Analyzing the Impact of Grey Data Quality in the Training Phase of Deep-Learning Models for Dwelling Extraction in Refugee Camps

Vanessa Streifeneder

People flee all over the world due to hazards, humanitarian disasters, and conflicts. They search for safety in refugee camps and finding their shelter sometimes for more than one generation. Managing and monitoring these camps is not simple, especially in remote areas. Remote sensing methods and geoinformation systems combined with deep learning (DL) offer an alternative for estimating camp population, structure, and dynamics. A major problem with DL, like convolutional neural networks (CNN), is the need for a high number of samples. Predefined samples for tent classification in the context of refugee camps are rarer and must be manually created. The idea is hence to use sample data from various sources. For this thesis, the data was provided by the gEOhum project in cooperation with Médecins Sans Frontiéres. However, this data includes semantic and geometrical noises. Figure 1 shows some of these noises including incorrect classification, geometrical shifts, incomplete delineation as well as image quality issues.

Figure 1: Preexisting dwelling samples and their associated issues

Due to the uncertainty of the quality and the partly unknown metadata this data is called 'grey-data'. In the case of utilizing grey data from refugee camps, manual evaluation, as well as semi-automated correcting of the data, is needed. In an operational situation, when time is limited, faster solutions must be implemented to identify and reject incorrect samples. The overall goal of this master thesis was hence to explore the impact of so-called ‘grey data’ in the training phase of deep learning models and implement a cyclic learning rate (CLR) to reduce noise in the sample data.

For this study, the Minawao Refugee camp in the northern part of Cameroon was chosen. Based on a World View 2 satellite image a semantic segmentation was conducted using the CNN U-Net model. To analyze the impact of different types of noise and the model robustness the training samples were artificially manipulated using GIS (Figure 2).

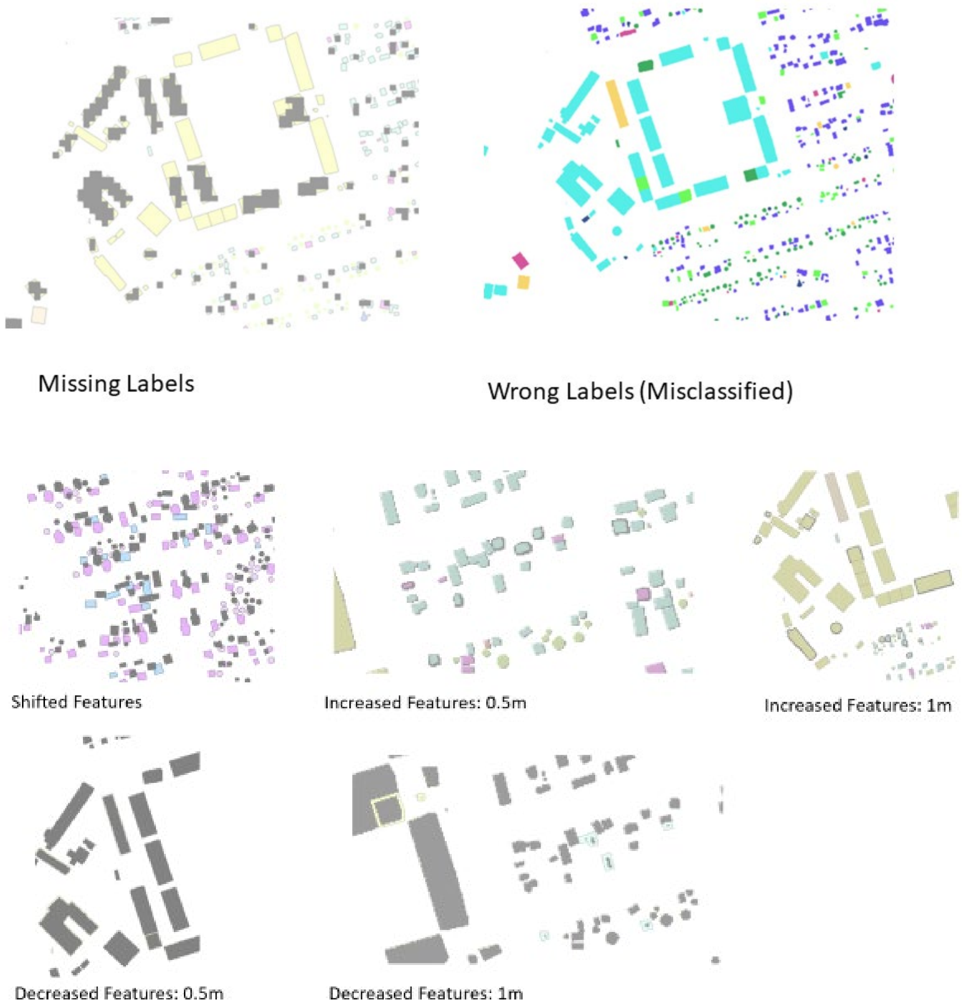

Figure 2: Examples of the label mask for semantic and geometrical feature manipulation (grey features = manipulated, colored features= original)

Figure 3 exemplary presents the results for the model runs with different amounts of unlabeled data in the training samples. The findings show the more data was missing the lower the model performance.

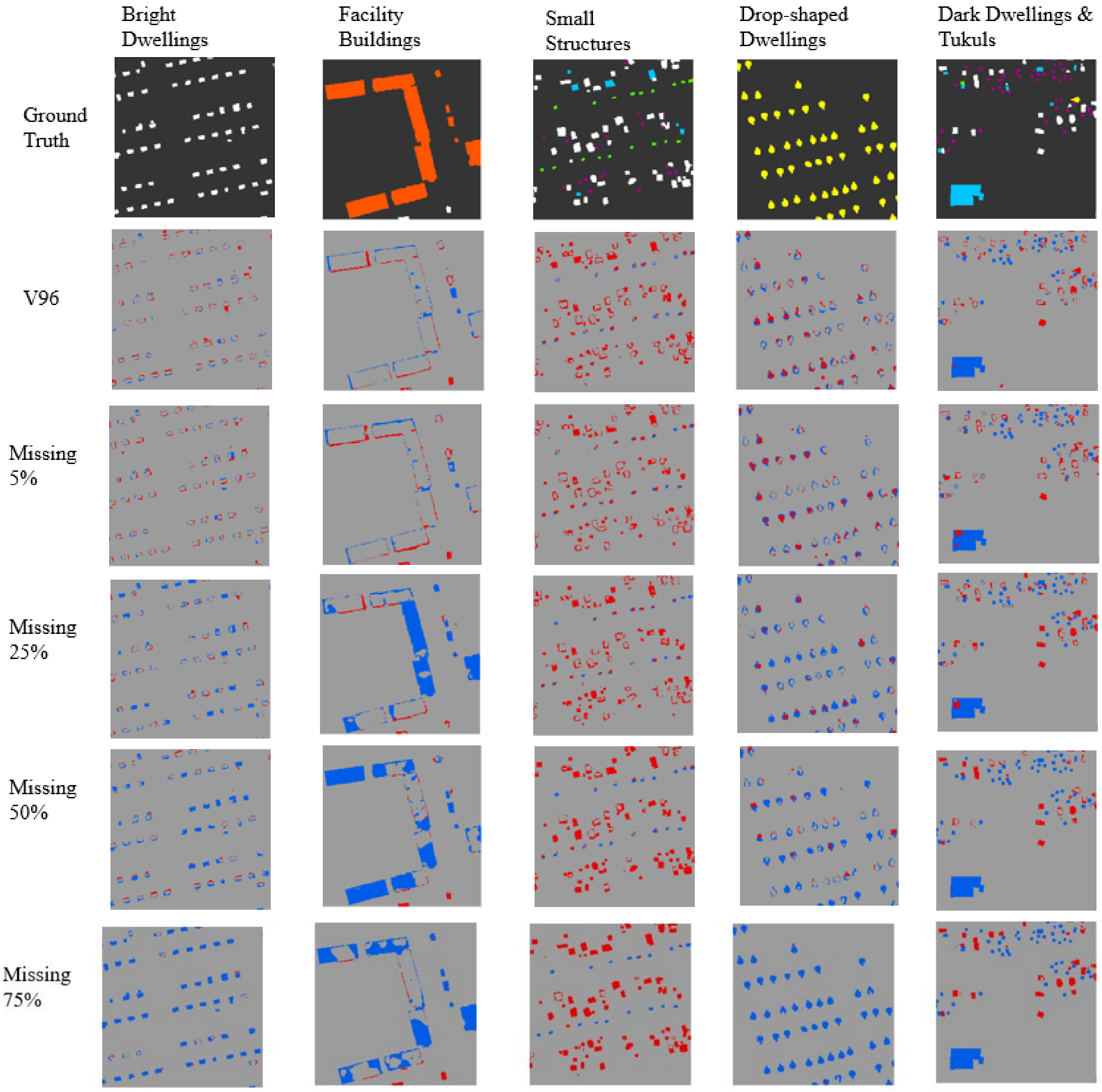

Figure 3: Predicted labels for a subset of data for the different dwelling classes (Bright Dwellings: White; Facility Buildings: Orange; Small Structures: Green; Drop-shaped Dwellings: Yellow; Dark Dwellings& Tukuls: Purple) showing TP: grey, FN pixels: Blue and FP: Red for the selected dwelling class

In the final step to delete noisy samples and improve the model performance, a CLR was implemented. This method detects noisy samples based on the higher loss value generated by the model when learning from such a sample.

Overall, the results showed that depending on the amount of noise and the kind of noise, grey data could even improve the model performance by providing more samples for certain dwelling classes. Still, especially annotation errors decreased the model accuracy drastically. Although, a slight enlarging of the dwelling delineation led to better detection of dwellings characterized by a lower contrast to the ground. Implementing a CLR could improve the model performance up to 20% for certain dwelling classes, but only if the percentage of noisy samples were not higher than 50%. All in all, implementing a CLR to detect and remove noisy samples enhanced the model performance, but the method still has to be improved to decrease for instance the processing time.