Supplementary Material: Single View Geocentric Pose in the Wild

Gordon Christie[1], Kevin Foster[1], Shea Hagstrom[1], Gregory D. Hager[2], Myron Z. Brown[1]
[1]The Johns Hopkins University Applied Physics Laboratory
[2]Department of Computer Science, The Johns Hopkins University

In this supplementary document we provide:

1: Summary and discussion of data statistics
2: Expanded discussion of metrics and results
3: Examples and discussion of limiting factors

1. Data Statistics

Statistics for the datasets used in our experiments are summarized in Tab. 1. Statistics for the public Urban Semantic 3D (US3D) dataset [1], including terrain variation and height distributions, are separately reported by [2]. We extended US3D with new public data for San Fernando, Argentina which presents additional challenges, with fewer tall buildings and increased architectural diversity.

Fig. 1 characterizes distributions of scale factor values that relate heights above ground to their respective vector field magnitudes for mapping surface-level features to ground level. Values are higher for more oblique images and close to zero for near-nadir viewing geometry. The train and test sets are well balanced.

Distributions for height above ground values are shown in Fig. 2. Our new train and test set for Argentina is well balanced, as are those from Atlanta and Omaha. While the overall DFC19 dataset [3] including both Jacksonville and Omaha are well balanced, the Jacksonville test set does not capture the full range of values represented in its train set.

2. Metrics

In our paper, for consistency we report accuracy with root mean square error (RMSE). Results by [2] were reported as mean absolute error (MAE), so for completeness we demonstrate our improvements in terms of MAE in Tab. 2 and Tab. 3. There are small differences between our numbers and those reported in [2] because of minor dataset changes they made before public release.

For relative assessment of performance for multiple cities, we adopt the R^2 metric defined below and report results in Tab. 4. R^2 clearly indicates relative prediction accuracy among cities, as shown in Fig. 3. In particular, R^2 correctly indicates that the predictive power of our regression model for ARG is much lower than for the other sites.

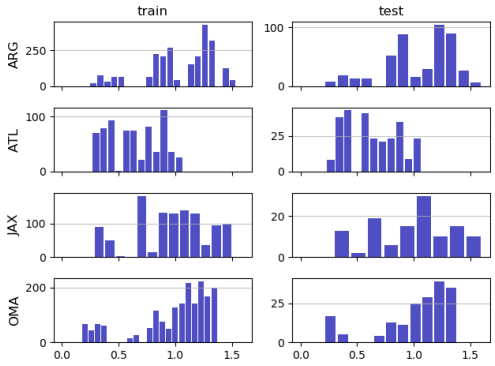

Figure 1: Histograms show distribution of image-level scale factors (pixels/meter) relating heights above ground to their respective vector field magnitudes for mapping surface pixels to ground level. Higher values represent images with more oblique viewing angle.

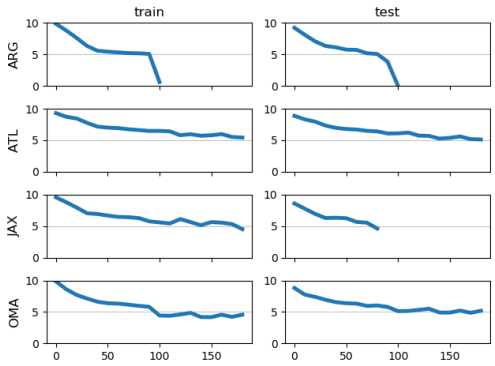

Figure 2: Plots show the distributions of height above ground (meters) with $log10$ pixel counts for all sites.

We define R^2 in terms of the residual sum of squares (RSS) of predicted values $f(x_i)$ for n observed samples x_i and reference values y_i in Equation 1. RMSE, $\sqrt{RSS/n}$, is useful for measuring accuracy in units of the dependent variable y_i (e.g., meters for heights and pixels for the vector field) for a single dataset; however, for multiple datasets

Table 1: Statistics for our train and test sets.

	Jacksonville, Florida (JAX)	Omaha, Nebraska (OMA)	Atlanta, Georgia (ATL)	San Fernando, Argentina (ARG)
Train image chips	1098	1796	704	2325
Test image chips	120	178	264	463
Source satellite images	24	43	30	39
Train geographic tiles	52	53	52	63
Test geographic tiles	5	5	22	13
Imaging satellite	WorldView-3	WorldView-3	WorldView-2	WorldView-3
Pixel size range (cm)	31 – 39	31 – 36	47 – 59	31 – 41
Azimuth angle range (deg)	3 – 268	3 – 349	0 – 358	14 – 352
Elevation angle range (deg)	57 – 84	63 – 86	56 – 81	54 – 84
Year range	2014 – 2016	2014 – 2015	2009	2015
Max height above ground (m)	200	200	200	100

Method	Train	Mag	Angle	Endpoint	Height
FLOW-HA [2]	DFC19	2.62	16.82	3.00	2.26
FLOW-H [2]	DFC19	2.32	15.58	2.80	2.14
Ours-NoAug	DFC19	1.81	11.62	2.18	1.66
Ours	DFC19	**1.71**	9.08	1.93	1.66
Ours-NoAug	All Cities	1.84	14.59	2.32	1.71
Ours	All Cities	1.72	**8.24**	**1.88**	**1.64**

Table 2: Our method improves on state of the art MAE errors for the DFC19 test set.

Method	Train	Mag	Angle	Endpoint	Height
FLOW-HA [2]	ATL	3.53	15.54	4.12	4.68
FLOW-H [2]	ATL	2.79	9.27	3.05	4.00
Ours-NoAug	ATL	**1.78**	7.77	2.13	**2.48**
Ours	ATL	1.87	9.50	2.23	2.72
Ours-NoAug	All Cities	1.81	9.02	2.16	2.52
Ours	All Cities	1.89	**7.15**	**2.12**	2.75

Table 3: Our method improves on state of the art MAE errors for the ATL test set.

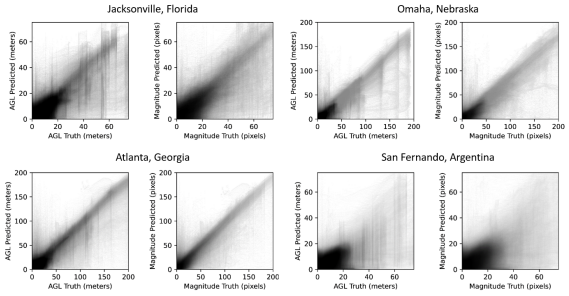

Figure 3: Above ground level (AGL) heights and vector field magnitudes from our model are compared to reference values for all test cities. Pixel intensity indicates count.

with varying value scales (e.g., large building height values in urban scenes and smaller values in suburban scenes), a normalized metric is more discriminating for measuring the estimator's predictive power. We normalize RSS by the total sum of squares (TSS) of the dependent variable in Equation 2, leading to the coefficient of determination R^2 in Equation 3.

	JAX	OMA	ATL	ARG
Height RMSE (m)	3.33	4.15	4.86	3.00
Endpoint RMSE (pix)	3.61	4.63	3.66	3.56
Height R^2	0.81	0.87	0.89	0.60
Endpoint R^2	0.84	0.88	0.90	0.68

Table 4: Our height and vector field prediction RMSE and R^2 are shown for four cities with significantly different value ranges. For RMSE, lower is better. Higher is better for $R^2 \in [0,1]$. R^2 much more clearly indicates relative prediction accuracy among cities.

$$RSS = \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2 \tag{1}$$

$$TSS = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 \tag{2}$$

$$R^2 = \max\left(0,\ 1 - \frac{RSS}{TSS}\right) \tag{3}$$

While R^2 is commonly applied for linear regression of trend lines, the general form measures the fraction of the total variance explained by any estimator's predictions. Since RSS can exceed TSS for a poor prediction, we clip negative values to zero such that $R^2 \in [0,1]$.

3. Examples and Limiting Factors

State of the art performance: Our method exploits invariant properties of affine imaging geometry to achieve state of the art performance, outperforming [2] by a wide margin. Comparisons for San Fernando, Argentina (ARG) in Fig. 7, Jacksonville, Florida (JAX) in Fig. 8, Omaha, Nebraska (OMA) in Fig. 9, and Atlanta, Georgia (ATL) in Fig. 10 all clearly show that our model produces more

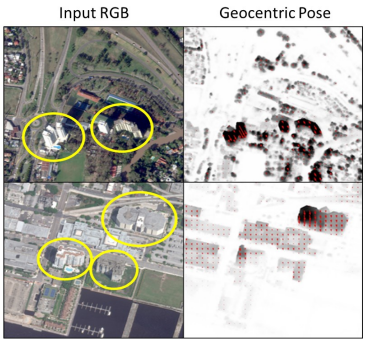

Figure 4: Heights of buildings with unusual appearance, highlighted in yellow, are not consistently well predicted.

consistently accurate height predictions and vector fields for rectification, particularly for tall buildings.

Variety of appearance: Our model performs very well for objects that are well-represented in the train set, including tall buildings; however, our model often under-predicts heights for buildings with unique appearance not captured in the train set (Fig. 4). Failure cases are often viewpoint-dependent, with less accurate predictions for more oblique views and for views without visible shadows (Fig. 5 and 7). To emphasize errors in Fig. 6, we converted RGB images to HSV and replaced intensity with $max(EPE, 20)$ normalized to fill the value range, where EPE is endpoint error (pixels).

We believe that more comprehensive geometric augmentations to render novel viewpoints and properly cast shadows may help address this limitation; however, the observed view-dependence of performance suggests that the variety of appearance for building facades must also be addressed. While our initial experiment using multi-view stereo instead of lidar for supervision is limited in scope, we believe that continuing with this approach will help address this challenge of diversity in visual appearance because satellite images can be acquired over much larger scales than lidar.

Partial occlusion: We show anecdotal evidence that smoke from chimneys or smokestacks induce gradual reduction in prediction accuracy (Fig. 11). Light haze in images also does not appear to significantly degrade performance (Fig. 9).

Small vertical features: Our model consistently under-predicts height for small vertical structures (Fig. 6); however, we do not consider this a failure case. We believe that inclusion of these structures in training inhibits learning for larger features that are more relevant to mapping applications, so we remove those reference heights in training. Interestingly, predictions from [2] depict the tall antenna shown in Fig. 10, though predicted heights are inaccurate.

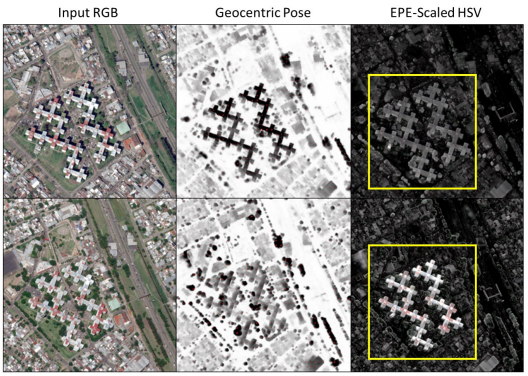

Figure 5: Prominent failure cases occur for some of the most oblique viewpoints without shadows (shown below).

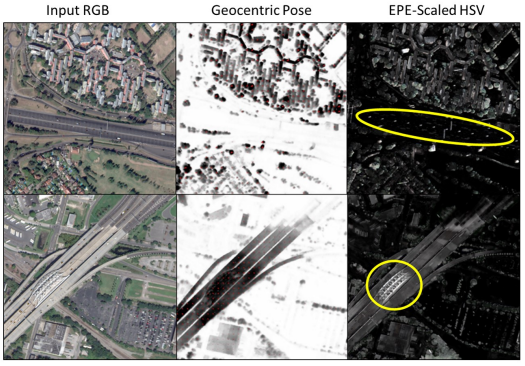

Figure 6: Small vertical structures such as lamp posts and bridge struts, highlighted yellow, are ignored by our model.

Acknowledgements

This work was supported by the National Geospatial-Intelligence Agency and approved for public release, 21-484, with distribution statement A – approved for public release; distribution is unlimited. Commercial satellite images were provided courtesy of DigitalGlobe.

References

[1] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic Stereo for Incidental Satellite Images. In *WACV*, 2019. 1

[2] Gordon Christie, Rodrigo Rene Rai Munoz Abujder, Kevin Foster, Shea Hagstrom, Gregory D Hager, and Myron Z Brown. Learning Geocentric Object Pose in Oblique Monocular Images. In *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 7

[3] S. Kunwar, H. Chen, M. Lin, H. Zhang, P. Dangelo, D. Cerra, S. M. Azimi, M. Brown, G. Hager, N. Yokoya, R. Hansch, and B. Le Saux. Large-scale semantic 3d reconstruction: Outcome of the 2019 ieee grss data fusion contest - part a. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020. 1

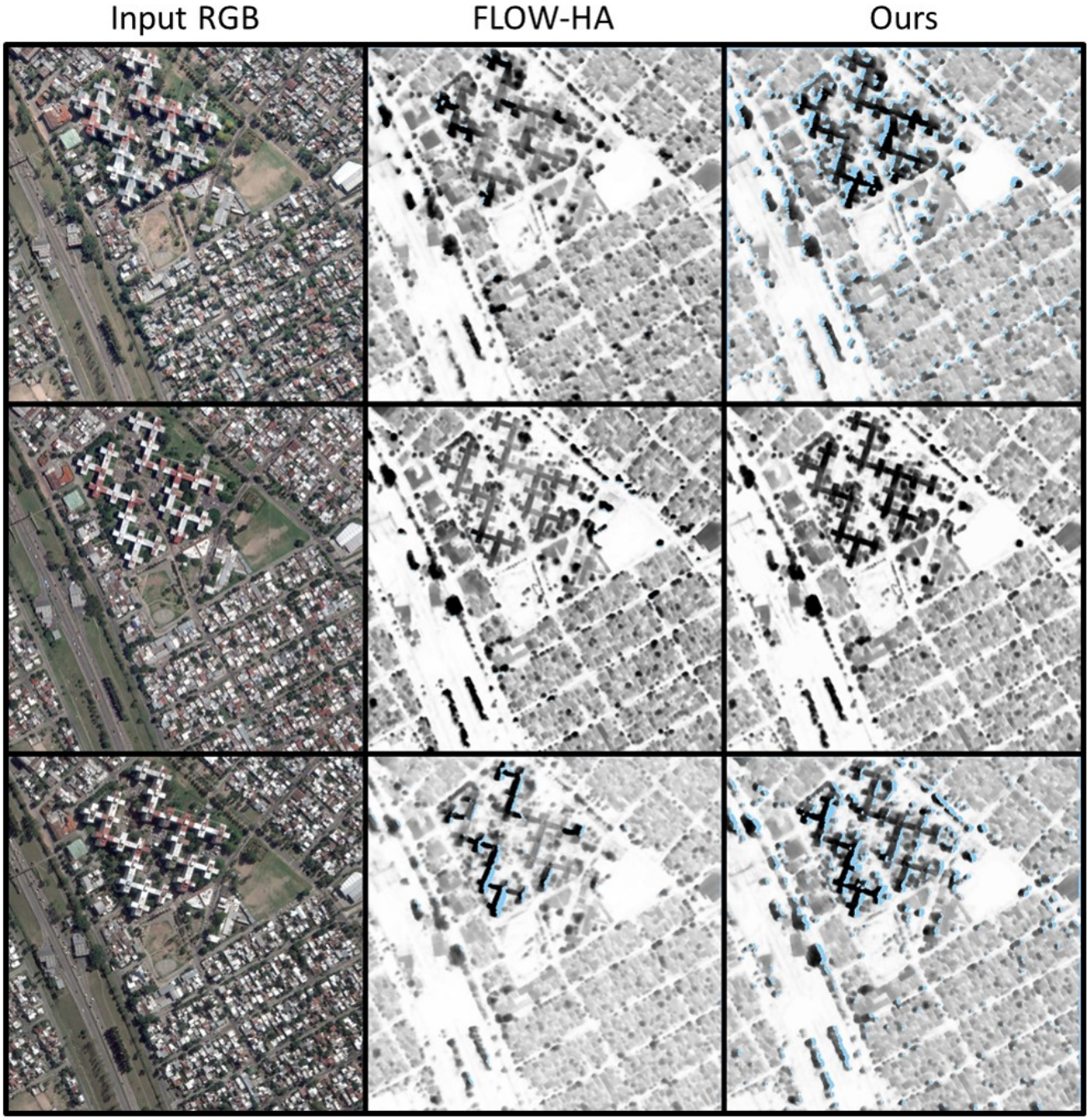

Figure 7: Rectified height images produced using predictions from FLOW-HA [2] and our model are compared for ARG tile 31, images 1, 10, and 12. Darker shades of gray represent larger height values. Occluded pixels are blue.

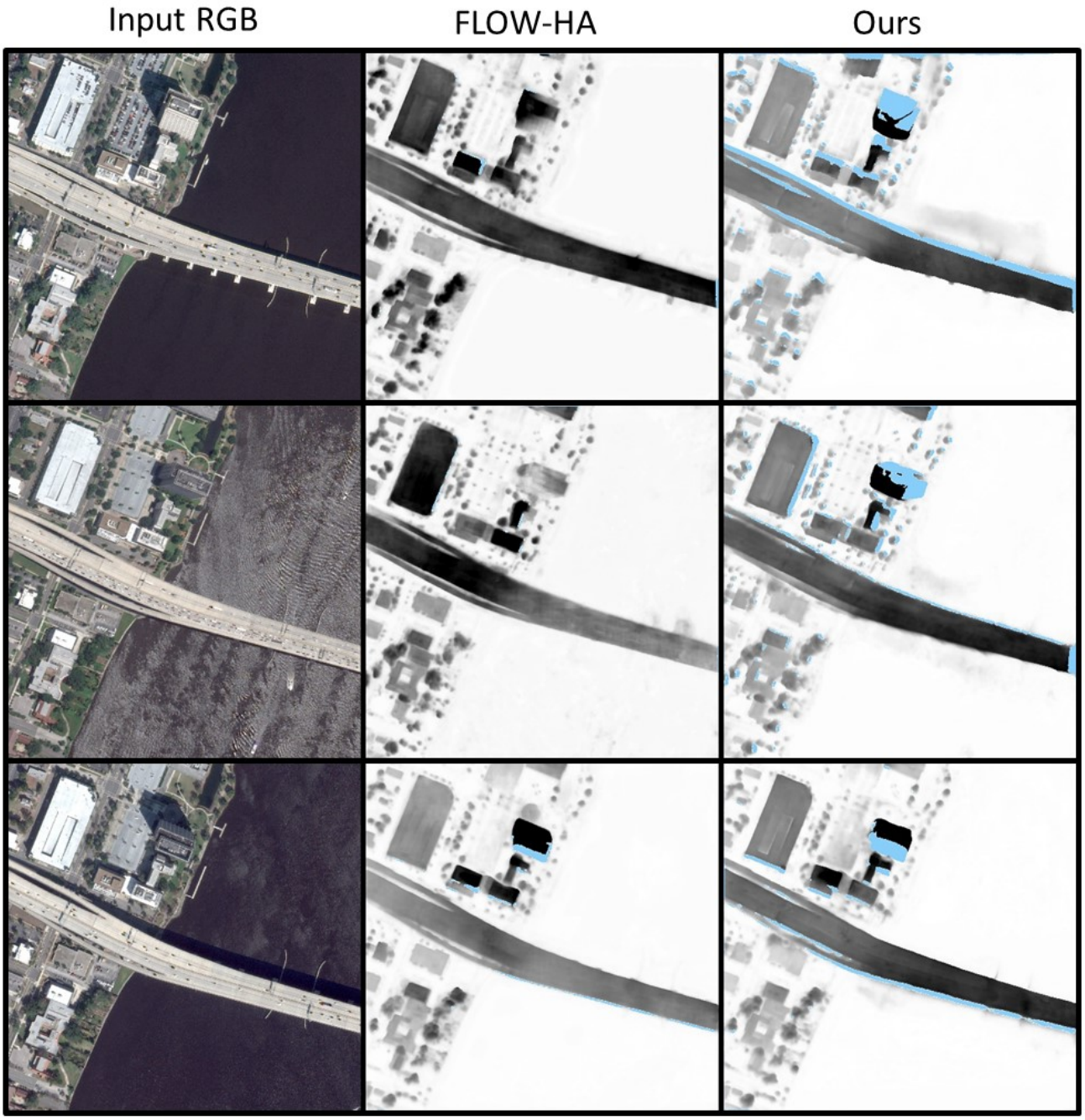

Figure 8: Rectified height images produced using predictions from FLOW-HA [2] and our model are compared for JAX tile 210, images 10, 12, and 20. Darker shades of gray represent larger height values. Occluded pixels are blue.

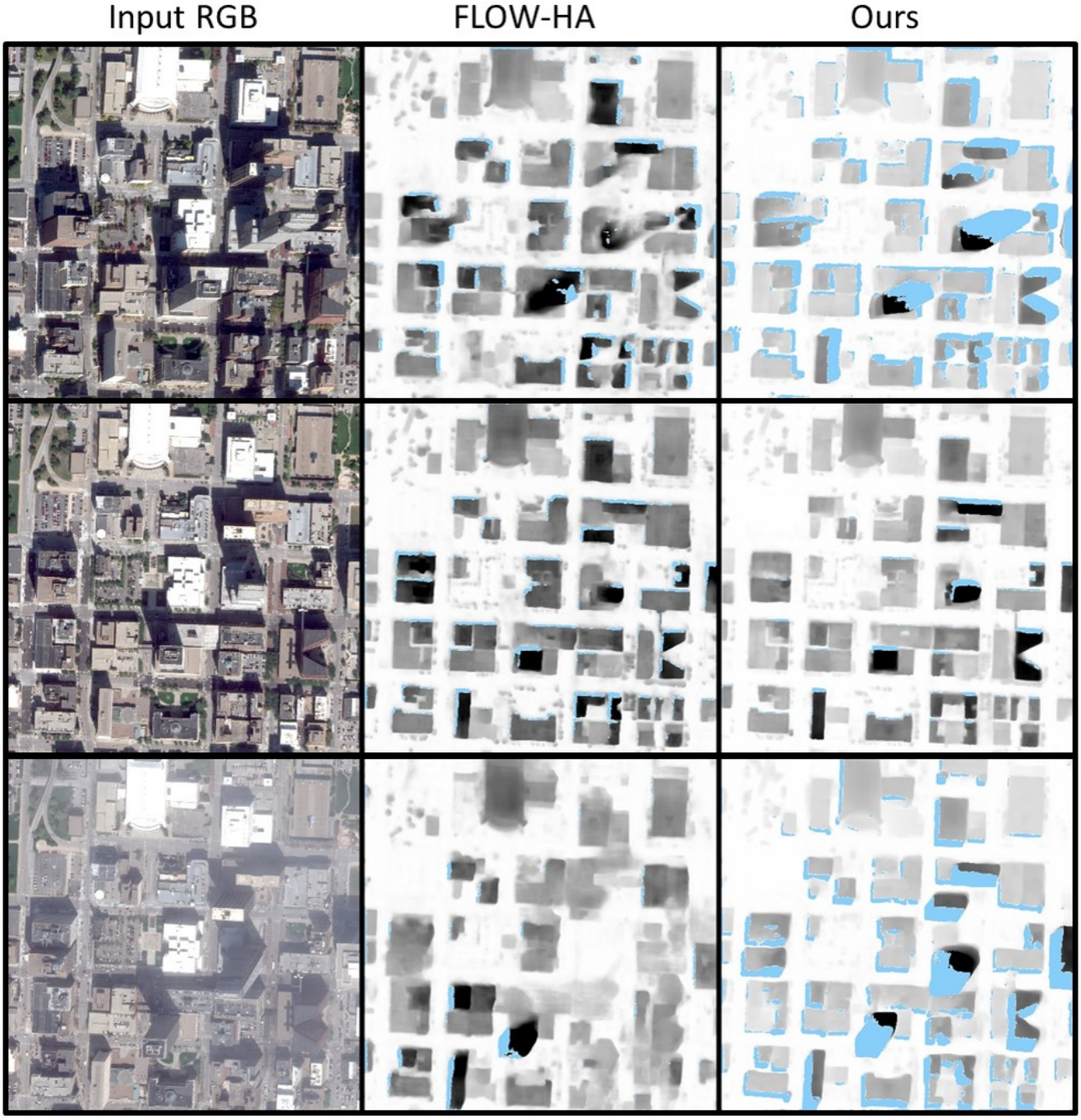

Figure 9: Rectified height images produced using predictions from FLOW-HA [2] and our model are compared for OMA tile 285, images 30, 35, and 39. Darker shades of gray represent larger height values. Occluded pixels are blue.

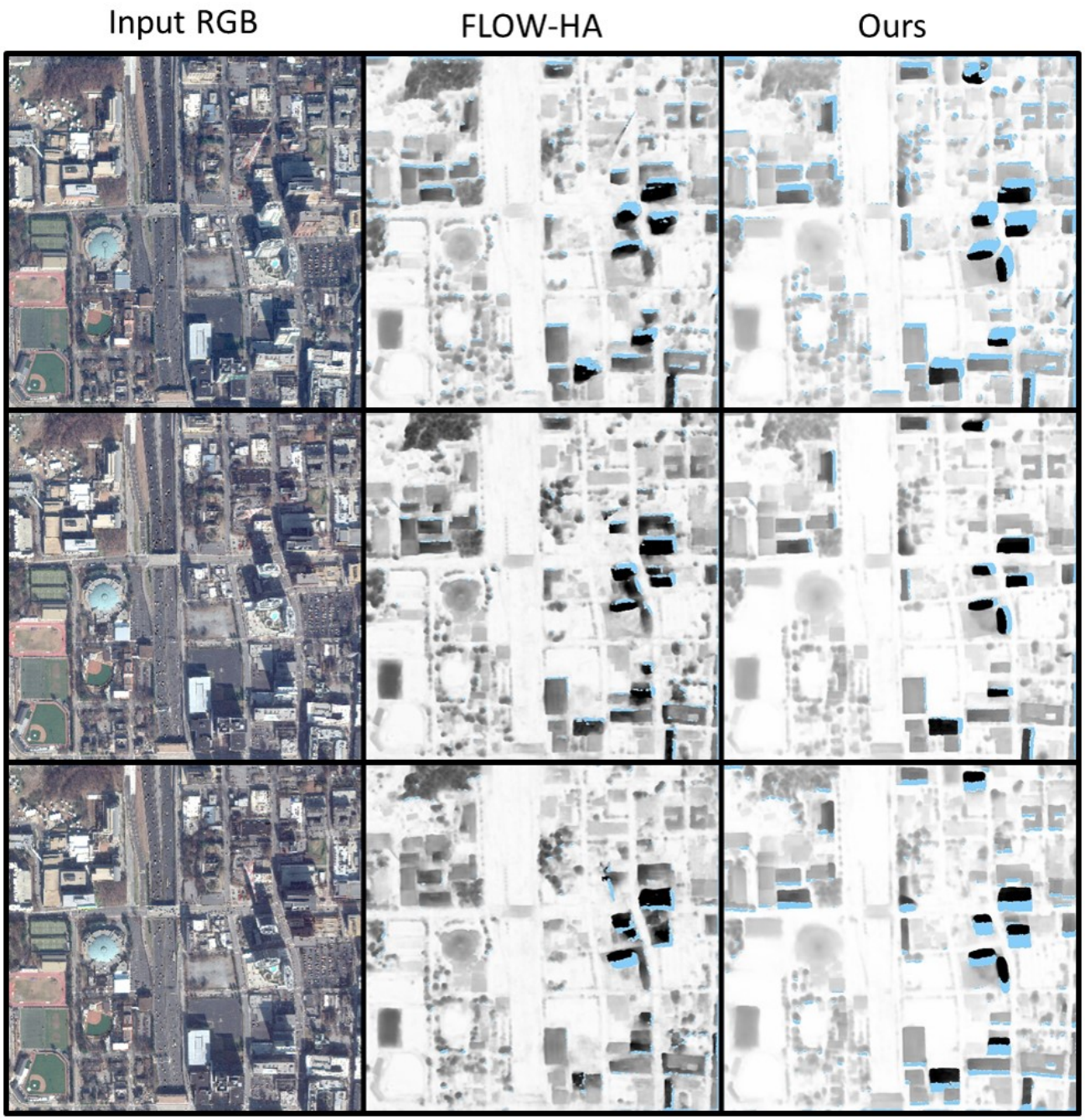

Figure 10: Rectified height images produced using predictions from FLOW-HA [2] and our model are compared for ATL tile 430, images 14, 26, and 37. Darker shades of gray represent larger height values. Occluded pixels are blue. Height of the tall antenna is captured in the FLOW-HA height predictions, though without sufficient accuracy for practical rectification. In our training, we apply a median filter to reference lidar height values to ignore tall narrowly occluding features.

Figure 11: Rectified height images produced using predictions from our model are shown for OMA tile 286, images 10-17, with varying seasons and increasing amounts of occluding smoke, top left to bottom right. Heavy smoke results in very inaccurate building height predictions, but performance appears to degrade gradually with the amount of occluding smoke. Darker shades of gray represent larger height values. Occluded pixels are blue.