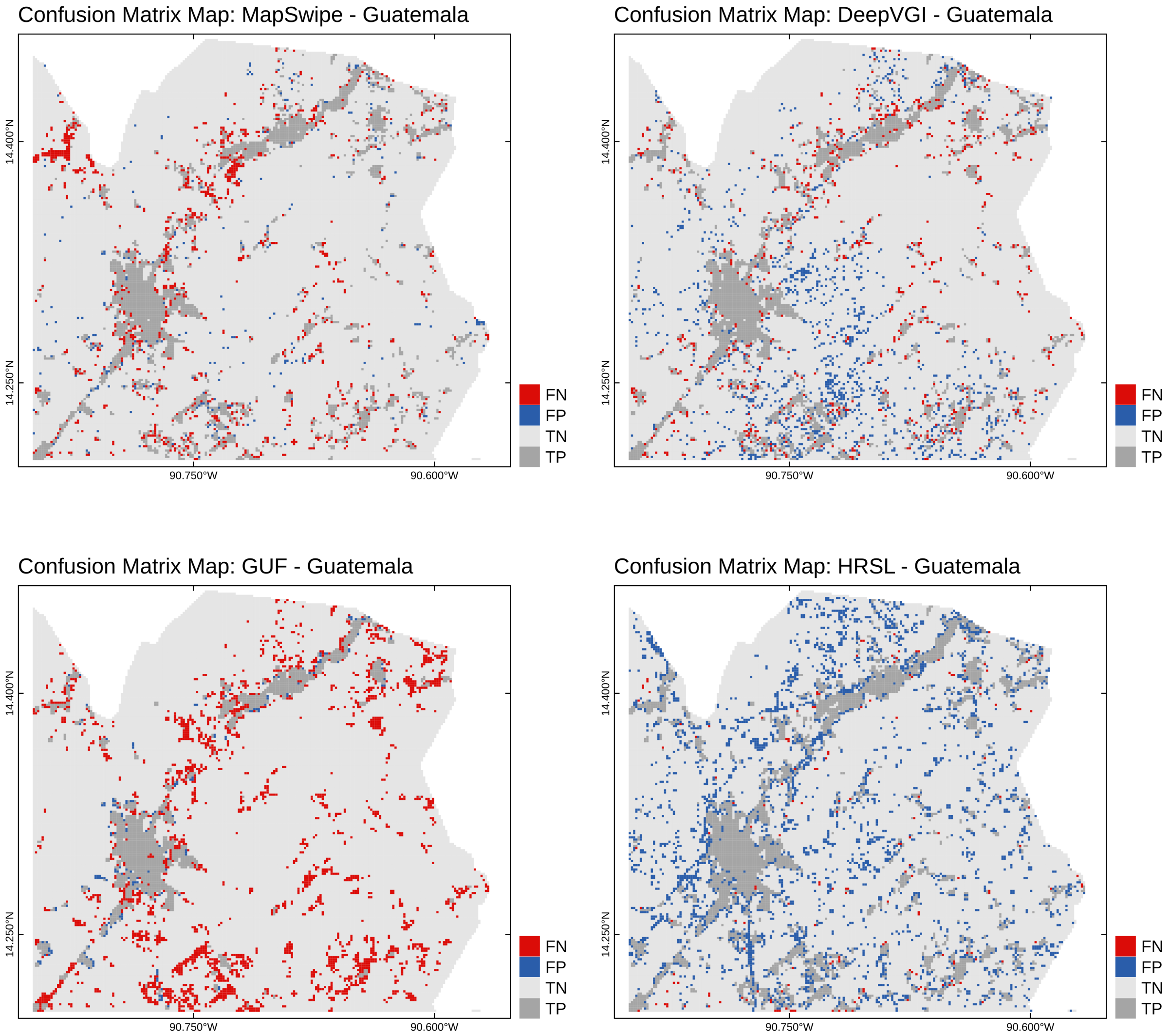
Confusion Matrix Map: MapSwipe - Guatemala
Confusion Matrix Map: DeepVGI - Guatemala
Confusion Matrix Map: GUF - Guatemala
Confusion Matrix Map: HRSL - Guatemala
14.400°N
14.250°N
90.750°W
90.600°W
FN
FP
TN
TP

Figure 3. Map representation of the confusion matrix for Guatemala: Each map shows the spatial distribution of correct "building" (TP) and "no building" (TN) classifications for a specific method in the testing area. Incorrect classifications are split into false positives (FP) and false negatives (FN). Each pixel corresponds to a single task/satellite imagery tile.

5.2. Spatial and Non-Spatial Characteristics of Misclassifications

This section focuses on the analysis of the characteristics of misclassified tiles of the DeepVGI approach.

For all study sites, the distribution density of "no building" tiles showed a clear peak, whereas the distribution of "building" tiles was much flatter. For instance for the Guatemala study site, Figure 6 (center) depicts that the distribution density of "no building" tiles peaked for a confidence score of around 0.25. "Building" tiles distributed equally between confidence scores of 0.0 and 0.45. Very high confidence scores (>0.3) are observed only for "building" tiles. At the same time, for very low confidence scores (<0.15) both "no building" and "building" tiles were present.

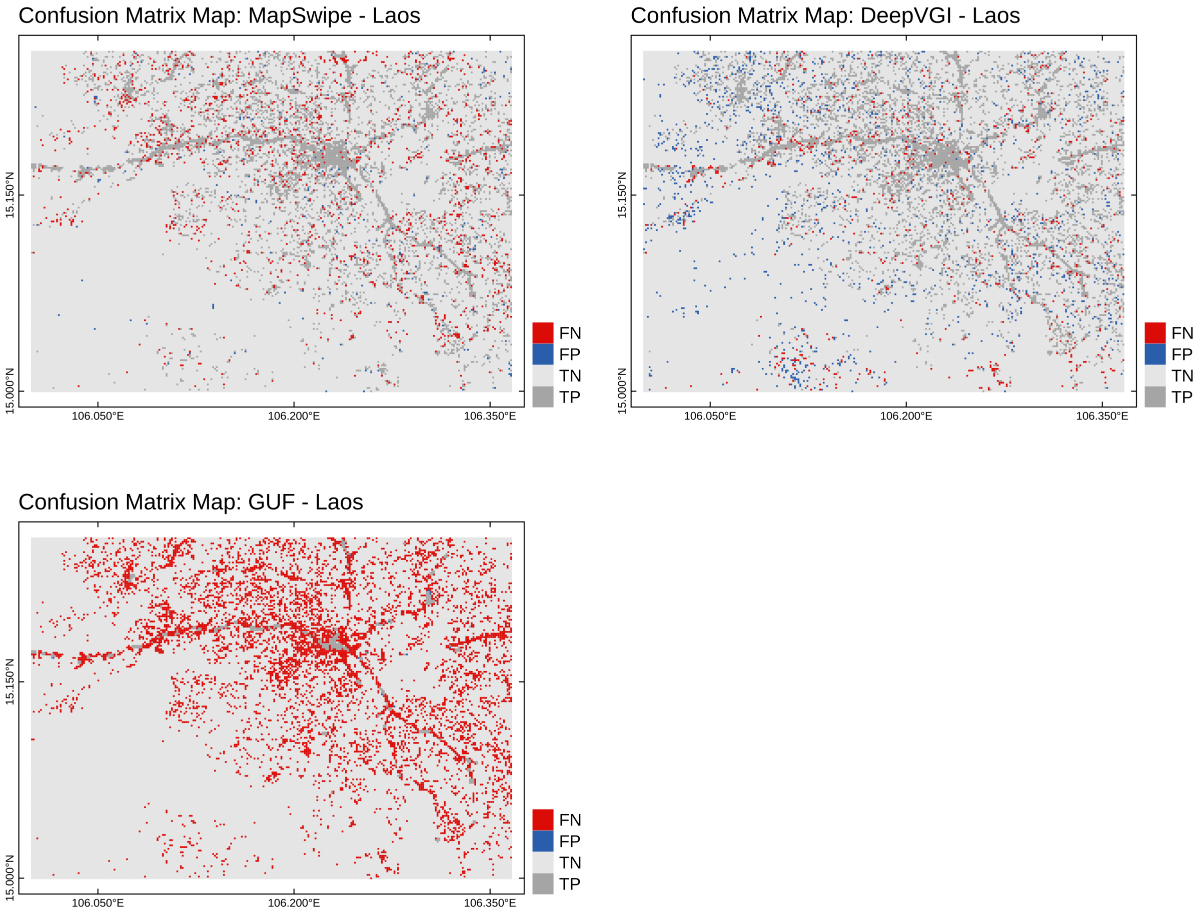

Figure 4. Map representation of the confusion matrix for Laos: Each map shows the spatial distribution of correct "building" (TP) and "no building" (TN) classifications for a specific method in the testing area. Incorrect classifications are split into false positives (FP) and false negatives (FN). Each pixel corresponds to a single task/satellite imagery tile.

The conditional density plots (Figure 6, blue axis) revealed the tendency that the accuracy of the DeepVGI approach increased with higher confidence scores. For all three study sites, accuracy increased steadily from around 50–60% to more than 95% when raising the confidence score from 0.0 to around 0.35. There was no such clear trend for the conditional density of the accuracy for the MapSwipe approach. MapSwipe's accuracy ranged between 80% and 95% and indicated no dependency from the confidence score. Additionally, the comparison of the conditional density plots from DeepVGI and MapSwipe underlined that tiles that were relatively easy for DeepVGI (high confidence scores) were on average not easy for MapSwipe users. Vice versa, tiles that were difficult for the DeepVGI approach (lower confidence scores), were on average not particularly more difficult for MapSwipe users. For example for the Guatemala study site, the DeepVGI approach provided more accurate results than MapSwipe for tiles with a confidence score higher than 0.25. For tiles with a confidence score below 0.2 MapSwipe reached higher accuracy.

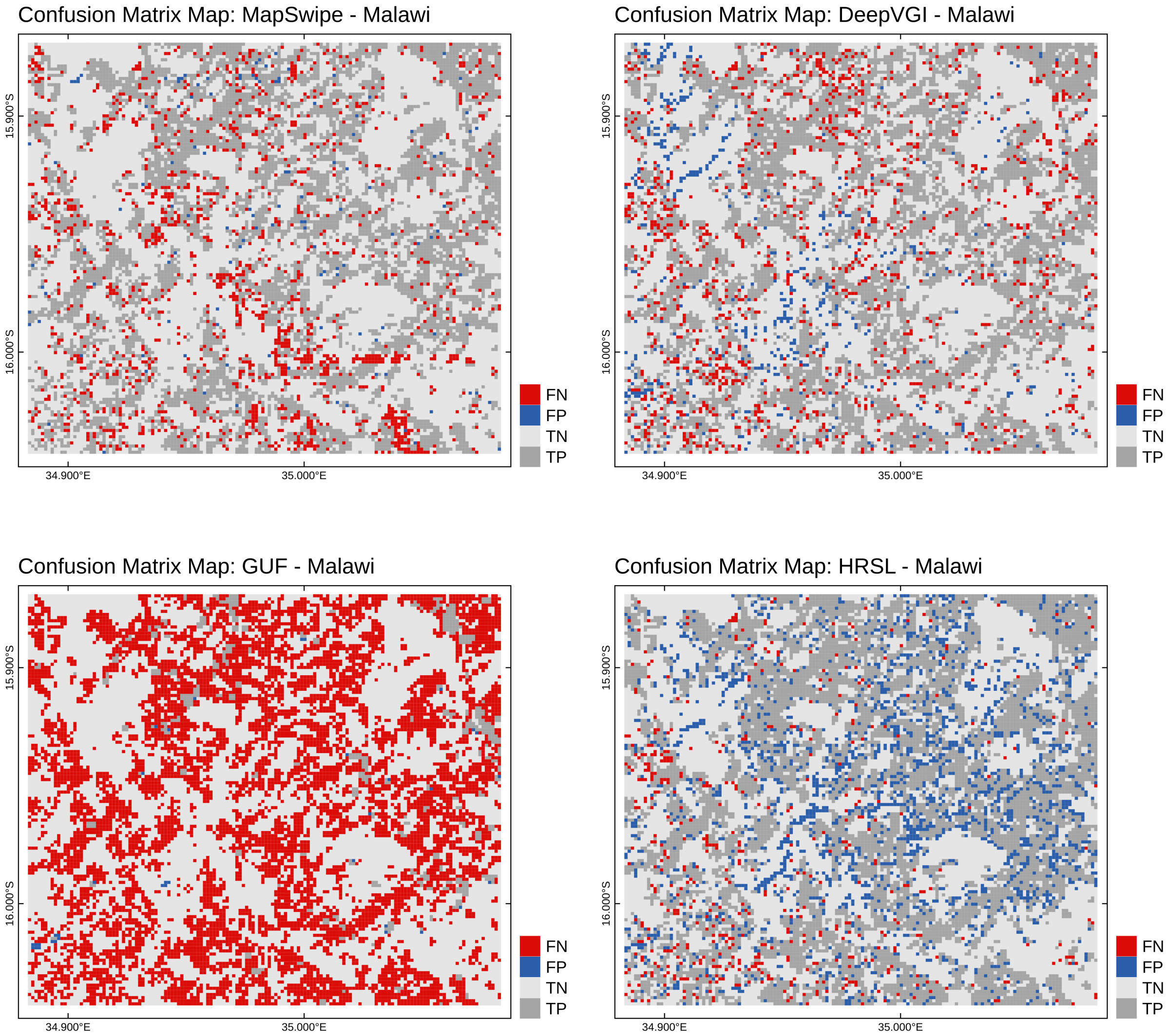
Confusion Matrix Map: MapSwipe - Malawi
Confusion Matrix Map: DeepVGI - Malawi
Confusion Matrix Map: GUF - Malawi
Confusion Matrix Map: HRSL - Malawi
15.900°S
16.000°S
34.900°E
35.000°E
🟥 FN
🟦 FP
⬜ TN
⬛ TP

Figure 5. Map representation of the confusion matrix for Malawi: Each map shows the spatial distribution of correct "building" (TP) and "no building" (TN) classifications for a specific method in the testing area. Incorrect classifications are split into false positives (FP) and false negatives (FN). Each pixel corresponds to a single task/satellite imagery tile

The spatial distribution of the confidence score revealed highest confidence scores for major settlements for all study sites. The most uncertain results were located in the areas with a mixed land-use, e.g., agricultural land and settlements. Furthermore, uncertain results seemed to be clustered, e.g., for the Laos study site (Figure 6) in most of the southern part of the study area and in a smaller area in the northern part.

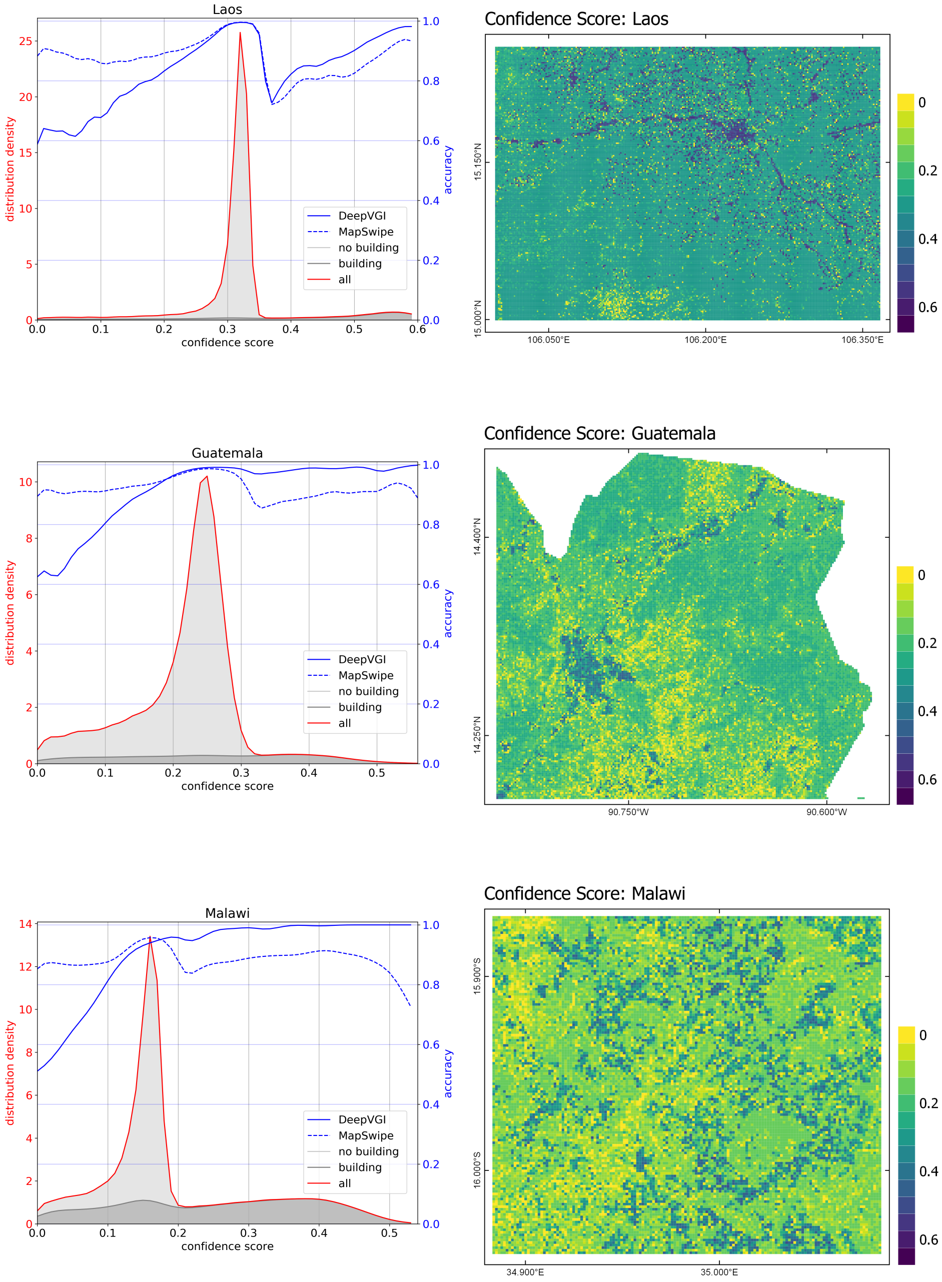

Figure 6. Spatial and Non-Spatial Distribution of the Confidence Score and Conditional Density of Accuracy for Laos, Guatemala and Malawi. The figure provides insights on the kernel density distributions of the confidence scores of “all” (red line), “no building” (light gray) and “building” (dark gray) tiles for the different study regions (red axis). The same plot shows the conditional distribution of accuracy (blue axis) for DeepVGI (solid blue line) and MapSwipe (dashed blue line). The map depicts the spatial distribution of the confidence scores.

The indications from the conditional density plot were confirmed by the results of the logistic regression analysis (Table 3): as depicted by the regression coefficient and the corresponding standard error and significance with increasing confidence score, probability of correct classifications increased significantly. However, as indicated by the low pseudo-r-squared values of (0.109–0.210) the confidence score explained only a fraction of all misclassified tiles. The results further confirmed that the confidence score had no explanatory power with respect to the accuracy of the MapSwipe approach (McFadden's Pseudo-r-squared of −0.211 to −0.130).

Table 3. Logistic Regression analysis.

		Guatemala	**Laos**	**Malawi**
DeepVGI	Coefficient	16.262	11.164	16.291
	Standard Error	0.165	0.086	0.210
	Significance	0.0 ***	0.0 ***	0.0 ***
	McFadden's Pseudo-r-squared	0.235	0.109	0.194
MapSwipe	Coefficient	15.224	10.454	11.676
	Standard Error	0.149	0.078	0.149
	Significance	0.0 ***	0.0 ***	0.0 ***
	McFadden's Pseudo-r-squared	−0.13	−0.154	−0.211

5.3. Combination of Crowdsourcing and Deep Learning

Figure 7 shows performance and effort in respect to crowd proportion. For all study sites allocating 10% to 20% of the tiles to MapSwipe (raising the crowd proportion from 0.0 to around 0.1–0.2) resulted in an overall performance increase in respect to accuracy ACC and Matthew's correlation coefficient MCC. Reducing the volunteer efforts to one fifth (labor reduction of 80 percentage points) resulted in a performance gain of 3–5 percentage points measured as MCC in all regions. For all study sites, this was caused mainly due to an increase in TNR (compared to the DeepVGI-only approach). For Guatemala and Laos, TPR remained stable, whereas for Malawi TPR increased as well (compared to the DeepVGI-only approach).

Vice versa, allocating 10% to 30% of the tiles to DeepVGI (decreasing the crowd proportion from 1.0 to around 0.7–0.9) also resulted in an overall performance gain. For all study sites, a gain in TPR (compared to the MapSwipe-only approach) was observed the more tiles have been allocated to DeepVGI. For Guatemala TNR remained stable, whereas for Laos and Malawi TNR decreased slightly at the same time.

The performance of the combined approach did not change, when allocating 20% to 70% of the sorted tiles to MapSwipe. For crowd proportions between 0.2 and 0.7 ACC, MCC, TNR and TPR remained mainly stable at a higher level compared to DeepVGI-only or MapSwipe-only approaches. For instance for the Guatemala study site, MCC gained around 10 percentage points (75% to 85%) compared to the performance of the DeepVGI-only approach by allocating 30% of the tiles to MapSwipe and 70% to DeepVGI. At the same time this resulted in an increase in MCC of around 5 percentage points (80% to 85%) and labor reduction of 70 percentage points compared to the performance of the MapSwipe-only approach. The MCC for a crowd proportion of 0.7 varied only slightly and reached around 84%.

When investigating the results of the random task allocation (see Figure 7) no such effects were observed. An increase in crowd proportion resulted in an improved TNR and decreased TPR with a uniform gradient for all study sites. Overall, ACC and MCC improved slightly with a homogeneous slope when allocation tasks from DeepVGI to MapSwipe. However, no performance gain was observed compared to the MapSwipe-only approach.

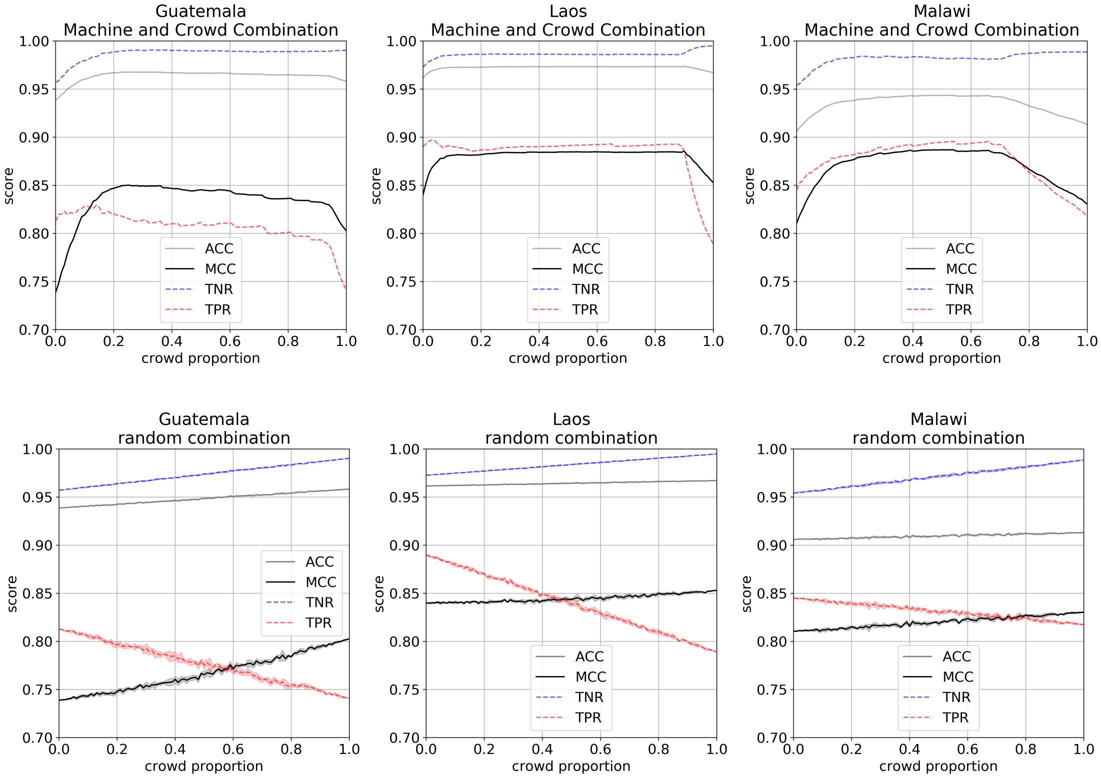

Figure 7. Data quality and effort in respect to crowd proportion and task allocation strategy for a combined MapSwipe-DeepVGI methods. The upper side of the figure represents the performance of the confidence score-based task allocation strategy, whereas on the bottom side the mean and standard deviation of the performance of 250 random allocations between DeepVGI and MapSwipe are shown. Performance is measured by accuracy (ACC), Matthew's correlation coefficient (MCC), sensitivity (TPR) and specificity (TNR). The x-axis of the plot shows the crowd proportion. A crowd proportion of 0.0 (left) refers to DeepVGI. Consequently, the performance for a crowd proportion of 1.0 (right) refers to the performance of MapSwipe. For a crowd proportion of 0.3, 30% of the results are obtained from MapSwipe (these tiles refer to the DeepVGI tiles with the lowest confidence) and 70% from DeepVGI.

6. Discussion

6.1. Overall Performance Evaluation

The crowdsourcing approach by MapSwipe generated the most accurate human settlement maps for all case study sites with Matthew's correlation coefficient between 80% to 85%. As expressed by the differences in specificity TNR and sensitivity TPR between the MapSwipe approach and the DeepVGI approach, our results supported the common hypothesis that humans rarely identify something as a building which is not a building, but tend to miss buildings. In contrast, our results indicate that deep learning approaches tend to miss fewer buildings at the cost of also falsely detecting a range of objects which are not buildings. In comparison to GUF, crowdsourcing and deep learning-based approaches demonstrated an improvement in data quality for our case study sites.

The performance of MapSwipe seemed to be consistent with previous findings from Albuquerque et al. (2016) [9], where an accuracy of 89%, a sensitivity of 73% and a precision of 89% are achieved for a very similar crowdsourced classification task for a study site in South Kivu (Democratic Republic of the Congo). For the case of automated village boundary detection Gueguen et al. (2017) [8] report an average precision of around 70% and a sensitivity of around 84%. Yuan et al. (2018) [30] compare a deep learning-based approach against GUF and GHSL for Kano city (Nigeria) and reach similar results. Their building extraction algorithm performs with a precision of 72% and a sensitivity of 70%. Their approach slightly improves the GUF, but significantly outperforms the Global Human Settlement Layer. Klotz et al. (2016) [7] show that GUF and GHSL significantly increased the completeness and precision of global build-area maps in comparison to previous low-resolution products such as MOD500 or GLOBC. Nevertheless, the authors also point out quantifiable weaknesses in rural areas, which could be confirmed by our study results as well. Whereas the GUF was of moderate quality for the sub-urban Guatemala study site, its weakness for the rural study sites in Laos and Malawi was immense.

Nevertheless, our results also suffered from limitations in our data sets and methods applied. Due to the imbalance of "no building" and "building" tiles, the accuracy ACC reported was biased towards identifying "no building" tiles correctly. This imbalance was strong for Guatemala and Laos, but less pronounced for Malawi. Whereas TNR and TPR show no bias, Matthew's correlation coefficient is biased as well (but not as strong as accuracy) [38]. This reduces the comparability of our results with the findings from other studies with less imbalanced data.

In this study, we did not investigate the effects of the imbalance on the training procedure of the DeepVGI approach. Furthermore, we decided to use a very specific network architecture and pre-trained model (SSD based on COCO data set, see Section 4.1). Our two-step approach (object detection first, then binary classification) also introduced further uncertainties. Whereas our results seemed to be consistent with the findings from other studies, further research is necessary to fully understand the impact of the data preparation on performance. For instance, Tiecke et al. (2017) [6] present a computer vision method to create population maps from satellite imagery with a very high resolution and provide further insights on potential systematic errors. The authors highlight the problem related to repetitive errors, such as the misinterpretation of large rocks, boats, or mountain ridges as buildings. Analyzing the potential sources of systematic errors would be of great benefit for our study to understand spatial clusters of false positives. New advances in machine learning research might produce architectures and training data sets which suit better to the specific use case of mapping human settlements. To reduce uncertainties in our approach, future studies should compare different architectures and training data set characteristics also regarding imbalanced classes.

Uncertainties were also present for the results of the MapSwipe approach. Using majority aggregation to generate binary labels from the individual user classifications favored higher specificity TNR, whereas choosing another method might have promoted higher TPR. The drawbacks of majority agreement are well described by Salk et al. (2016) [43]. The confusion matrix maps showed that wrong classifications were not randomly distributed, but revealed spatial pattern. For MapSwipe this has already been confirmed in Herfort et al. (2017) [15]. For our study, this implies that individual user behavior, geographical characteristics of the surrounding of building features might be major causes of wrong classifications limiting the transferability of this approach. A more detailed analysis is necessary to understand the factors which drive the quality of crowdsourcing and its implications for human settlement maps.

Satellite imagery quality is another major concern for both crowdsourcing and deep learning approaches. Better satellite imagery (e.g., in terms of resolution), might strongly influence the performance of MapSwipe and DeepVGI. Our current study was limited to satellite imagery tiles at zoom level 18. For the regions analyzed in this study, satellite imagery tiles with a higher image resolution at zoom level 19 were not available from Bing Maps. Nevertheless, new earth observation satellites such as WorldView3 would potentially provide sufficient imagery data for this zoom level. The scarce availability of up-to-date satellite imagery, emphasizes another drawback of MapSwipe and DeepVGI: the quality of the human settlement maps is closely tied to the structures visible in the satellite images. In situations where settlement patterns change rapidly, e.g., due to forced displacement, on the ground data is irreplaceable unless up-to-date satellite data becomes available.

6.2. Spatial and Non-Spatial Characteristics of Misclassifications

Our analysis provided insights into the spatial and non-spatial characteristics of misclassified tiles of the DeepVGI method. For all study sites the conditional density plots and logistic regression analyses revealed a significant correlation between confidence score and accuracy. Additionally, the results showed no such correlation for the MapSwipe data set. This indicates that DeepVGI and MapSwipe tended to detect different tiles with different characteristics at varying accuracy. We interpreted this as potential complementary value of both approaches.

However, the limited explanatory power of accuracy for imbalanced data sets needs to be considered for our study (especially for Guatemala and Laos). Hence, the increase in accuracy mainly depicted the correlation between specificity TNR and confidence score. The design of the confidence score (using absolute values, see Section 4.3) hampered the differentiation of false positives and false negatives. Whereas this simplification turned out to be beneficial for the logistic regression analysis, using a more sophisticated method, e.g., relying on a quadratic function, would have reduced the bias introduced due to class imbalance. Furthermore, higher confidence scores might be also related to the number of detected buildings per tile. In our approach, the confidence score was based only on the most probable building detection and did not consider multiple detections per tile. This approach increased uncertainties for tiles, for which only a few buildings were located, e.g., in rural areas.

The results of the logistic regression analysis highlighted that confidence score contributed to the probability of tiles being misclassified, but only to a minor fraction. For the case of land cover mapping with a focus on urban areas Kampffmeyer et al. (2016) [27] provide similar findings and show that pixels with low uncertainty are more likely to be classified correctly. As in our study, areas of class boundaries were a cause of wrong classifications. However, our study showed as well that most misclassified tiles had a different cause not captured by our design. Our methods fell short especially in differentiating the confidence of "no building" classifications, which however constituted most tasks in our study areas.

Further research is needed to expand our understanding of the systematic errors underlying our approach. For example, we would investigate to what degree deep learning-based methods are able to map various building types, e.g., in relation to shape and size or to characteristics related to ethnic or social groups.

6.3. Combination of Crowdsourcing and Deep Learning

Combining the MapSwipe and DeepVGI methods using the confidence score-based task allocation strategy increased performance by around 3 - 5 percentage points measured by MCC (compared to the MapSwipe-only approach). At the same time, the approach reduced the volunteer efforts to one fifth (labor reduction of 80 percentage points). Our results suggest that the task allocation strategy helped to exploit the complementary value of a sensitive method (DeepVGI) and a specific method (MapSwipe) and would improve the existing crowdsourcing approach MapSwipe uses.

For a similar set up but limited geographic scope Chen et al. (2018) [31] show that a combination of results from machine learning and crowdsourcing can result in a labor reduction of 85 percentage points and achieves a similar accuracy. Gueguen et al. (2017) [8] report similar findings regarding semi-supervised village boundary mapping and are able to improve the precision of automated data extraction by introducing a crowdsourced validation step.

The combination of crowdsourcing and deep learning showed promising results for our selected case studies, but limitations of the presented workflow must be considered. First, the individual performance of both methods might have a great impact on the performance of the combined approach. Our approach was able to improve results because of the complementary characteristics MapSwipe (high TNR) and DeepVGI (high TPR) hold. Due to the uncertainties of crowdsourcing and deep learning (discussed in Section 6.1) these differences between the two methods might be less pronounced or even reversed in other regions.

Considering the needs of humanitarian organizations (e.g., getting information on human settlements for which no other data sets exist) our approach could be used in real applications. However, the current workflow was not able to provide a clear estimation which crowd proportion would generate optimal results beforehand. Project managers organizing humanitarian mapping campaigns would still need to adjust the proportion of tasks mapped by the crowd manually, e.g., in respect to the given time frame and complexity of the mapping task.

We tested the workflow for three rather diverse study sites, nevertheless a more detailed investigation of the influence of geographic characteristics is necessary. Quantifying the differences between the study sites (e.g., in respect to land cover) would be a first step towards contextualizing the results. Together with an analysis of the quality of the satellite imagery and quality of the crowdsourced classifications this would help to understand for which study site characteristics a combination might result in better data.

7. Conclusions

Human settlement maps produced by crowdsourcing (MapSwipe) or deep learning (DeepVGI) showed large overlaps and for most areas both methods generated results with a similar accuracy. In general, both methods outperformed existing EO-based products such as the Global Urban Footprint in terms of MCC and TPR. The proposed confidence score indicator helped to explain misclassified tiles of the DeepVGI method and revealed the complementary value of DeepVGI and MapSwipe. Combining crowdsourcing and deep learning by applying the proposed task allocation strategy facilitated the complementary values of both methods and provided a promising extension to the existing crowdsourcing approach MapSwipe incorporates.

Further research needs to validate these findings also for other study regions and various settlement types and shapes and contextual features such as vegetation and land use. The large amount of finished MapSwipe projects provides an obvious starting point for such an extended geographical analysis.

Our study focused on the binary classification of satellite imagery tiles to map human settlements. Due to the structure of the MapSwipe results, the analysis was limited to the tile level. Future research should overcome this drawback and investigate human settlement classifications at a more fine-grained resolution and/or move on to investigating automatically generated building footprint geometries. Initial research in this direction has been conducted by Vargas-Munoz et al. (2019) [32]; however national and regional investigations are necessary.

Reaching the targets of the sustainable development, planning disaster responses more efficiently and reducing the vulnerability of people at risk before disasters occur will remain challenges for the upcoming decades. The proposed combined use of satellite data, deep learning technology and citizen-based observations showed great potential to contribute to those efforts and future applications should consider the lessons learned from this research. However, to take fully advantage of the new opportunities there is a need to further understand the technical and non-technical challenges that come with them. The presented approach might help to identify data quality issues during or immediately after an object has been mapped to provide real-time or near real-time feedback for mappers. However, understanding and communicating the quality of automatically generated results and ensuring that tools and data are open and accessible are the very basis for this.

Integrating machine learning techniques into existing crowdsourcing applications will also create an increased need for technical knowledge for project managers and data users. From the citizen science and VGI projects perspective, introducing these new techniques might also lead to an increased demand for experienced validators, which are already few in number, presently. Whereas we envision to use these new tools to improve overall data quality and availability, they also constitute a new potential source of bias introduced into data sets such as OpenStreetMap. This bias might also be caused by class imbalances, which were present also in our study.

Taking all this into account, our results endorsed that for the creation of human settlement maps, we should rely on automated approaches (e.g., machine learning) when possible, but rely on human skills (e.g., citizens science and crowdsourcing) when needed.

Author Contributions: Conceptualization, B.H. and H.L.; Methodology, B.H., H.L. and S.F.; Validation, B.H., H.L. and S.F.; Formal Analysis, B.H.; Writing—Original Draft Preparation, B.H. and H.L.; Writing—Review and Editing, B.H., S.F. and S.L.; Visualization, B.H.; Supervision, A.Z.; Project Administration, S.F. and A.Z.

Funding: The authors acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding programme Open Access Publishing, by the Baden-Württemberg Ministry of Science, Research and the Arts and by Ruprecht-Karls-Universität Heidelberg. This research was partly funded by the Klaus Tschira Stiftung, Heidelberg.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. United Nations Department of Economic and Social Affairs Population Division. *World Urbanization Prospects: The 2018 Revision*; Technical Report; United Nations Department of Economic and Social Affairs Population Division: New York, NY, USA, 2018.
2. United Nations. Transforming Our World: The 2030 Agenda for sUstainable Development. Technical Report. 2015. Available online: http://xxx.lanl.gov/abs/arXiv:1011.1669v3 (accessed on 31 July 2019).
3. United Nations Office for Disaster Risk Reduction. *Sendai Framework for Disaster Risk Reduction 2015–2030*; Technical Report; United Nations Office for Disaster Risk Reduction: Geneva, Switzerland, 2015.
4. Pesaresi, M.; Huadong, G.; Blaes, X.; Ehrlich, D.; Ferri, S.; Gueguen, L.; Halkia, M.; Kauffmann, M.; Kemper, T.; Lu, L.; et al. A global human settlement layer from optical HR/VHR RS data: Concept and first results. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 2102–2131. [CrossRef]
5. Esch, T.; Marconcini, M.; Felbier, A.; Roth, A.; Heldens, W.; Huber, M.; Schwinger, M.; Taubenbock, H.; Muller, A.; Dech, S. Urban footprint processor-Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 1617–1621. [CrossRef]
6. Tiecke, T.G.; Liu, X.; Zhang, A.; Gros, A.; Li, N.; Yetman, G.; Kilic, T.; Murray, S.; Blankespoor, B.; Prydz, E.B.; et al. Mapping the world Population One Building at a Time. *arXiv* **2017**, arXiv:1712.05839.
7. Klotz, M.; Kemper, T.; Geiß, C.; Esch, T.; Taubenbóck, H. How good is the map? A multi-scale cross-comparison framework for global settlement layers: Evidence from Central Europe. *Remote Sens. Environ.* **2016**, *178*, 191–212. [CrossRef]
8. Gueguen, L.; Koenig, J.; Reeder, C.; Barksdale, T.; Saints, J.; Stamatiou, K.; Collins, J.; Johnston, C. Mapping Human Settlements and Population at Country Scale from VHR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 524–538. [CrossRef]
9. Albuquerque, J.; Herfort, B.; Eckle, M. The Tasks of the Crowd: A Typology of Tasks in Geographic Information Crowdsourcing and a Case Study in Humanitarian Mapping. *Remote Sens.* **2016**, *8*, 859. [CrossRef]
10. Hachmann, S.; Jokar Arsanjani, J.; Vaz, E. Spatial data for slum upgrading: Volunteered Geographic Information and the role of citizen science. *Habitat Int.* **2017**. [CrossRef]

11. Scholz, S.; Knight, P.; Eckle, M.; Marx, S.; Zipf, A. Volunteered Geographic Information for Disaster Risk Reduction—The Missing Maps Approach and Its Potential within the Red Cross and Red Crescent Movement. *Remote Sens.* **2018**, *10*, 1239. [CrossRef]
12. See, L.; Mooney, P.; Foody, G.; Bastin, L.; Comber, A.; Estima, J.; Fritz, S.; Kerle, N.; Jiang, B.; Laakso, M.; et al. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 55. [CrossRef]
13. Hecht, R.; Kunze, C.; Hahmann, S. Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 1066–1091. [CrossRef]
14. Chen, J.; Zipf, A. DeepVGI: Deep Learning with Volunteered Geographic Information. In Proceedings of the WWW '17 Companion: Proceedings of the 26th International Conference Companion on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 771–772. [CrossRef]
15. Herfort, B.; Reinmuth, M.; de Albuquerque, J.P.; Zipf, A. Towards evaluating crowdsourced image classification on mobile devices to generate geographic information about human settlements. In Proceedings of the 20th AGILE, Wageningen, The Netherlands, 9–12 May 2017.
16. Esch, T.; Heldens, W.; Hirner, A.; Keil, M.; Marconcini, M.; Roth, A.; Zeidler, J.; Dech, S.; Strano, E. Breaking new ground in mapping human settlements from space – The Global Urban Footprint. *ISPRS J. Photogramm. Remote Sens.* **2017**, *134*, 30–42. [CrossRef]
17. Zook, M.; Graham, M.; Shelton, T.; Gorman, S. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Med. Health Policy* **2010**, *2*, 6–32. [CrossRef]
18. Degrossi, L.C. Potential of Collaborative Mapping for Disaster Relief : A Case Study of OpenStreetMap in the Nepal Earthquake 2015. In Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, USA, 5–8 January 2016; doi:10.1109/HICSS.2016.31. [CrossRef]
19. Crooks, A.; Pfoser, D.; Jenkins, A.; Croitoru, A.; Smith, D.; Karagiorgou, S.; Efentakis, A.; Crooks, A.; Pfoser, D.; Jenkins, A.; et al. Crowdsourcing urban form and function. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 720–741. [CrossRef]
20. Goodchild, M.F.; Li, L. Assuring the quality of volunteered geographic information. *Spat. Stat.* **2012**, *1*, 110–120. [CrossRef]
21. Ballatore, A.; Zipf, A. A Conceptual Quality Framework for Volunteered Geographic Information. In *Proceedings of the Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, 12–16 October 2015*; Fabrikant, S.I., Raubal, M., Bertolotto, M., Davies, C., Freundschuh, S., Bell, S., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 89–107._5. [CrossRef]
22. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719. [CrossRef]
23. Comber, A.; Mooney, P.; Purves, R.S.; Rocchini, D.; Walz, A. Crowdsourcing: It matters who the crowd are. The impacts of between group variations in recording land cover. *PLoS ONE* **2016**, *11*. [CrossRef]
24. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
25. Jean, N.; Burke, M.; Xie, M.; Davis, W.M.; Lobell, D.B.; Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science* **2016**. [CrossRef]
26. Li, H.; Herfort, B.; Zipf, A. Estimating OpenStreetMap Missing Built-up Areas using Pre-trained Deep Neural Networks. In Proceedings of the 22nd AGILE, At Limassol, Cyprus, 17–20 June 2019.
27. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 680–688. [CrossRef]
28. Tracewski, L.; Bastin, L.; Fonte, C.C. Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization. *Geo-Spat. Inf. Sci.* **2017**, *20*, 252–268. [CrossRef]
29. Vakalopoulou, M.; Karantzalos, K.; Komodakis, N.; Paragios, N. Building detection in very high resolution multispectral data with deep learning features. *Int. Geosci. Remote Sens. Symp. (IGARSS)* **2015**, *2015*, 1873–1876. [CrossRef]

30. Yuan, J.; Roy Chowdhury, P.K.; McKee, J.; Yang, H.L.; Weaver, J.; Bhaduri, B. Exploiting deep learning and volunteered geographic information for mapping buildings in Kano, Nigeria. *Sci. Data* **2018**, *5*, 180217. [CrossRef] [PubMed]
31. Chen, J.; Zhou, Y.; Zipf, A.; Fan, H. Deep Learning From Multiple Crowds: A Case Study of Humanitarian Mapping. *IEEE Trans. Geosci. Remote Sens.* **2018**, doi:10.1109/TGRS.2018.2868748. [CrossRef]
32. Vargas-Muñoz, J.E.; Lobry, S.; Falcão, A.X.; Tuia, D. Correcting rural building annotations in OpenStreetMap using convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *147*, 283–293. [CrossRef]
33. Maso, J.; Pomakis, K.; Julia, N. *OpenGIS Web Map Tile Service Implementation Standard*; Open Geospatial Consortium Inc.: Wayland, MA, USA, 2010; pp. 4–6.
34. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Basel, Switzerland, 2016; Volume 9905 LNCS, pp. 21–37.
35. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: tensorflow.org (accessed on 31 July 2019).
36. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 740–755. doi:10.1007/978-3-319-10602-1_48. [CrossRef]
37. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE CVPR, Honolulu, HI, USA, 21–26 July 2017 ; Volume 4.
38. Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [CrossRef]
39. Jones, E.; Oliphant, T.; Peterson, P. *SciPy: Open Source Scientific Tools for Python*. Available online: https://scholar.google.com/scholar?cluster=2086009121748039507&hl=en&oi=scholarr (accessed on 31 July 2019).
40. Scott, D.W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
41. McFadden, D. *Conditional Logit Analysis of Qualitative Choice Behavior*. Available online: https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf (accessed on 31 July 2019).
42. Seabold, S.; Perktold, J. Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28–30 June 2010.
43. Salk, C.F.; Sturn, T.; See, L.; Fritz, S. Limitations of Majority Agreement in Crowdsourced Image Interpretation. *Trans. GIS* **2016**. [CrossRef]