Boundary-aware 3D Building Reconstruction from a Single Overhead Image: Supplementary

Jisan Mahmud True Price Akash Bapat Jan-Michael Frahm
University of North Carolina at Chapel Hill
{jisan,jtprice,akash,jmf}@cs.unc.edu

Figure 1: From left: Overlapping proposals from Task 1, predicted BPSH, building masks after NMS, additionally recovered building proposals, and final building masks. The overlapping proposals are the purple and teal buildings in the top left, where the teal proposal actually covers two buildings, and the purple proposal overlaps the part of the teal that covers the smaller building.

In this supplementary material, we present additional details about the network, optimization, and building overlap refinement strategies of our method. We also present additional qualitative results and ablative analysis.

1. Shared feature representation between T2, T3, T4

Our proposed multi-task learning framework for Tasks 2, 3 and 4 utilizes a common feature representation, as discussed in Sections 3.1.2, 3.3, and 3.4 of the main paper. From the feature layer $P2$ of the Feature Pyramid Network (FPN), we perform two (3×3) convolutions, followed by a (1×1) convolution and a skip connection to obtain this shared feature representation. We obtain the outputs of Task 2, 3, and 4 by using this shared representation, and by performing lightweight task-specific convolutions. This construct is illustrated in Fig. 2 of the main paper.

2. Optimization for multi-task learning

We initialize parts of our network from pre-trained weights. The FPN and the region proposal network are initialized using Mask R-CNN trained on the SpaceNet dataset [3]. Since many overhead datasets have multi-spectral images (for example, 8-channel images instead of typical 3-channel RGB) we initialize the parameters of the first layer using He initialization [6]. All of the newly added layers are also initialized with this method. All batch normalization layers are frozen during the training. Random flipping, cropping, and rotation are used as data augmentation techniques. We optimize the loss using stochastic gradient descent with an initial learning rate of 2×10^{-3}, momentum of 0.9, weight decay of 10^{-4}, and batch size of 6 images of (512×512) px. Loss gradients are clipped to $[-0.5, 0.5]$, and gradient norms are clipped at 1.

3. Overlap refinement – recovering buildings after NMS

In this section, we provide additional detail on how we recover valid building proposals that were removed during overlapping proposal detection, which is described in Section 3.1.3 of the main paper. Recall that Task 1 of the multitasking framework provides us with building proposals, where the proposals may overlap with each other. Task 2 learns the boundary proximity signed heatmap (BPSH) of the image, which is used to refine the overlapping proposals of Task 1. For this refinement, each building proposal is scored using (1) the proposal confidence from Task 1 and (2) the proposal's agreement with the building boundaries indicated by the zero-level set of the BPSH. We apply a non-maximum suppression (NMS) based on these new scores to

remove proposals whose predicted building masks overlap.

However, this NMS may in turn suppress valid buildings. For example, consider the two buildings shown in the top left corner of the left subfigure in Fig. 1, above. For two buildings B_1 and B_2 that are very close to each other, Task 1 may generate two building proposals: (1) one proposal that covers B_1 but not B_2 (the purple-colored proposal in the figure), and (2) a proposal that combines B_1 and B_2 together into a single building (the teal-colored proposal; parts of the teal-colored mask are covered by the purple-colored proposal's mask). The scoring mechanism described in Eq. (2) of the main paper will give a higher score to the first proposal, as it will have a stronger agreement to the BPSH zero-level set. The NMS will suppress the second proposal as a result. To recover building B_2, we regenerate proposals as explained next.

Let $\mathcal{B}$ denote the set of building proposals output by Task 1, with each building $B \in \mathcal{B}$ having a binary pixel-wise mask M_B and confidence c_B. Let M be the aggregated pixel-wise mask of all building proposals after NMS (Fig. 1, middle). We create a "recovered" mask M^+ (Fig. 1, second from right) for suppressed buildings containing pixels that: (1) have no overlap with M, (2) are inside a building according to the BPSH, and (3) are part of at least one highly confident proposal in $\mathcal{B}$. We specify these criteria at each pixel p:

$$M^+(p) = \Bigg(\overline{M}(p) \wedge (BPSH(p) > \alpha_1) \\ \wedge \exists_{B \in \mathcal{B}} \left(M_B(p) \wedge c_B > \alpha_2 \right) \Bigg) \qquad (1)$$

For the second condition enforcing that the pixel is inside a building as indicated by the predicted BPSH, we use $\alpha_1 = 0.5$. The last condition enforces that at least one building mask proposal covers pixel p with some minimum confidence α_2. We experimentally found that $\alpha_2 = 0.7$ offers good building retention without introducing many false-positive detections. The connected components in M^+ with covering fewer than 256 pixels are suppressed. Finally, the remaining connected components of M^+ are added to the predicted building set to obtain the final set of building outlines.

4. Evaluation datasets

In this section, we give a brief overview of the different datasets used in our experiments.

GRSS_DFC_2019 dataset The GRSS_DFC_2019 dataset [1, 4] contains multi-spectral satellite images, nDSMs, and semantic segmentations for parts of Jacksonville, FL and Omaha, NE with over 100,000 building instances. We split the dataset containing buildings into a training set containing 88 regions and a test set containing 11 other regions. The ground truth contains semantic segmentations over building, ground, vegetation, water, bridge deck, and unclassified regions. To extract the ground-truth building footprints from the slightly noisy semantic segmentation, we apply a morphological closing operation on the building mask, followed by removing objects with areas smaller than $23m^2$ (256 pixels).

USSOCOM Urban 3D dataset The USSOCOM Urban 3D dataset [5] contains RGB satellite images and nDSMs for parts of Jacksonville and Tampa, Florida, with over 180 km^2 of land coverage containing over 74000 buildings. We split the dataset into a training set containing 130 regions and a test set containing 44 other regions.

SpaceNet Buildings Dataset v2 SpaceNet [3] contains multi-spectral satellite images with ground-truth building masks for over 300,000 building instances, which we use to evaluate 2D building outline detections, without training for height or semantic segmentation. For our experiments, we split the dataset with 7128 training and 1254 testing images.

Potsdam and Vaihingen datasets These datasets are released by the ISPRS 2D Semantic Labeling contest [2] and contain high-resolution but lower-spatial-coverage aerial images. The Potsdam dataset contains 4-channel (R,G,B,NIR) aerial images, while the Vaihingen dataset provides 3-channel (IR,R,G) aerial images. Semantic segmentations of buildings, low vegetation, trees, impervious surfaces, cars, and background are available for both datasets, and both datasets include normalized DSMs.

Following [7], for the Potsdam dataset we select 10 images for training, and the remaining 7 images (image IDs: 02_11, 02_12, 04_10, 05_11, 06_07, 07_08, and 07_10) are used to test all models. Similarly, for the Vaihingen dataset we follow [7] and select 11 images for training, and the remaining 5 images (image IDs: 11, 15, 28, 30, and 34) are used for testing our model.

5. Additional ablation

We study the effectiveness of our proposed overlap refinement technique compared to an naïve NMS technique. In the näive NMS setting, building outline proposals generated by Task 1 are suppressed if they overlap another proposal with higher confidence. Table 1 shows that, in comparison of F1 scores, the proposed refinement technique is more robust in retaining true building instances compared to the näive NMS. The table also contrasts between performing the overlap refinement using the BPSH learned by the multi-task network, and using the improved BPSH using instance-level reasoning (see Section 3.5 in the main paper). Instance-level reasoning improves the BPSH across all datasets, which is evidenced by the fact that this step consistently improves refinement.

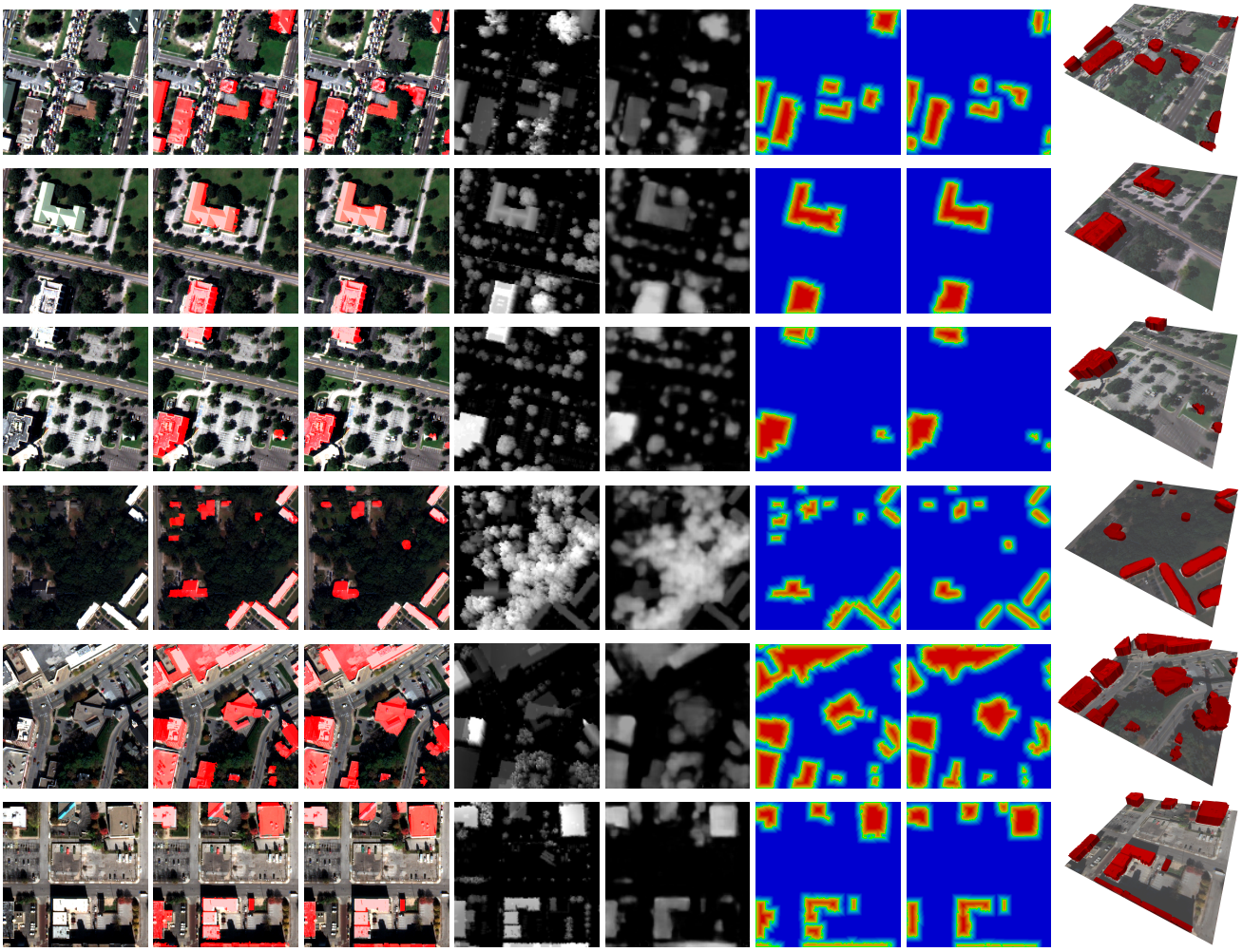


Figure 2: Ground-truth (GT) and predicted masks, height maps, and BPSH maps for different satellite images from the GRSS_DFC_2019 dataset.

	Naïve NMS	With Overlap Refinement	Overlap & BPSH Refinement
SpaceNet	67.12	68.50	68.87
GRSS_DFC	63.70	67.53	68.34
Urban 3D	80.87	82.67	82.89
Potsdam	67.47	69.81	71.98
Vaihingen	67.39	70.23	72.85

Table 1: Comparison of building detection F1 scores (higher is better) for different overlap refinement techniques, and with our BPSH refinement network.

6. Additional qualitative results

Fig. 2 provides example building mask, nDSM, and BPSH outputs of our network for images from the GRSS_DFC_2019 dataset, along with 3D reconstruction visualizations for each image. We provide additional qualitative results for our method versus the methods of Wang and Frahm [9] and Srivastava *et al*. [8] on the Potsdam, Vaihingen, and Urban3D datasets. We compare the building outline detection performance of our method versus Wang and Frahm [9] in Figs. 3 and 8. Our proposed method is contrasted against Srivastava *et al*. [8] in Figs. 4, 5, 6, 7, 9, 10, 11, and 12. Results on the Urban3D dataset with zoom-in crops are shown for both methods, as well as our own, in Fig. 13. For nDSM visualizations, the black-to-white visualization range is clamped between 0 and 25.5 meters. Additional 3D scene reconstructions for our method are shown in Figs. 14 and 15 for the Potsdam and Vaihingen datasets, respectively.

References

[1] 2019 IEEE GRSS Data Fusion Contest. http://www.grss-ieee.org/community/technical-committees/data-fusion. Accessed on: 2019-03-15. 2

[2] Isprs 2d semantic labeling contest. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html. Last modified: 2019-10-25, Accessed on: 2019-10-25. 2

[3] SpaceNet on Amazon Web Services (AWS). datasets. the spacenet catalog. https://spacenetchallenge.github.io/datasets/datasetHomePage.html. Last modified: 2018-04-30, Accessed on: 2019-03-15. 1, 2

[4] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D. Hager, and Myron Z. Brown. Semantic stereo for incidental satellite images. *CoRR*, abs/1811.08739, 2018. 2

[5] Hirsh Goldberg, Myron Brown, and Sean Wang. A benchmark for building footprint classification using orthorectified rgb imagery and digital surface models from commercial satellites. In *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop 2017*, 2017. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

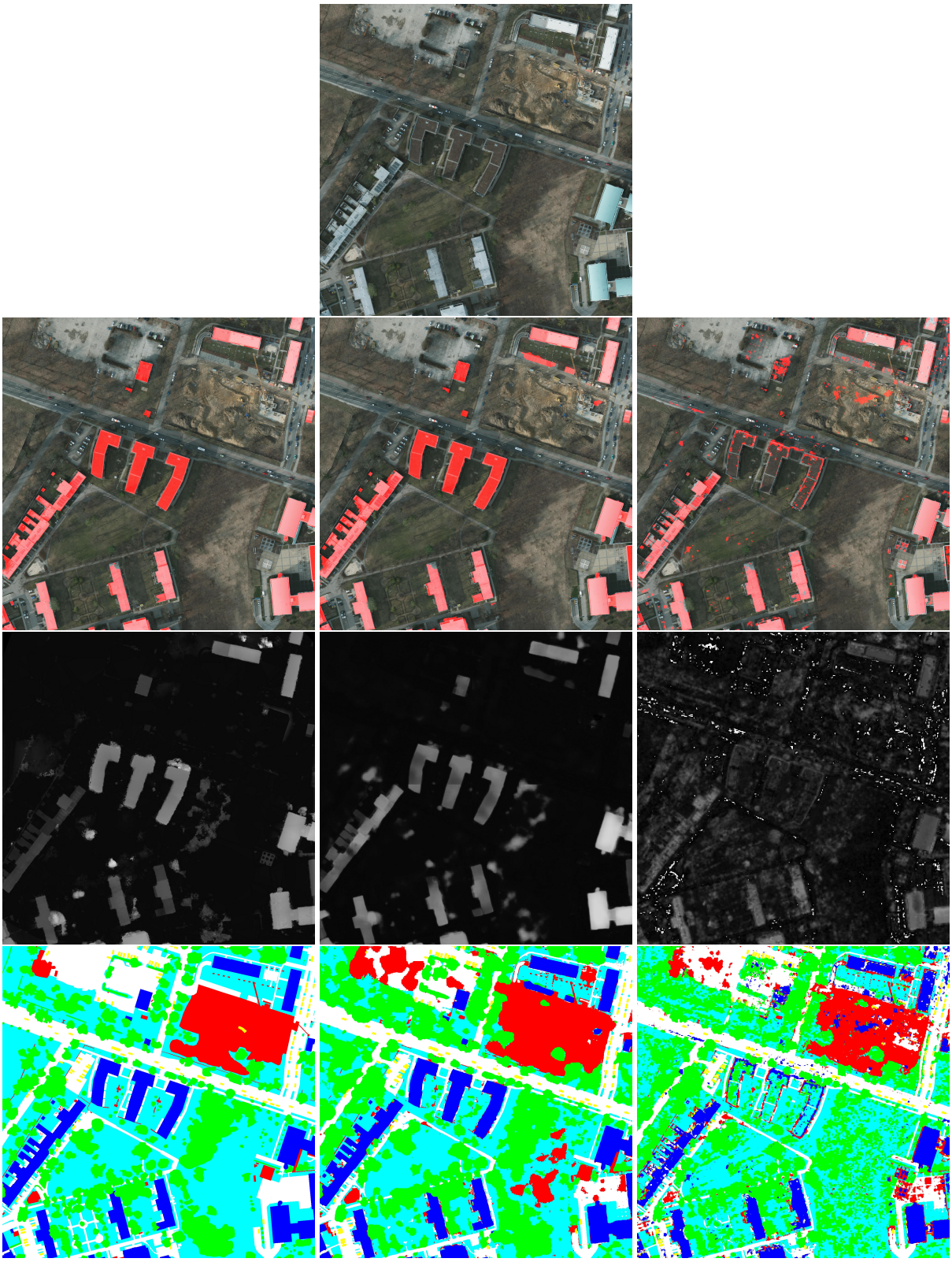
[7] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2019. 2

[8] Shivangi Srivastava, Michele Volpi, and Devis Tuia. Joint height estimation and semantic labeling of monocular aerial images with cnns. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5173–5176. IEEE, 2017. 3, 6, 7, 8, 9, 11, 12, 13, 14, 15

[9] Ke Wang and Jan-Michael Frahm. Single view parametric building reconstruction from satellite imagery. In *2017 International Conference on 3D Vision (3DV)*, pages 603–611. IEEE, 2017. 3, 5, 10, 15

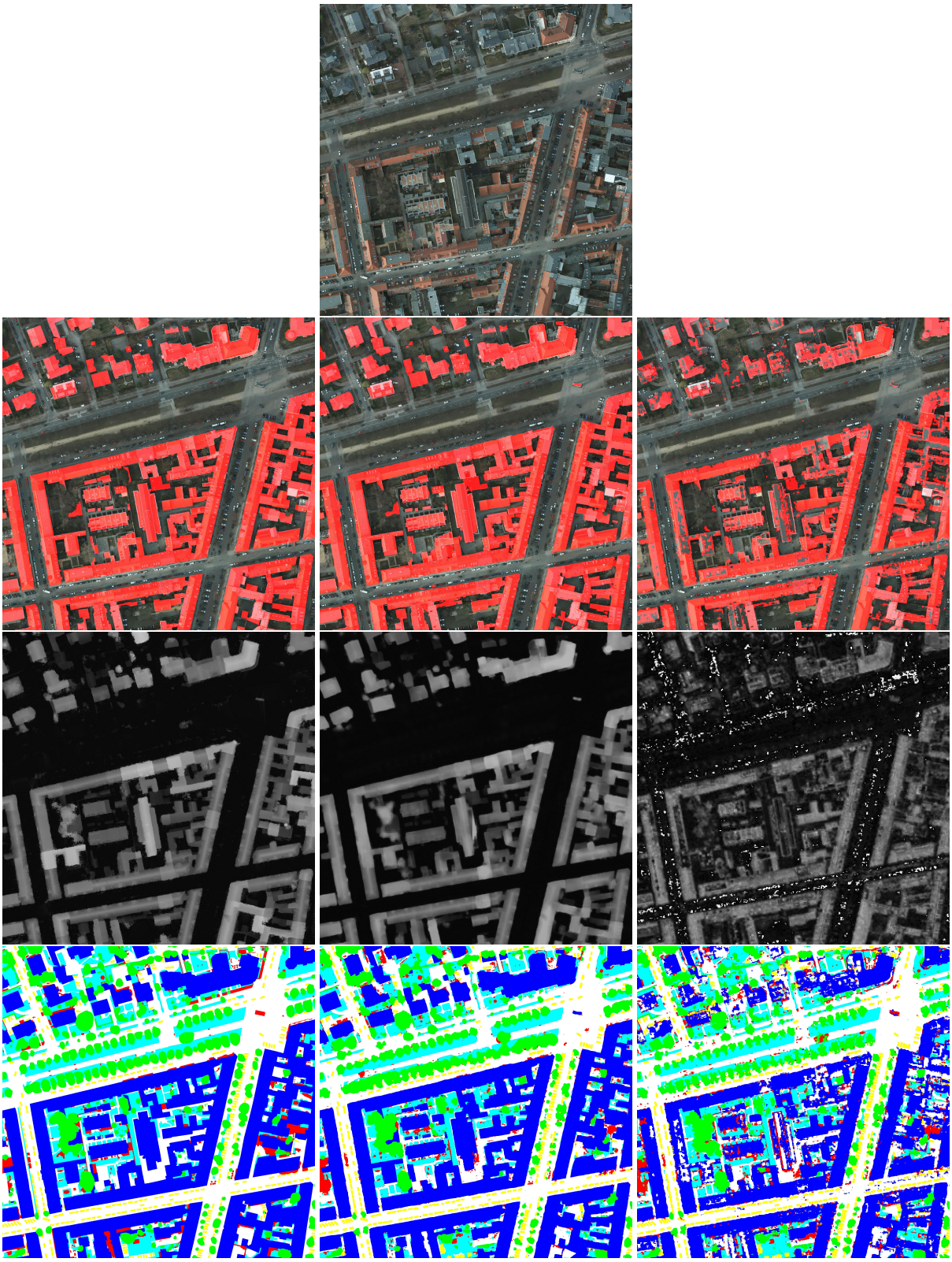
Ground truth | Ours | Wang and Frahm [9]

Figure 3: Qualitative result on our method versus Wang and Frahm's [9] method. First two rows: Potsdam dataset. Bottom two rows: Vaihingen dataset.



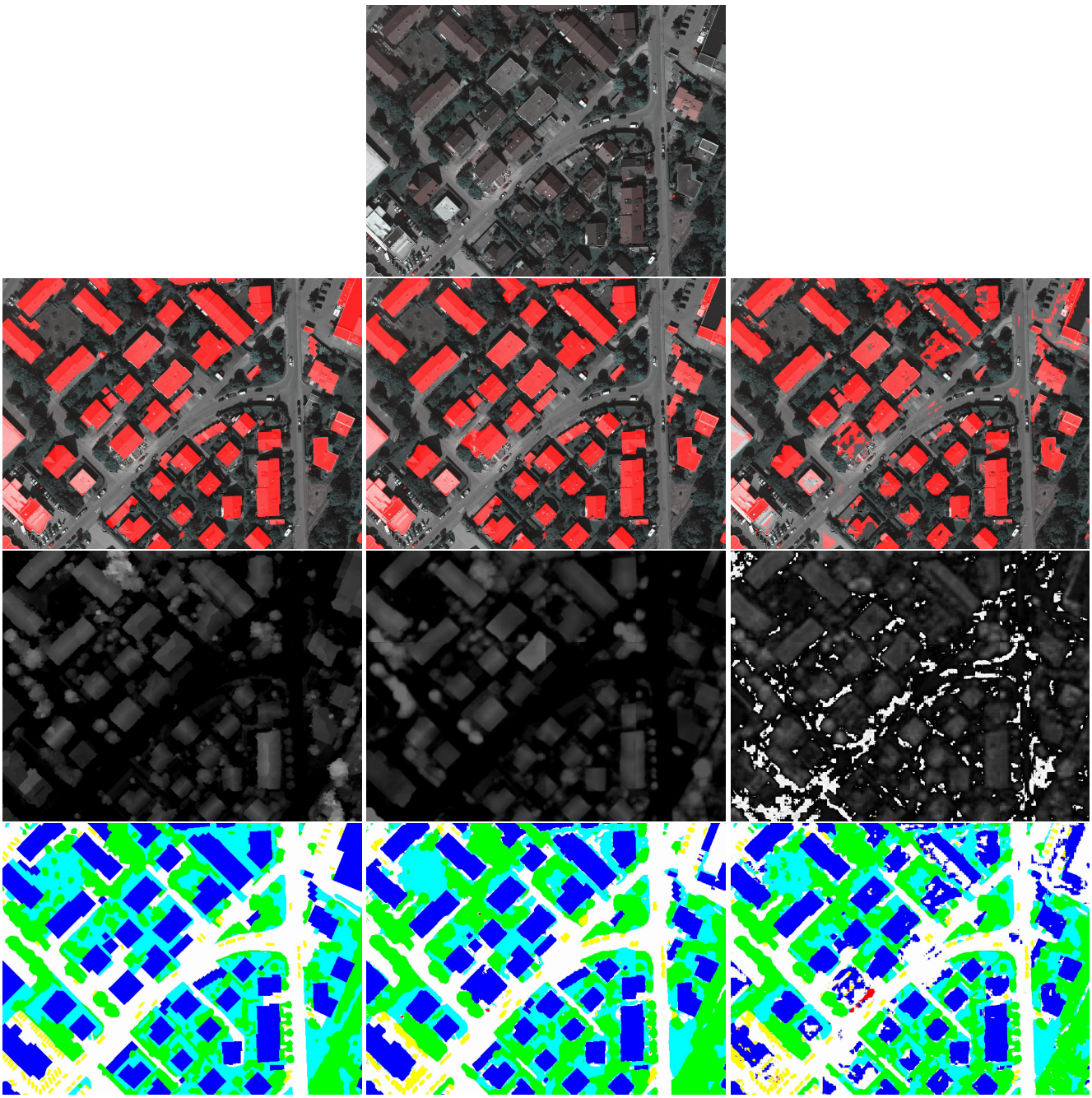
Ground truth | Ours | Srivastava *et al*. [8]

Figure 4: Results on the Potsdam dataset for our method versus [8]. First row: aerial image. Second row: predicted building instances overlaid on the overhead image. Third row: nDSM. Fourth row: semantic segmentation.



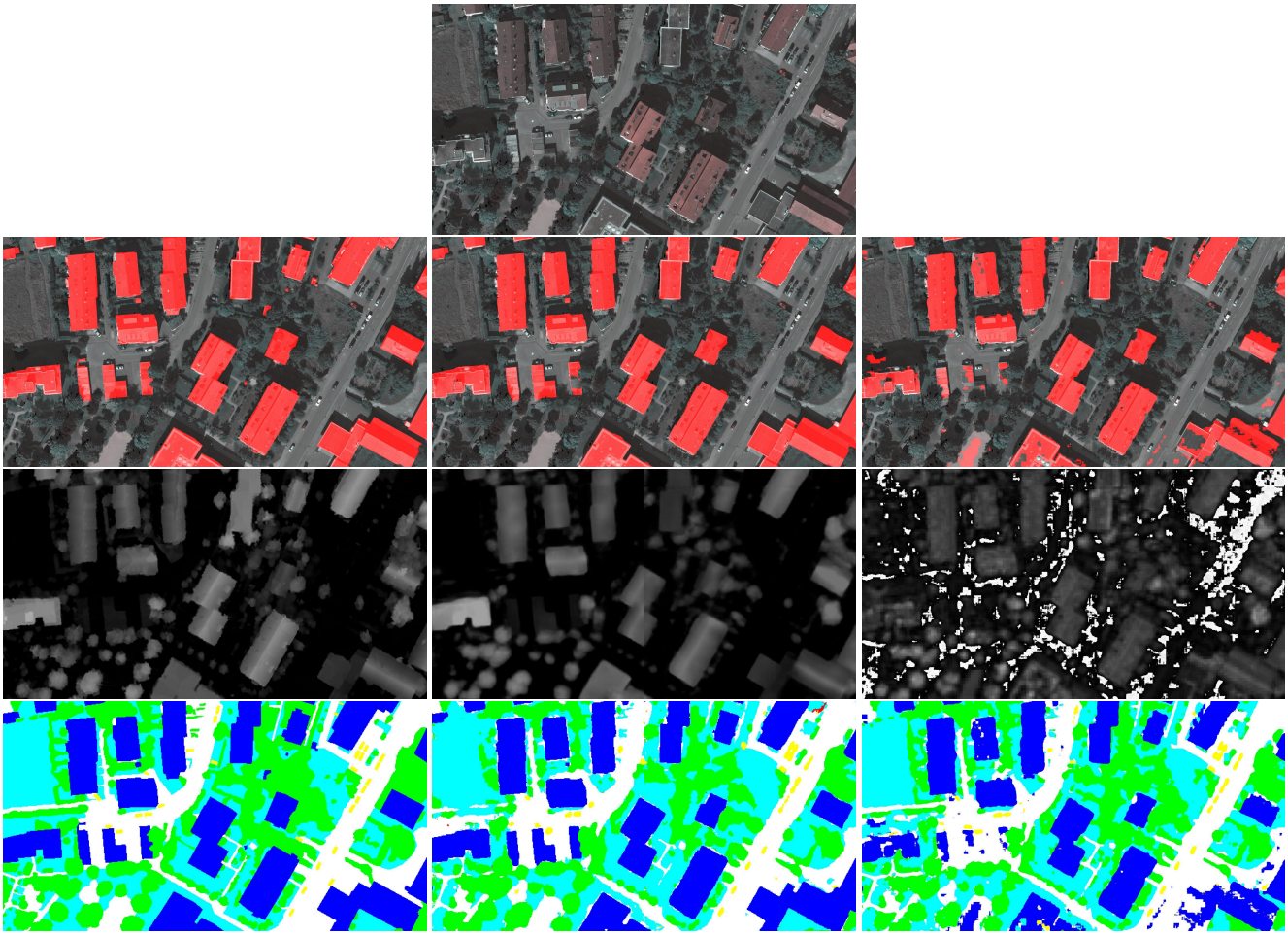
Ground truth | Ours | Srivastava *et al*. [8]

Figure 5: Additional results on the Potsdam dataset for our method versus [8]. First row: aerial image. Second row: predicted building instances overlaid on the overhead image. Third row: nDSM. Fourth row: semantic segmentation..



Ground truth | Ours | Srivastava *et al.* [8]

Figure 6: Results on the Vaihingen dataset for our method versus [8]. First row: aerial image. Second row: predicted building instances overlaid on the overhead image. Third row: nDSM. Fourth row: semantic segmentation.

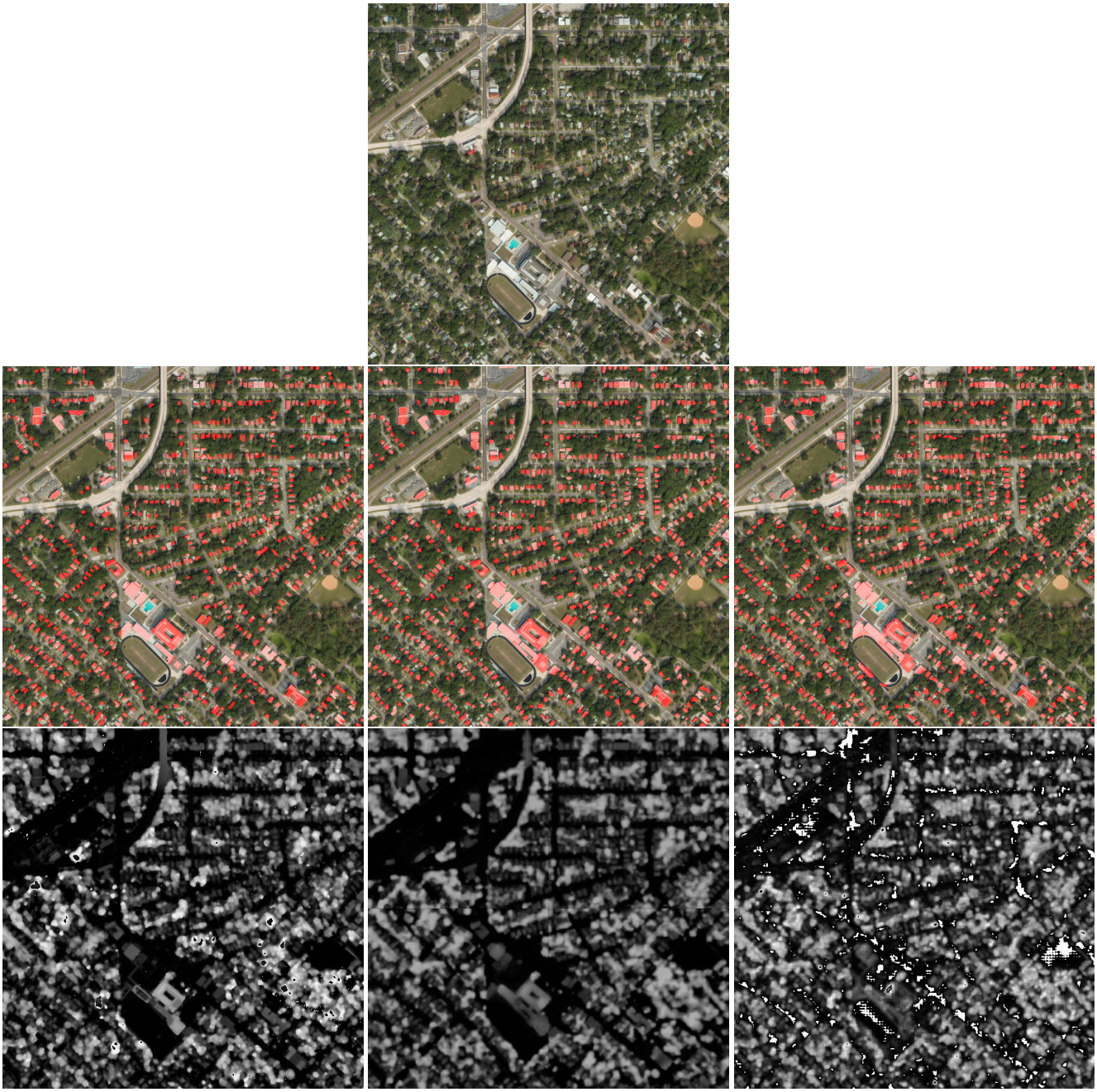


Ground truth | Ours | Srivastava *et al.* [8]

Figure 7: Additional results on the Vaihingen dataset for our method versus [8]. First row: aerial image. Second row: predicted building instances overlaid on the overhead image. Third row: nDSM. Fourth row: semantic segmentation.

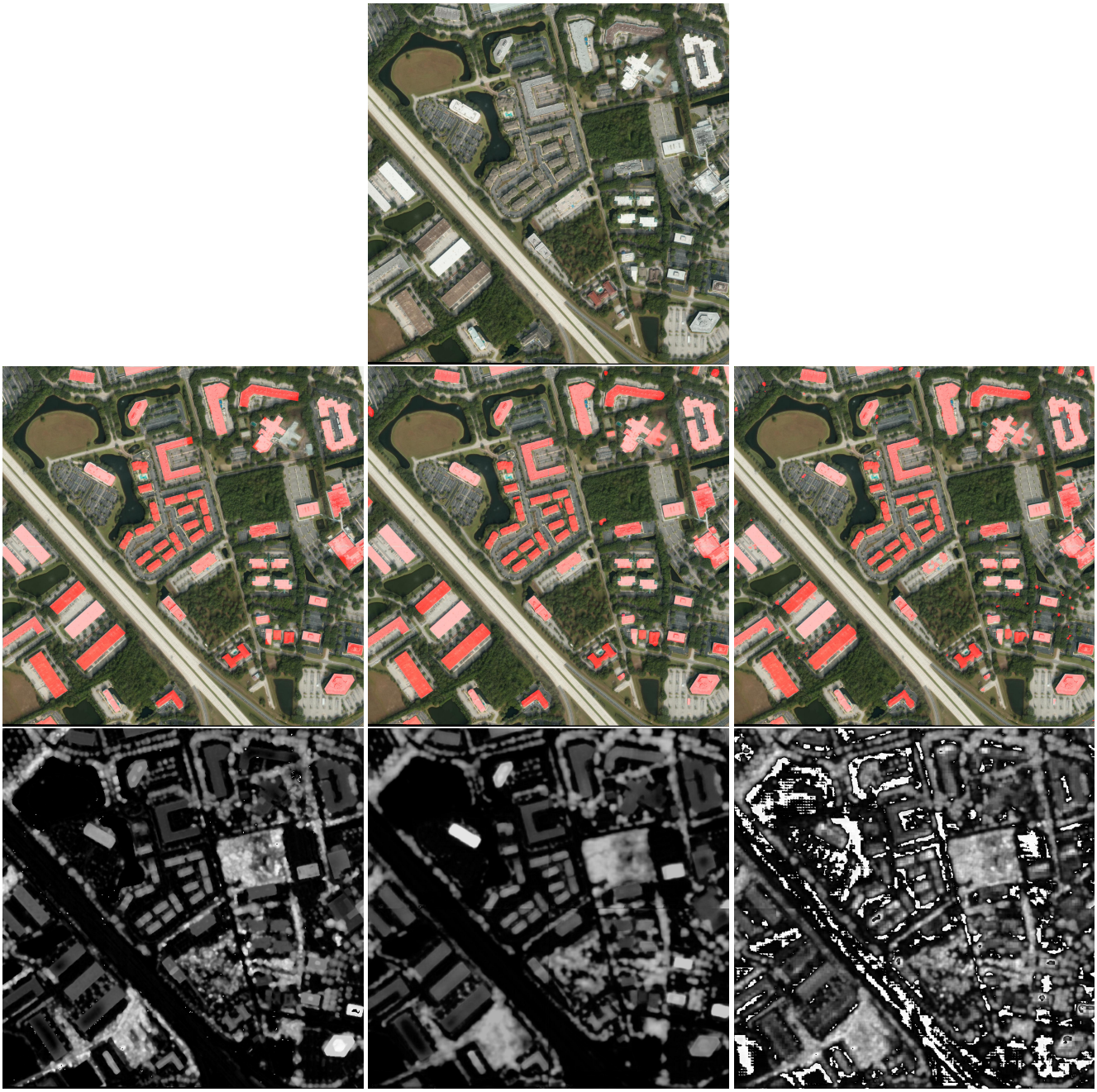
Ground truth | Ours | Wang and Frahm [9]

Figure 8: Qualitative results on our method versus Wang and Frahm's [9] method on four images from the Urban3D dataset.



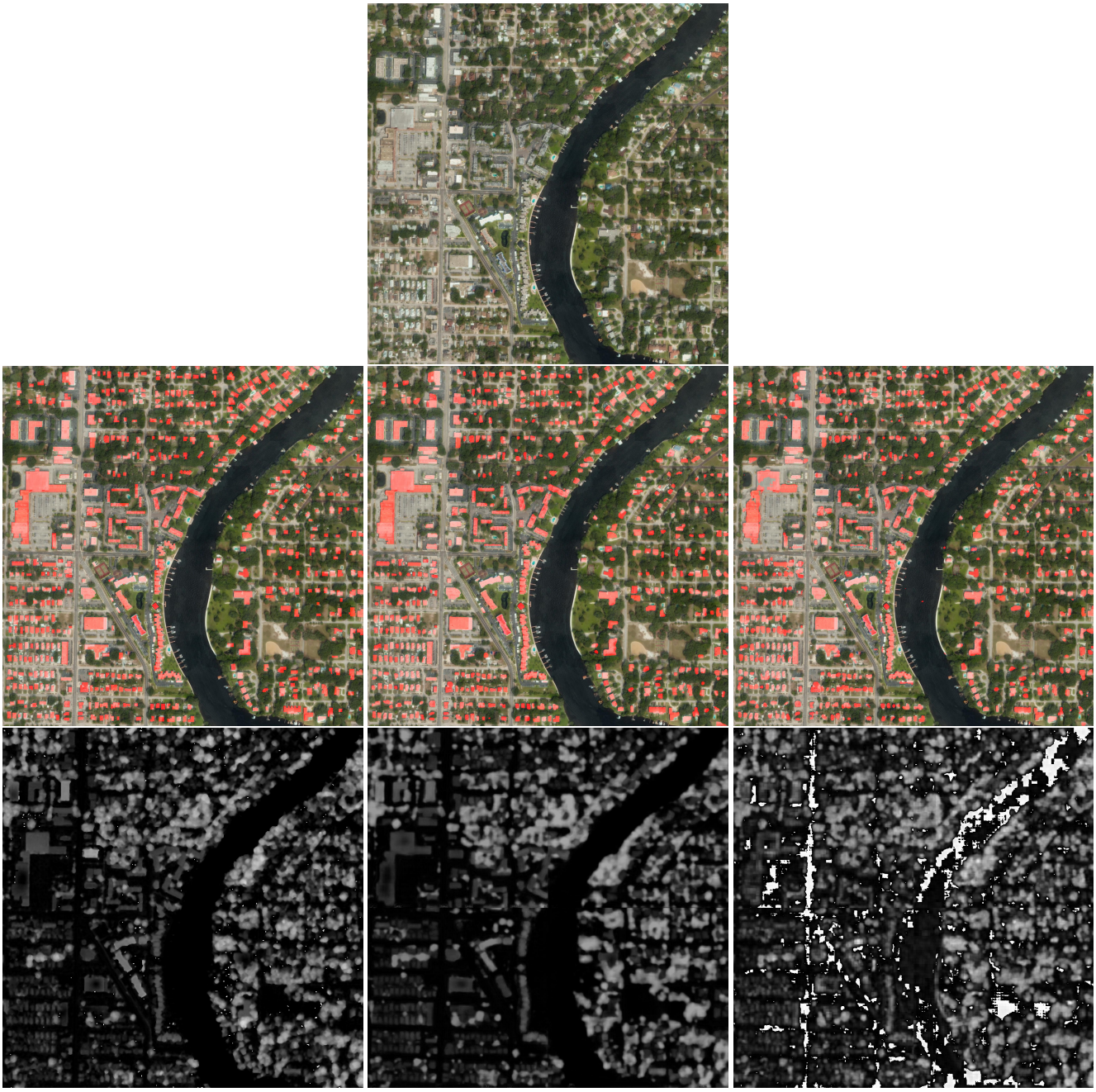
Ground truth | Ours | Srivastava *et al.* [8]

Figure 9: Results on the Urban3D dataset for our method versus [8]. First row: aerial image. Second row: predicted building instances overlaid on the overhead image. Third row: nDSM.



Ground truth | Ours | Srivastava *et al.* [8]

Figure 10: Additional results on the Urban3D dataset for our method versus [8]. First row: aerial image. Second row: predicted building instances overlaid on the overhead image. Third row: nDSM.



Ground truth | Ours | Srivastava *et al.* [8]

Figure 11: Additional results on the Urban3D dataset for our method versus [8]. First row: aerial image. Second row: predicted building instances overlaid on the overhead image. Third row: nDSM.

Ground truth | Ours | Srivastava *et al.* [8]

Figure 12: Additional results on the Urban3D dataset for our method versus [8]. First row: aerial image. Second row: predicted building instances overlaid on the overhead image. Third row: nDSM. Our prediction apparently does not capture two ground-truth building instances in the top-right corner. However, the ground truth may be incorrect for both of these instances, as we can observe roads passing through them. When visualized in Google Earth, as well, these instances do not appear to be buildings.

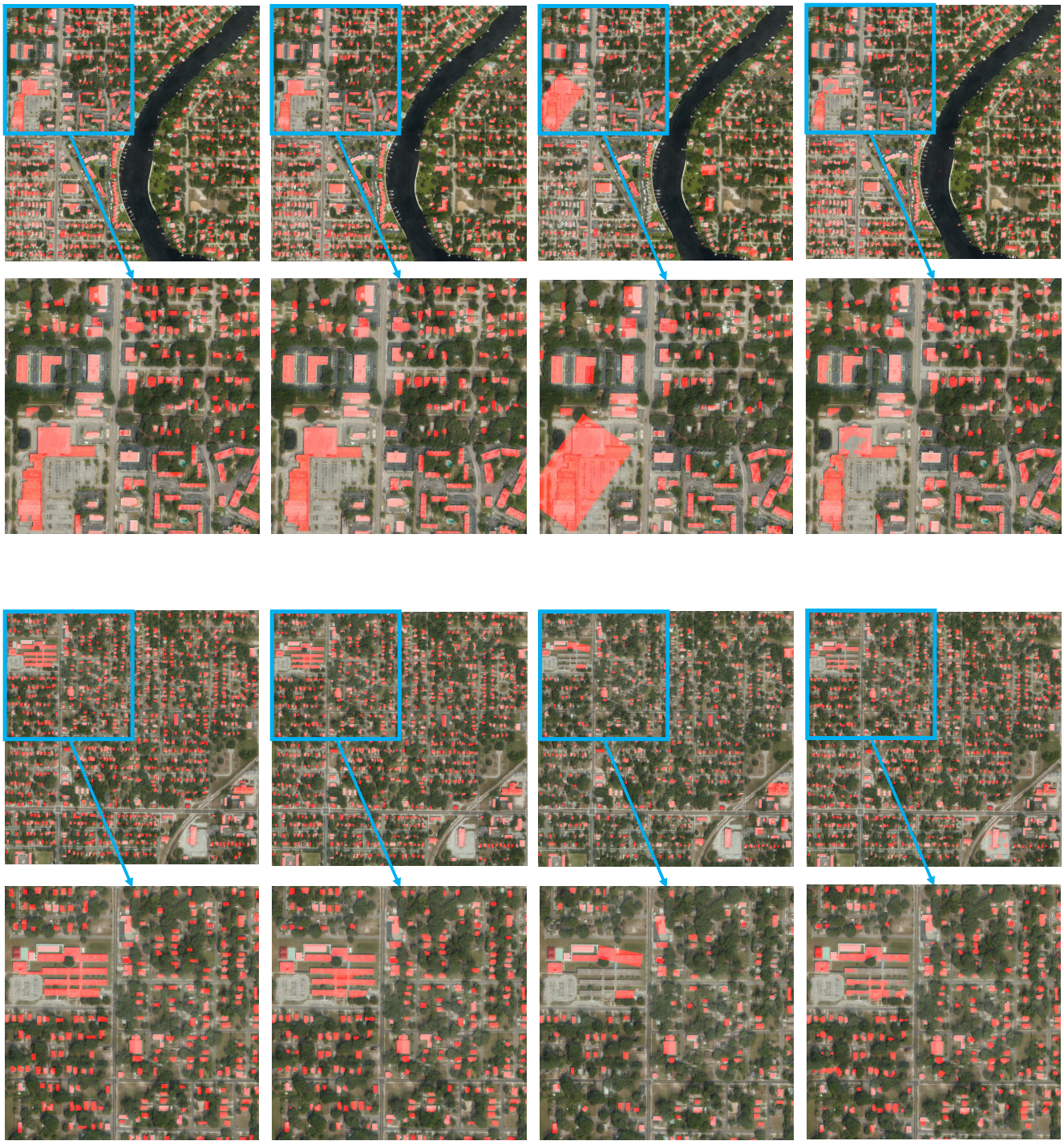


Figure 13: Comparison of building mask predictions, plus zoom-in crops, for our method versus [9] and [8] on two images from the Urban3D dataset.

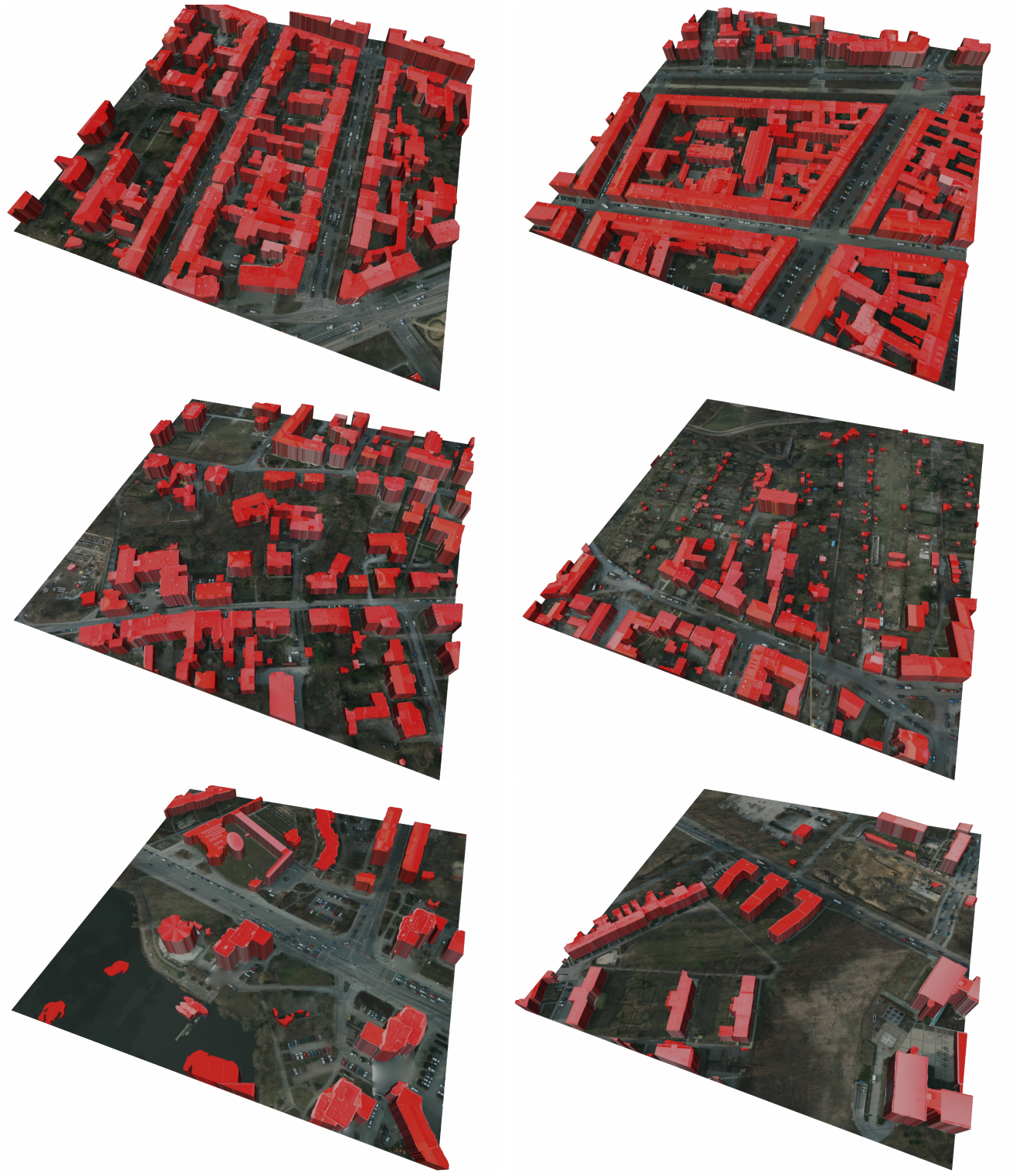


Figure 14: 3D building models generated for the Potsdam dataset.

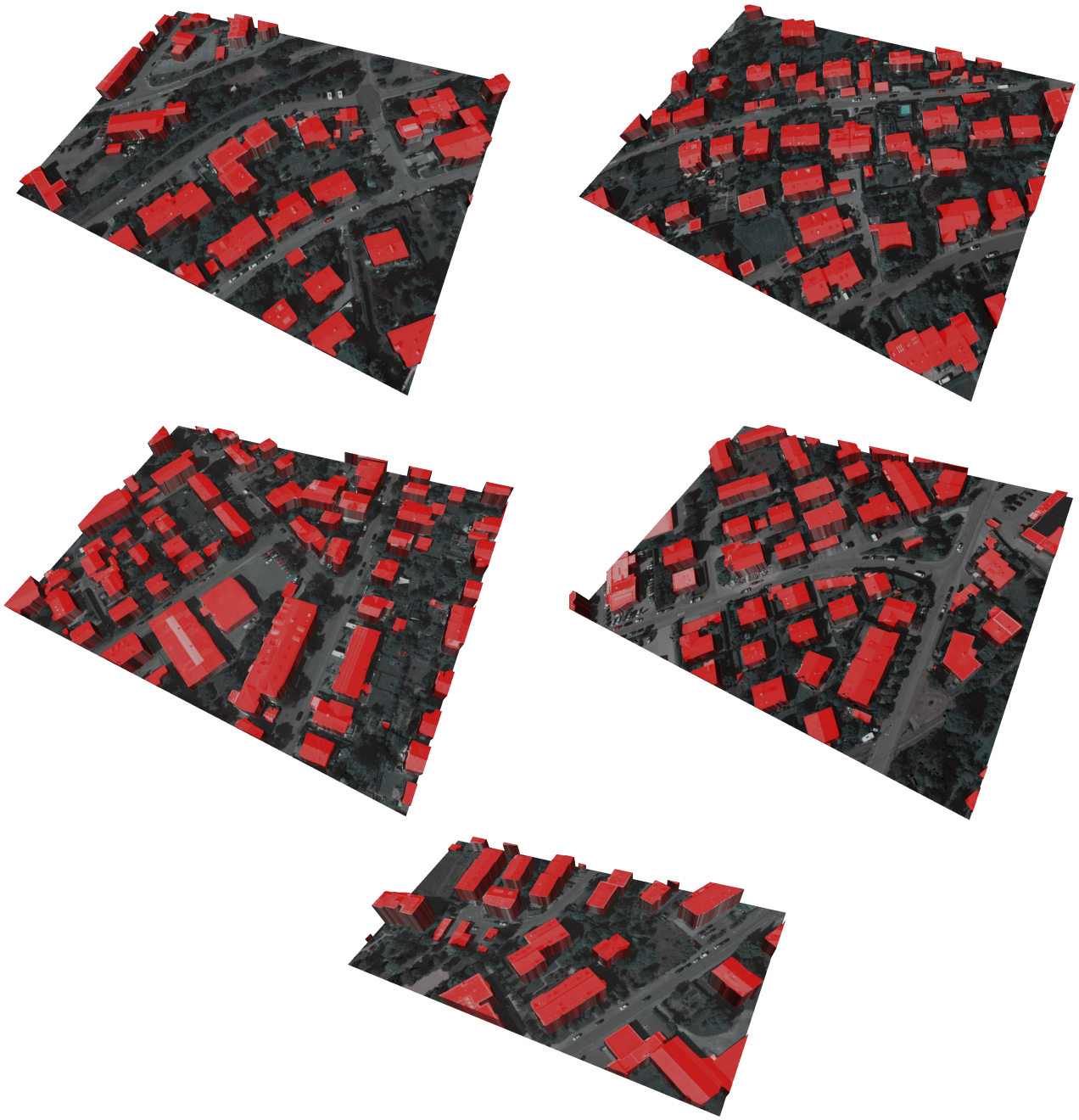


Figure 15: 3D building models generated for the Vaihingen dataset.