Appendix A. Supplementary Material

A.1. Training parameters

For **road segmentation**, our Dense-U-Net-121 was trained using an NVIDIA RTX Titan GPU, with ImageNet pre-training for the encoder, a cross-entropy loss, an ADAM optimizer, an initial learning rate of $1e-4$ and an exponential learning rate decay of 0.8 applied after each epoch. We applied random horizontal flips and 90° rotations. A quantile truncation and a normalization were applied on each input channel separately to remove the upper and lower 2% of outliers.

For **building segmentation**, ImageNet weights were used to initialize the model. The training was performed on four NVIDIA RTX Titan GPUs using a cross entropy loss with online hard example mining [14], stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01, weight decay of 0.001, and Nestrov momentum of 0.9. The learning rate was decayed and the training patches are randomly flipped and rescaled. The image processing is the same as for the road segmentation.

For training the **person detection** network we use a batch size of 50 and adopted a learning rate of $5e{-}7$, using the scheduling mechanism from [12]. Adam optimizer with a decay of $5e{-}4$ and a pre-trained model from the MS COCO dataset [10] was used. We performed a statistical analysis for selecting the anchor boxes. To calculate the loss, we used a combination of logistic regression for objectness error and complete IoU [18] for bounding box error, similar to [12].

A.2. Details about the test data for road and building segmentation

Besides the MACS camera system [6], image data from various systems were used. For these scenes, the data was acquired by the 4K camera system [8] and from Germany's Federal Agency for Cartography and Geodesy (Digital Orthophotos, DOP20). All test images are resampled to 50 cm/pixel for the road segmentation model and to 20 cm/pixel for the building segmentation model. In addition, in order to test the generalizability of the models in different regions of operation, particularly in disaster-prone developing countries, the trained models are tested in Beira, Mozambique, and Kathmandu, Nepal, following cyclones and earthquakes. The UAV image of Beira was provided by WFP and Mozambique's National Institute for Disaster Management (INGC), and the aerial imagery of Kathmandu was captured by the MACS system.

A.3. Detailed results for road segmentation

We tested the road segmentation method on 21 test areas from Epeisses in Switzerland, from the Ahr Vally in Germany, from Beira in Mozambique, and from Kathmandu in Nepal. For all scenes except Kathmandu, we annotated the images by hand following the centerline of the roads and saved them as vector graphics. Figure 6, Figure 7, Figure 8 and Figure 9 show the predictions of our model and a dilated version of our ground truth if available. Note that although our model outputs pixel-wise segmentation, our metrics do not compare them pixel to pixel to the labels as shown in the figures. Rather, they are evaluated on a topological basis after the predictions have been thinned to a 1-pixel thickness, equivalent to vectorizing them into centerlines.

In the Epeisses scene (see Figure 6), most roads were correctly identified and accurately extracted, except for some sections located close to the edges of the mosaic. This is due to the lack of context given to the model, which expects the roads to be continuous as it has not been trained to overcome sudden disruptions by the background areas in the images.

In the Ahr Valley scene (see Figure 7), the model managed to detect most roads in both the pre- and post-disaster images, though we could not report results as no ground truth was available for this area yet. In the pre-disaster image, our model shows its capacity for generalizing well to sub-urban scene types unseen during its training, as it was only given to see regions from Southeast Asia. In the post-disaster image, it has shown some confusion as to which road to consider as still intact: there is in fact much water, mud, and debris on the surface of the roads, making it more difficult to draw a line between damaged and usable road sections.

In the Beira scene (see Figure 8), while the images are particularly challenging due to the presence of unpaved roads or streets covered in sand in the aftermath of Cyclone Idai, the model still manages to extract all the roads except a few narrower ones. However, it did detect roads that the annotators did not include in the labels due to occlusion or the lack of clues as to their usability by vehicles. This begs the question of the annotation policy and the boundaries between road and non-road objects in difficult scenarios where they might either be not visible or require local knowledge for a specific region.

In the Kathmandu scene (see Figure 9), the model was faced with a complex urban infrastructure, featuring many narrow, irregular, and therefore occluded streets, and often unpaved roads. Nevertheless, it was capable of extracting the vast majority of the roads with great accuracy, from large arteries to small alleyways, even though the connectivity of the mask may be improved in locations where the road segments are kept apart by the occlusion of buildings. In such scenarios, it actually becomes challenging to define a fair, comprehensive road annotation policy, as expert on-the-ground knowledge is required to define the difference between a road and a simple large- drivable area not dedicated to vehicles.

A.4. Detailed results for building segmentation

The building segmentation method is tested on 15 areas, including those mentioned in the main text plus one area in Kathmandu, Nepal (same as for the road segmentation).

In the Epeisses scene (see Figure 10), most of the buildings are extracted and only two buildings are omitted. On the other side, large tents are mistakenly segmented due to their similarity to real buildings. In addition, damaged and collapsed buildings are classified as buildings but are not labeled in the annotation. Overall, an F1 and IoU score of 47.72% and 31.33% are acquired. **In the Ahrtal scene** (see Figure 11), we selected 10 regions (10.5km^2) and manually annotated the building ground truth of the pre-flood images. In the 10 annotated regions we achieved 86.66% and 76.46% for building F1 and IoU scores respectively. Figure 11 illustrates three small regions with pre- and post-flood images. All pre-event images are DOP20, the post-event image (b) is captured by the MACS camera system, and (d) and (f) illustrate images captured by the 4k system. Due to the difference in flight altitude and viewing angles between the training and test data, the network was unable to detect a couple of damaged buildings from a very oblique view (see Figure 11 (d)), but still managed to accurately extract most of the buildings.

Within the Beira scenes (refer to Figure 12), the model successfully identifies larger buildings, but fails to detect the majority of smaller structures as shown in Figure 12 (i), which is evidenced by the precision score of 76.51% and recall score of 44.92%. This observation highlights the need to address domain shifts between the training and test datasets. In particular, when the building characteristics differ between the training and test areas, as in the case of Beira, factors such as different size distributions and different roof materials contribute to the observed drop in performance. Furthermore, it is also crucial to properly account for the varying imaging conditions. The Beira test data was acquired using a low-altitude UAV, resulting in a centimeter-level GSD that has unique spectral features even after downsampling. The black lines in Figure 12 (a) resulted from co-registration with the pre-event imagery (not shown). Due to memory constraints, the UAV images are cut into smaller patches, resulting in unaligned image boundaries.

In contrast to the outcomes observed in Beira, the visual outcomes achieved in **Kathmandu** (see Figure 13) appear promising, with the majority of buildings being successfully identified despite their dissimilarity to the training data. As ground truth data is unavailable for this scene, our analysis is conducted exclusively through qualitative evaluation. We study three different urban zones characterized by different building types and densities. Figure 13 (a) depicts a typical densely populated urban area with small-scale residences. Despite significant differences in building characteristics such as roof materials and density, a visual assessment shows a similar level of performance to that achieved in the European test areas. Similar results are found in the samples of Figure 13 (c) and Figure 13 (e). The success observed in the Kathmandu results underlines the robust generalizability of the method to MACS images despite the regional differences.

A.5. Detailed results for person detection

In Figure 14, we present selected examples from our annotated training set, where each person is individually annotated with a bounding box. Additionally, we present the results of our person detection algorithm applied to image mosaics in Figure 15 for the near real time scenario and to single images for onboard processing in Figure 16. In the figures, we show zoomed areas that demonstrate both successes and failures. For example, in Figure 15 (d), the very high resolution of the image confuses the model, leading to false detections of small objects such as stones, which can resemble the appearance of people in images with larger GSDs.

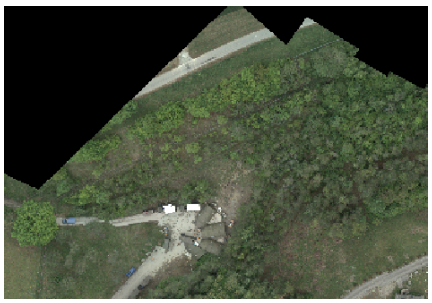
Details of the training dataset for person detection: The dataset for person detection contains 311 annotated aerial and drone images acquired between 2012 and 2022 over different regions in Germany, the Netherlands, Switzerland, Spain, France, and Nepal. The sizes of the images vary between 4864×3232 px, 5184×3456 px, 5616×3744 px, and 8000×6000 px. Care was taken during the image selection process to guarantee that images with different GSD, cloud cover, and acquired with different weather conditions, sun positions, viewing angles, seasons, times of day, types of scene (urban, suburban, rural, park, and recreation sites), and application scenarios (rescue, crowd events, construction) were included in the dataset. We divided the 311 images of our dataset into three disjoint sets: 1) the training set consisting of 259 images with 6934 annotations, 2) the validation set consisting of 25 images with 2706 annotations, and 3) the test set consisting of 27 images with 410 annotations. The samples in Figure 14 illustrate the diversity of the images within the dataset.

(a) The complete scene in Epeisses with overlaid predictions

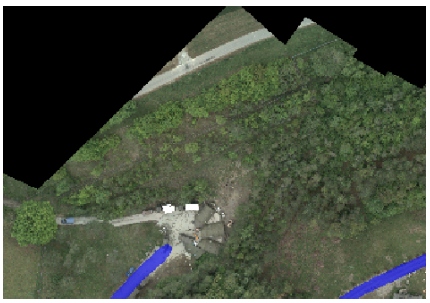
(b) success case: zoom-in view of the image

(c) success case: zoom-in view of the predictions

(d) success case: zoom-in view of the ground truth



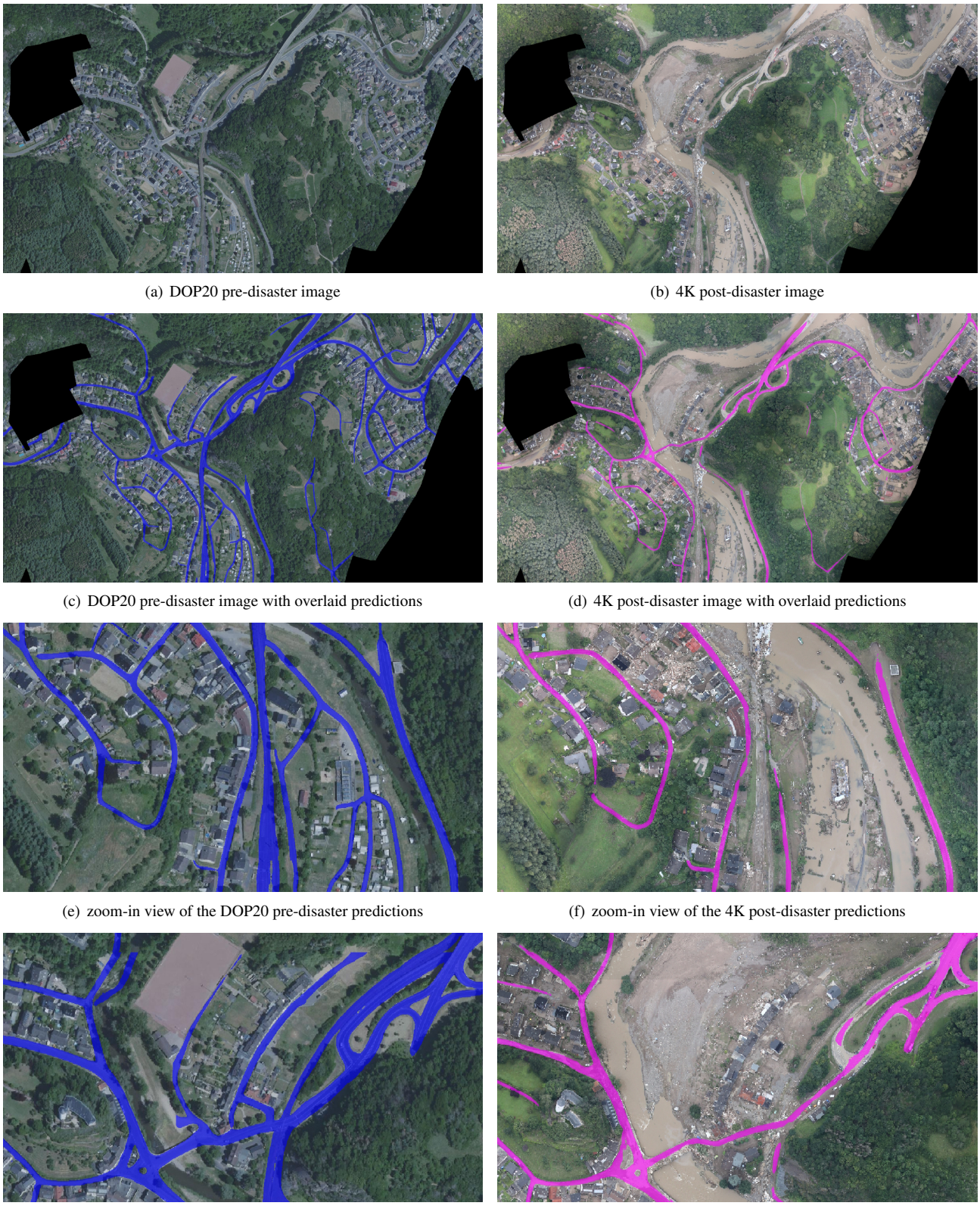
(e) failure case: zoom-in view of the image



(f) failure case: zoom-in view of the predictions

(g) failure case: zoom-in view of the ground truth

Figure 6: Road segmentation results for the test area in Epeisses, Switzerland.



(a) DOP20 pre-disaster image

(b) 4K post-disaster image

(c) DOP20 pre-disaster image with overlaid predictions

(d) 4K post-disaster image with overlaid predictions

(e) zoom-in view of the DOP20 pre-disaster predictions

(f) zoom-in view of the 4K post-disaster predictions

(g) zoom-in view of the DOP20 pre-disaster predictions

(h) zoom-in view of the 4K post-disaster predictions

Figure 7: Road segmentation results for the test areas in the Ahr Valley, Germany.

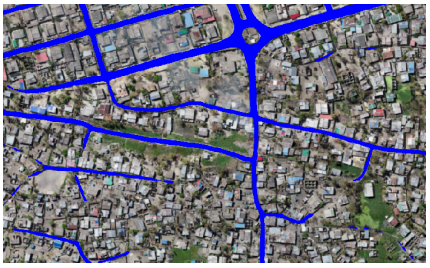
(a) A selected scene from Beira with overlaid predictions

(b) success case: zoom-in view of the image

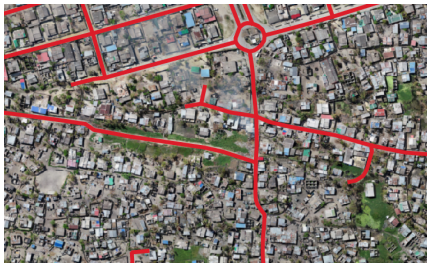
(c) success case: zoom-in view of the predictions

(d) success case: zoom-in view of the ground truth

(e) failure case: zoom-in view of the image



(f) failure case: zoom-in view of the predictions



(g) failure case: zoom-in view of the ground truth

Figure 8: Road segmentation results for a scene of the test areas in Beira, Mozambique.

Figure 9: Road segmentation results for a scene from Kathmandu in Nepal captured from a UAV at 8 cm/px and resampled to 50 cm/px.

(a) The complete scene in Epeisses with overlaid predictions

(b) failure case: zoom-in view of the image

(c) failure case: zoom-in view of the predictions

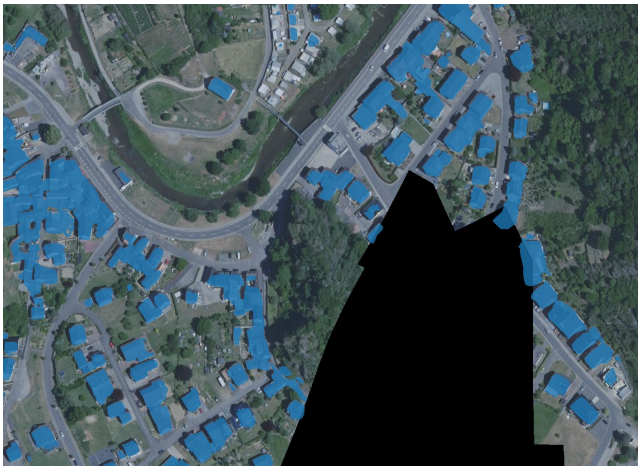
(d) failure case: zoom-in view of the ground truth

(e) failure case: zoom-in view of the image

(f) failure case: zoom-in view of the predictions

(g) failure case: zoom-in view of the ground truth

Figure 10: Building segmentation results for the test area in Epeisses, Switzerland.

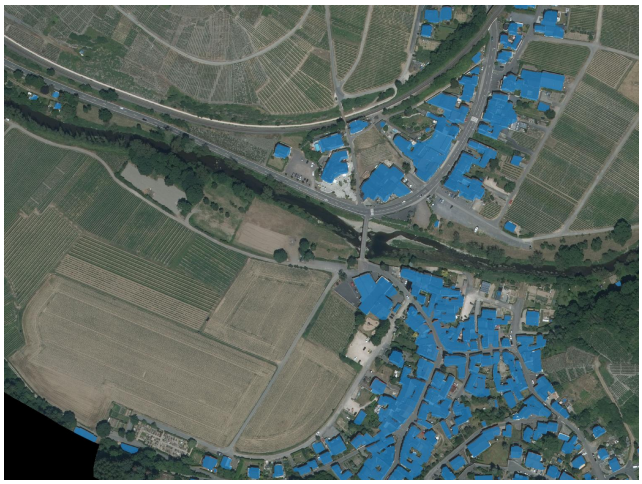


(a) DOP20 pre-disaster image with overlaid predictions

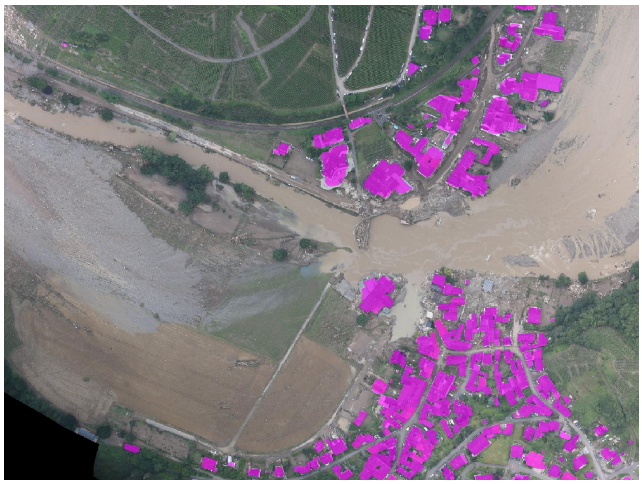
(b) MACS post-disaster image with overlaid predictions

(c) DOP20 pre-disaster image with overlaid predictions

(d) 4K post-disaster image with overlaid predictions

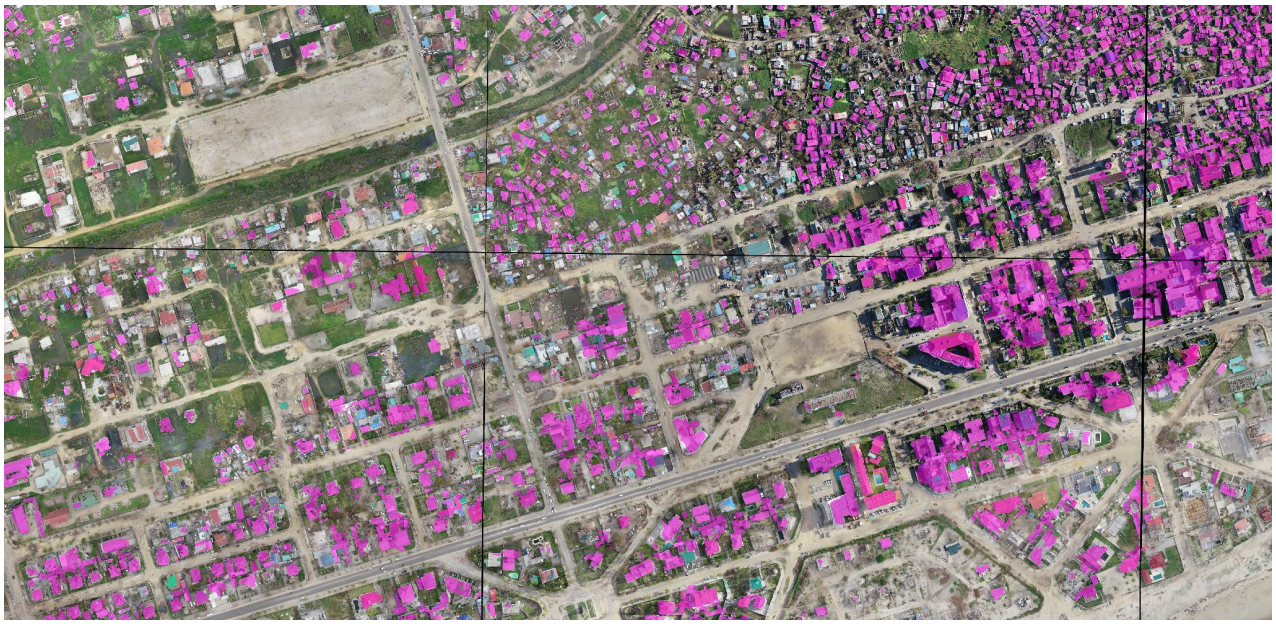


(e) DOP20 pre-disaster image with overlaid predictions



(f) 4K post-disaster image with overlaid predictions

Figure 11: Building segmentation results of the selected areas in the Ahr Valley, Germany.



(a) A larger scene selected from Beira, Mozambique with overlaid predictions

(b) zoom-in view of the image

(c) zoom-in view of the predictions

(d) zoom-in view of the ground truth

(e) zoom-in view of the image

(f) zoom-in view of the predictions

(g) zoom-in view of the ground truth

(h) failure case: zoom-in view of the image

(i) failure case: zoom-in view of the predictions

(j) failure case: zoom-in view of the ground truth

Figure 12: Building segmentation result visualization of the test area in Beira, Mozambique.

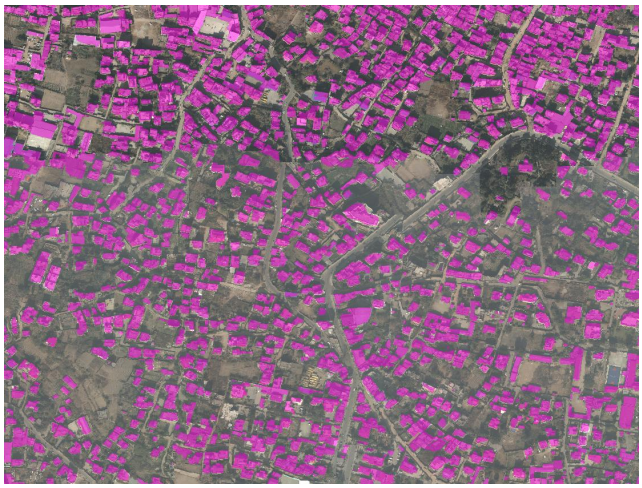
(a) Zoom-in view of the image in dense urban area

(b) Zoom-in view with the prediction in dense urban area

(c) Zoom-in view of the image in urban area with public park

(d) Zoom-in view with the prediction in urban area with public park

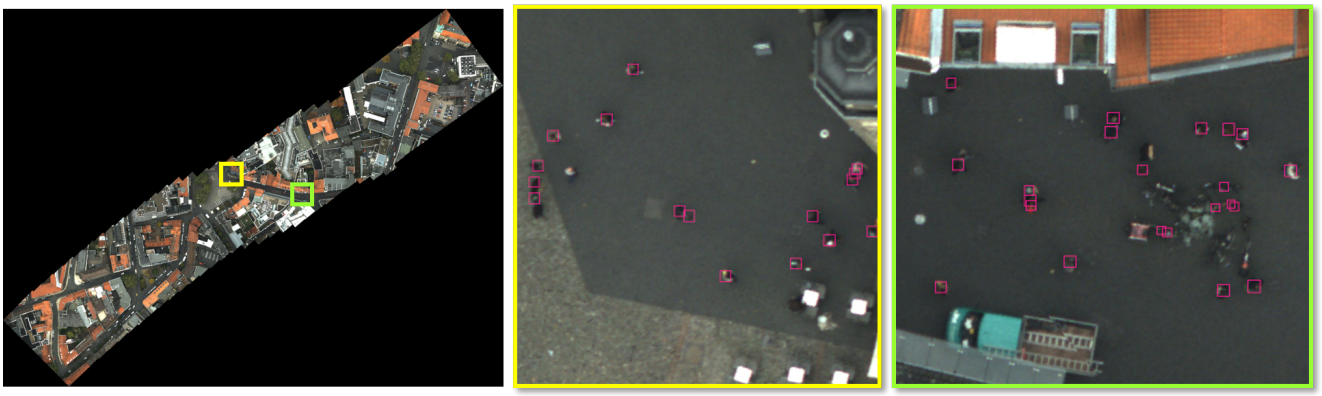
(e) Zoom-in view of the image in urban area with more detached houses



(f) Zoom-in view with the prediction in urban area with more detached houses

Figure 13: Building segmentation result visualization of the test area in Kathmandu, Nepal. Three different urban areas are selected for demonstration.

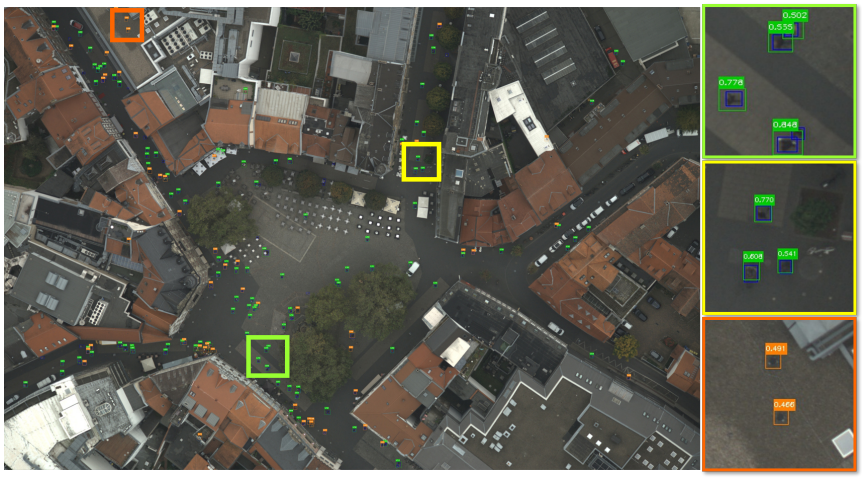
Figure 14: Samples of the person detection dataset. Each person is annotated with an individual bounding box. This figure shows image samples from the training set.



(a) Downtown Brunswick, Germany. GSD = 3 cm, coverage = 0.1 km^2, process time = 99s.

(b) Epeisses, Switzerland. GSD = 3 cm, coverage = 0.08 km^2, process time = 47s.

Figure 15: Person detection on two example image mosaics processed in near real time.



(a) Downtown Brunswick, Germany. GSD = 4.2 cm, AP = 67%.

(b) Flood ruins Stolberg, Germany. GSD = 1 cm, AP = 46%.

(c) Villejust, France. GSD = 0.7 cm, AP = 79%.

Figure 16: Sample results from our test set for the onboard person detection with high and low confidence predictions marked in green and orange, respectively. The ground truth annotations are represented by blue bounding boxes.