Learning Geocentric Object Pose in Oblique Monocular Images

Gordon Christie[1,*] Rodrigo Rene Rai Munoz Abujder[1,*] Kevin Foster[1] Shea Hagstrom[1]
Gregory D. Hager[2] Myron Z. Brown[1]
[1]The Johns Hopkins University Applied Physics Laboratory [2]The Johns Hopkins University
{gordon.christie,rai.munoz,kevin.foster, shea.hagstrom, myron.brown}@jhuapl.edu
hager@cs.jhu.edu

Abstract

An object's geocentric pose, defined as the height above ground and orientation with respect to gravity, is a powerful representation of real-world structure for object detection, segmentation, and localization tasks using RGBD images. For close-range vision tasks, height and orientation have been derived directly from stereo-computed depth and more recently from monocular depth predicted by deep networks. For long-range vision tasks such as Earth observation, depth cannot be reliably estimated with monocular images. Inspired by recent work in monocular height above ground prediction and optical flow prediction from static images, we develop an encoding of geocentric pose to address this challenge and train a deep network to compute the representation densely, supervised by publicly available airborne lidar. We exploit these attributes to rectify oblique images and remove observed object parallax to dramatically improve the accuracy of localization and to enable accurate alignment of multiple images taken from very different oblique viewpoints. We demonstrate the value of our approach by extending two large-scale public datasets for semantic segmentation in oblique satellite images. All of our data and code are publicly available[1].

1. Introduction

In this paper, we study the problem of rectifying oblique monocular images from overhead cameras to remove observed object parallax with respect to ground, enabling accurate object localization for Earth observation tasks including semantic mapping [6], map alignment [32, 3], change detection [7], and vision-aided navigation [11]. Current state-of-the-art methods for these tasks focus on near-nadir images without the confounding effect of parallax; however, the vast majority of overhead imagery is oblique. For response to natural disasters and other dynamic world events, often only oblique images can be made available in a timely manner. The ability to rectify oblique monocular images to remove parallax will enable a dramatic increase in utility of these methods to address real-world problems.

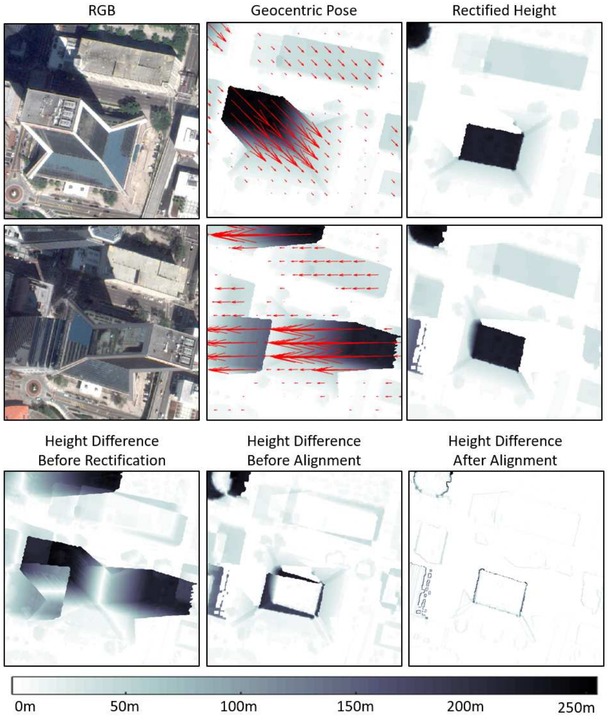
RGB
Geocentric Pose
Rectified Height
Height Difference Before Rectification
Height Difference Before Alignment
Height Difference After Alignment
0m
50m
100m
150m
200m
250m

Figure 1: Our method takes monocular RGB images, predicts object height (meters) and geocentric pose, and rectifies height to geospatially accurate 3D models suitable for reliable alignment by a conventional method.

To address this very challenging problem, we first draw inspiration from Gupta et al. [12] who proposed geocentric pose, or height above ground and orientation with respect to gravity, as a powerful representation to impose real-world structure on object detection, segmentation, and localization tasks using RGBD images. Hand-crafted features based on

[1] https://github.com/pubgeo/monocular-geocentric-pose
* denotes equal contribution

this idea have even featured prominently in state-of-the-art deep learning methods for indoor semantic segmentation [13, 20, 14, 4, 24, 18, 22, 30]. For close-range vision tasks, height and orientation have been derived directly from stereo-computed depth and more recently from monocular depth predicted by deep networks [17]. For long-range vision tasks such as Earth observation, depth cannot be reliably estimated with monocular images, so we further draw inspiration from recent work in monocular height above ground prediction [26, 21, 10, 1, 2, 16, 15, 33] and optical flow prediction from static images [23, 29, 28, 9]. We develop an encoding of geocentric pose and train a deep network to compute the representation densely. Our model jointly learns to predict height above ground and dense flow vectors mapping surface features to ground level. Figure 1 illustrates the use of our method to rectify overhead images taken from very different oblique viewpoints and then align the rectified height images – for this example, by affine homography. Height and flow for this example were derived from lidar, but example predictions from our model are shown in Section 4.3. While our experimental results are demonstrated for satellite images, we believe our method can also be successfully applied to airborne cameras and even ground-based cameras.

Our contributions are summarized as follows:

- We adopt geocentric pose as a general representation for geometry in oblique monocular images and are the first to report the following: 1) a method to supervise its learning, and 2) a method for prediction without reliance on depth estimates which cannot be reliably determined from monocular images at longer ranges.
- We extend the Urban Semantic 3D (US3D) dataset [2] to include labels for the geocentric pose task, enabling public research and comparative analysis of methods. We further extend US3D to include additional images with a wide range of oblique viewing angles from the SpaceNet 4 (SN4) contest [31] to enable more comprehensive parametric evaluation of this task.
- We demonstrate that our model designed to jointly learn height and orientation performs better than a model trained for each task independently, and increases efficiency through shared weights. We further demonstrate the need for rotation augmentations to overcome bias from severely limited viewpoint diversity due to sun-synchronous satellite orbits.
- We demonstrate the efficacy of our method for image rectification to improve intersection over union (IoU) scores for semantic segmentation with oblique images.
- All of our data and code are publicly available.

2. Related Work

Our approach draws inspiration from a large body of work exploiting object height and orientation with respect to ground to improve semantic segmentation and related tasks for RGBD images. Our encoding of this representation in a deep network is inspired by recent progress in predicting height above ground from single images and predicting optical flow from static images. Before introducing the details of our method, we review these motivating works.

2.1. Geocentric Pose

Gupta et al. [12] proposed geocentric pose – height and orientation with respect to ground – as a general feature for object recognition and scene classification. Gupta et al. [13] further proposed to encode horizontal disparity (or depth), height above ground, and orientation with respect to gravity as the popular three-channel HHA representation and demonstrated significant performance improvements for object detection, instance segmentation, and semantic segmentation tasks. Hand-crafted HHA features have since featured prominently even in deep learning state-of-the-art methods for indoor semantic segmentation [20, 14, 4, 24, 18, 22, 30] as well as object detection [20, 25] and semantic scene completion [19]. All of these works involve close-range indoor vision tasks and derive geocentric pose from depth, with height above ground approximated relative to the lowest point in an image [12]. In our work, we learn to predict these attributes directly in complex outdoor environments based on appearance without depth which is difficult to estimate reliably from images captured at long range. We also accurately predict absolute height above ground from monocular images. This is necessary for accurately rectifying the images, removing observed object parallax to improve accuracy of localization and enable accurate alignment of multiple images taken from very different oblique viewpoints.

2.2. Monocular Height Prediction

The successes of deep learning methods for monocular depth prediction [17] have motivated recent work to directly learn to predict height from appearance in a single image. The earliest work to our knowledge was conducted by Srivastava et al. (2017) who proposed a multi-task convolutional neural network (CNN) for joint height estimation and semantic segmentation of monocular aerial images [26]. Mou and Zhu (2018) also proposed a CNN for height estimation and demonstrated its use for instance segmentation of buildings [21]. Each of these early works was evaluated using a single overhead image mosaic from a single city. Ghamisi and Yokoya (2018) proposed a conditional generative adversarial network (cGAN) for image to height translation and reported results with a single image from each of three cities [10]. Amirkolaee and Arefi (2019) proposed a CNN trained with post-earthquake lidar and demonstrated its use to detect collapsed buildings by comparing model predictions for pre- and post-event images [1]. To promote research with larger-scale supervision, Bosch et al. (2019) produced the Urban Semantic 3D (US3D) dataset which includes sixty-nine satellite images over Jacksonville, FL and Omaha, NE, each covering approximately one hundred square kilometers [2]. Le Saux et al. (2019) leveraged this dataset to conduct the 2019 Data Fusion Contest focused on semantic 3D reconstruction, including a novel challenge track for single-view semantic 3D [16]. The winning solutions by Kunwar [15] and Zheng et al. [33] both exploited semantic labels as priors for height prediction. In this work, we demonstrate comparable accuracy without semantic priors. We also show improved height prediction accuracy by jointly learning to predict orientation flow vectors. In addition to our experiments, we leverage and extend the US3D dataset using public satellite images from the 2018 SpaceNet 4 (SN4) contest that span a wide range of viewing angles over Atlanta, GA [31], and we demonstrate that our method to predict geocentric pose significantly improves building segmentation accuracy for oblique images.

2.3. Optical Flow Prediction from a Static Image

Our approach to learning geocentric pose is inspired by recently demonstrated methods to predict dense optical flow fields from static images with self-supervision from optical flow methods applied to videos. Pintea et al. (2014) proposed regression of dense optical flow fields from static images using structured random forests [23]. Walker (2015) proposed a CNN for ordinal regression to better generalize over diverse domains [29]. Walker et al. (2016) proposed a generative model using a variational auto-encoder (VAE) for learning motion trajectories from static images [28]. Gao et al. (2018) also explored a generative model using a cGAN but reported state-of-the-art results for optical flow prediction and action recognition with their Im2Flow regression model, a modified U-Net CNN encoder/decoder trained by minimizing both a pixel L2 loss and a motion content loss derived from a separate action recognition network that regularizes the regression network to produce realistic motion patterns [9]. To learn geocentric pose, we employ a similar U-Net architecture and demonstrate improved performance by jointly learning to predict height. We also highlight orientation bias for our task by performing rotation augmentations during training. We produce reference flow fields for supervision automatically using lidar as discussed in Section 3.3.

3. Learning Geocentric Pose

3.1. Representation

Our representation of geocentric pose encodes height above ground and flow vectors that map surface features to ground level. A satellite pushbroom sensor model is well-approximated locally by affine projection which preserves the invariant property of parallelism [5]. We exploit this property in representing flow fields with pixel-level magnitudes and image-level orientation. Similar to [9], we represent orientation (θ) as a two-element vector, $[\sin(\theta),\ \cos(\theta)]$, representing the horizontal and vertical components of the flow vectors. We observe that each feature's height above ground is intrinsic and the magnitude of its flow vector is related to that height by each image's projection. We thus employ height as a prior in our model for learning magnitude.

3.2. Model

Our model, illustrated in Figure 2, jointly predicts image-level orientation, as well as dense above-ground-level heights and flow vector magnitudes. The base architecture utilizes a U-Net decoder with a ResNet34 encoder. At the last layer of the encoder, the image-level orientation is predicted as $\sin(\theta)$ and $\cos(\theta)$. The output of the decoder is used to predict heights, which are concatenated with the decoder output for predicting magnitudes. MSE is used for all output heads (image-level orientation, magnitude, and height), where each loss is weighted equally during training. At test time, flow vectors can be calculated by multiplying the predictions of image-level orientation and per-pixel magnitudes. We present an ablation study where height prediction is removed from the model to show its importance for learning to predict orientation and magnitude. Height is intrinsic to objects in the image, where pixels representing the same physical locations on a building in different images should have the same heights. However, magnitudes for these pixels will vary with changes to viewing geometry. We believe the intrinsic properties of height provide valuable context for predicting magnitude. We also show that the accuracy of our height predictions is comparable to state-of-the-art solutions for a public challenge dataset, and note that our network shares weights for multiple tasks, making it more efficient than having separate networks for each task.

3.3. Supervision

To enable supervised learning of our model, we have developed a pipeline for producing non-overlapping overhead RGB image tiles with lidar-derived attributes projected into each oblique image pixel, as illustrated in Figure 3. We utilized this pipeline to produce training and test datasets for our task, augmenting public data from US3D [2] and SN4 [31]. For each geographic tile, we first align each overhead image with lidar intensity using the mutual information metric and update the image translation terms in the RPC camera metadata [5]. To improve reliability of image matching, we cast shadows in each lidar intensity image using solar angle image metadata to match the shad-

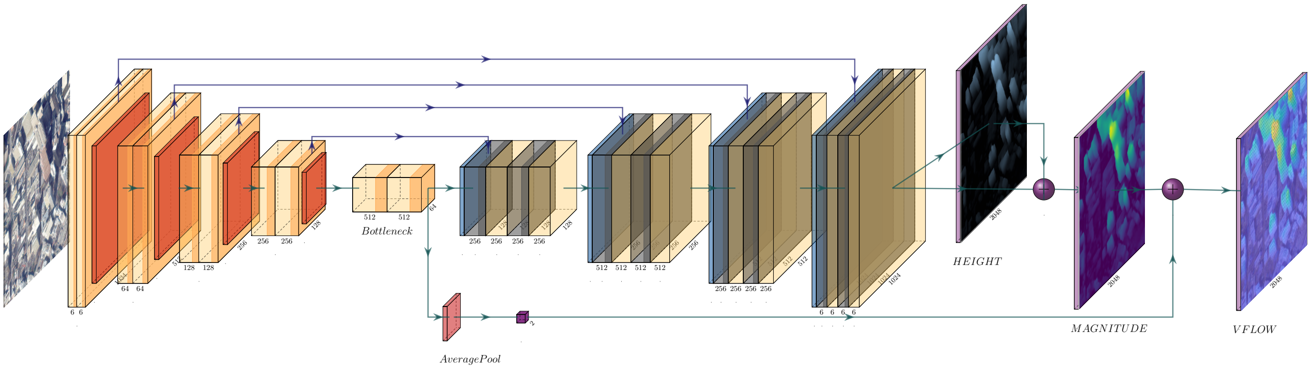

Figure 2: This shows the architecture of our full approach, which uses a U-Net decoder with a ResNet34 encoder. At the last layer of the encoder, we predict the image-level orientation as $\sin(\theta)$ and $\cos(\theta)$. At the output of the decoder, we predict per-pixel above-ground-level height values, which are concatenated with the decoder's output and used to predict per-pixel magnitudes. An MSE loss is used for all output heads. At test time, flow vectors can be calculated by multiplying image-level orientation predictions with the per-pixel magnitudes.

ows observed in the RGB image. Layers produced include UTM geographic coordinates, ground-level height from the Digital Terrain Model (DTM), surface-level height from the Digital Surface Model (DSM), height above ground computed from the difference of the DSM and DTM, the shadow mask produced for image matching, and image flow vectors mapping surface-level feature pixels to their ground-level pixel coordinates. Our representation of geocentric pose is composed of height above ground and orientation with respect to ground as defined by the dense flow vectors. Both rely on knowledge of ground level in the DTM. For the lidar data used in our experiments, DTM layers were produced by professional surveyors with manual edits, but automated methods for ground classification in lidar and even in DSMs produced using satellite images also work well [8].

For our experiments, we also employ semantic labels derived from public map data to demonstrate the value of our model for rectifying map features in oblique images. We project this map data into each image with the same procedure used for lidar attributes. Layers include semantic label for each pixel and ground-level footprints for buildings. Building facades are labeled separately from roofs.

4. Experiments

4.1. Datasets

For our experiments, we extended two publicly-available datasets – US3D [2] and SN4 [31] – using the method described in Section 3.3 and illustrated in Figure 3. We train with the full resolution for each dataset.

- DFC19. We use the same 2,783 training images and 50 testing images of Jacksonville, FL and Omaha, NE from US3D used for the 2019 Data Fusion Contest [16]. We also use an extended test set with 300 images including more view diversity for the same geographic tiles. Images are each 2048x2048 pixels.
- ATL-SN4. We produced 25,500 training images and 17,554 testing images of Atlanta, GA using public unrectified source images to closely match the rectified image tiles used for SN4, as shown in Figure 4. We used 7,702 training images and 310 testing images, cropped to 1024x1024 pixels, for our experiments.

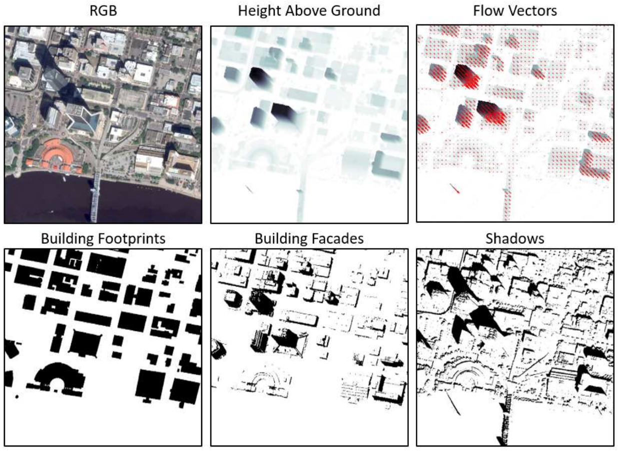

Figure 3: Lidar-derived attributes for each RGB image include height above ground, geocentric pose flow vectors, and shadow masks. Map attributes include semantic labels, building facades, and ground-level building footprints.

Viewpoint diversity and pixel resolution for images in the DFC19 and ATL-SN4 datasets are shown in Figure 5. Jacksonville and Omaha images were collected by MAXAR's WorldView-3 satellite on multiple dates with a variety of azimuth angles and limited off-nadir an-

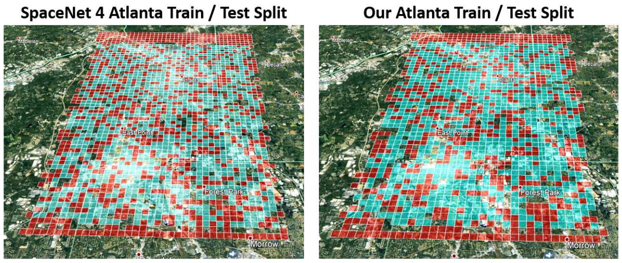
SpaceNet 4 Atlanta Train / Test Split
Our Atlanta Train / Test Split

Figure 4: Train (blue) and test (red) tiles for the ATL-SN4 unrectified images (right) were selected to closely match the split for SpaceNet 4 orthorectified image tiles (left). Images shown are from Google Earth.

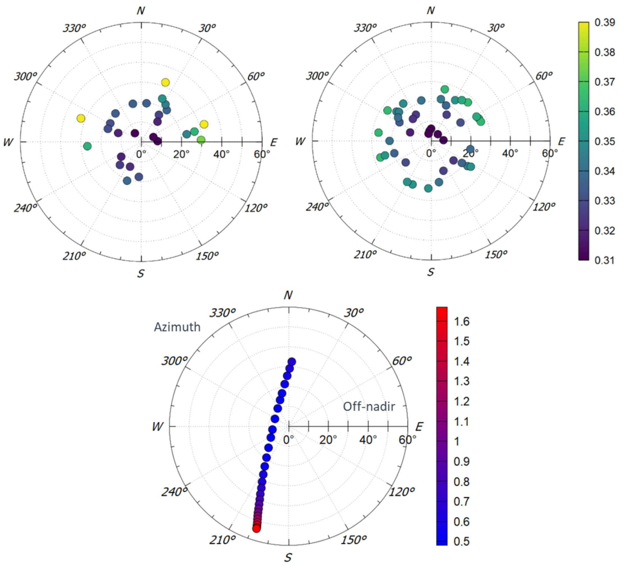

Figure 5: Azimuth angle, off-nadir angle, and resolution (meters) is shown for Jacksonville and Omaha images from DFC19 (top) and ATL-SN4 images (bottom).

gles. ATL-SN4 images were collected by MAXAR's WorldView-2 satellite during a single orbit with very limited azimuth diversity and a wide range of off-nadir angles. Together, these datasets enable thorough evaluation.

4.2. Methods

Flow Vector Regression For each test set, we present four sets of results. These include combinations of models trained with and without height supervision, and with and without train-time rotation augmentations. As discussed earlier, our datasets consist of orientation bias because of the sun-synchronous satellite orbits. To make our model generalizable to unseen orientations, we perform train-time flips and rotations randomly, which can introduce new orientation ground truth for each image at different epochs during training. Our approaches are described as follows:

- **FLOW** Model with height prediction head removed and trained without augmentations.
- **FLOW-H** Full model trained without augmentations.
- **FLOW-A** FLOW trained with augmentations.
- **FLOW-HA** FLOW-H trained with augmentations.

For completeness, we present image-level orientation (angle) and pixel-level magnitude (mag) errors for our predictions, as they are learned separately during training. Orientation errors are measured in degrees, while magnitude errors are measured in pixels. However, we note that orientation and magnitude are typically not appropriate metrics for this task. As an example, in a nadir image where all pixel magnitudes are zero, predicting the orientation is meaningless. Similarly, in a highly-oblique image where the magnitudes are high, it is extremely important to predict the orientation accurately. We therefore measure per-pixel endpoint errors (EPE), which measure the Euclidean distance between the endpoints of the predicted and ground truth flow vectors. However, note that mag errors are equal to EPE when orientation is known from the sensor metadata, which is sometimes the case with satellite imagery. Therefore, mag errors can be an appropriate metric when orientation is known.

These metrics are calculated with and without test-time rotations to show how models that do not include train-time rotations over-fit to the limited set of orientations in the train set. We also calculate per-category EPE to show how semantics affect performance. Categories from DFC19 are used, as well as a separate layer with shadow masks.

Building Footprint Extraction One of the goals of this work is to enable more accurate automated-mapping from overhead imagery. With our flow vector predictions, outputs from any segmenter or detector can be input into our model and transformed to ground level. To demonstrate the accuracy of our model, we use building annotations and footprints from the DFC19 and ATL-SN4 test sets. Building annotations consist of the roof and facade labels in the image, while the footprints represent the base of the building identified from top-down lidar. Using our predicted flow vectors, we warp the building annotations to ground level and compare to the ground truth footprints.

We also demonstrate the reverse capability, where we start with footprints and warp them into building annotations using our predicted flow vectors. This is useful in situations where there is a desire to overlay map data (e.g., OpenStreetMap) on imagery as an initial set of annotations. For example, when a new image is captured of an area actively being developed, we may want to pull in existing annotations so annotators do not start from scratch.

We compare three results for each of the two tasks: 1) transform building annotations to footprints, and 2) transform footprints to building annotations. First, we measure

IoU between the building annotations and the footprints to understand what the accuracy is when we do nothing. Second, we warp the source mask (building annotations or footprints) to the target mask using the ground truth flow vectors to get an upper bound for the IoU on what can be achieved if we perfectly predict the flow vectors. Note that we do not get perfect overlap in this case because of occluded ground pixels. Finally, we measure IoU for the warped versions of the source masks using our predicted flow vectors.

4.3. Results

Height Prediction We assess our current method, which takes the height outputs of FLOW-H, compared to two recent strong baselines [15, 33] for the very challenging DFC19 test set [16], measuring mean and root mean square (RMS) error (meters) for height predictions compared to above ground height measured from lidar. Results are shown in Table 1. Both baseline methods anchor height predictions using semantic category, and both exploit test-time ensembles to improve performance. While semantic anchors appear to improve accuracy for categories with low height variance, they do not account for the variance observed in urban scenes. Our model performs better overall without semantic priors or test-time ensembles.

Figure 6 depicts building height statistics for the train and test sets, with some building heights approaching 200 meters. Achieving more reliable predictions for those rare tall objects is a topic for ongoing research. Height prediction performance in the presence of significant terrain relief has also yet to be characterized. Statistics for ground-level terrain height variation in the DFC19 and ATL-SN4 data sets are shown in Figure 7.

	mean	mean bldgs	RMS	RMS bldgs
Kunwar [15]	**2.69**	8.33	9.26	19.65
Zheng et al. [33]	2.94	8.72	9.24	19.32
Ours	2.98	**7.73**	**8.23**	**16.87**

Table 1: Our regression model produces height predictions with lower RMS error (meters) than baseline models that anchor height predictions with semantic category.

Flow Vector Regression Our results for each of the approaches on the DFC19 test set without test-time augmentations can be seen in Table 2. The results from the same approaches applied to the test set containing rotation augmentations are shown in Table 3. The per-category results are EPE. Results in shadows, which are a separate layer (i.e., not included as part of the DFC category layer) are also included. Table 4 and Table 5 show similar results for ATL-SN4, but exclude a semantic breakdown, as the same human-validated semantic labels are not available for this dataset. The test sets consist of the original DFC19 and ATL-SN4 test sets along with 9 additional rotation augmentations per image at intervals of 36 degrees.

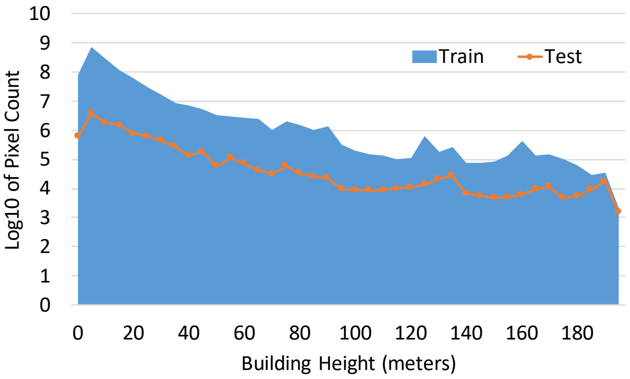

Figure 6: Height distributions in train and test sets are comparable, with some buildings approaching 200 meters.

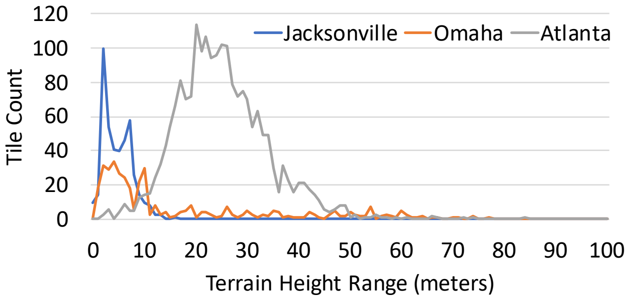

Figure 7: Ground-level terrain height variation statistics.

Two key observations can be made about these results. 1) It is clear from Table 3 and Table 5 that models trained without rotation augmentations over-fit to the orientation bias of the train set, and that train-time rotation augmentations are currently needed to create generalizable models for this task. 2) Jointly learning to predict above-ground-height improves metrics across most categories when test-time rotations are applied. Unsurprisingly, we observe the lowest EPE values for ground pixels, and some of the highest EPE errors on facades, roofs, and elevated roads, where ground truth magnitudes are highest.

We show the importance of train-time rotations qualitatively in Figure 8. In the first column, where no test-time rotation was performed, we can qualitatively observe similar performance between FLOW-H and FLOW-HA. However, in the second column, when we rotate the image to an orientation not originally represented in the train set, we see FLOW-H qualitatively performing worse than FLOW-HA.

Building Footprint Extraction In this section, we demonstrate the ability to transform semantic segmentations in the image space to ground-level map data, as well as pulling map data into imagery. Table 6 and Table 7 show

Method	mag	angle	EPE	ground	veg	roof	water	elevated roads	facade	shadow
FLOW	2.71	16.11	3.08	1.39	**3.68**	**5.44**	1.78	6.86	**7.11**	**4.03**
FLOW-H	**2.40**	16.14	**2.92**	**0.92**	3.86	5.70	1.54	**6.42**	7.37	3.98
FLOW-A	2.91	17.52	3.24	1.15	4.04	6.17	1.57	7.66	8.32	4.42
FLOW-HA	2.69	**15.09**	3.04	1.06	4.06	5.89	**1.41**	6.89	7.83	4.25

Table 2: Results **without** test-time rotations for DFC19. Lower is better for all numbers. Per-category values are all end point errors (EPE). This table highlights that models trained to generalize perform worse than models that learn the orientation bias of the train set. However, we note that the model trained without rotation augmentations and with height supervision has the best overall EPE.

Method	mag	angle	EPE	ground	veg	roof	water	elevated roads	facade	shadow
FLOW	4.15	79.52	6.11	2.39	7.34	11.99	3.01	12.67	13.80	7.50
FLOW-H	4.07	78.15	5.95	2.06	7.29	12.18	2.94	12.82	13.86	7.35
FLOW-A	3.02	17.48	3.35	1.18	**4.12**	6.22	1.56	8.06	8.35	4.51
FLOW-HA	**2.83**	**16.79**	**3.21**	**1.10**	4.17	**6.10**	**1.44**	**7.55**	**8.08**	**4.42**

Table 3: Results **with** test-time rotations for DFC19. Lower is better for all numbers. Per-category values are all end point errors (EPE). This table highlights that train-time rotation augmentations are currently needed to overcome the orientation bias caused by sun-synchronous satellite orbits and perform well in the presence of test-time rotations. These results also highlight that training with height supervision improves overall EPE performance over most categories. These improvements are most notable for roof, elevated roads, and facades, where accurate flow vector prediction is more important.

Method	mag	angle	EPE
FLOW	3.88	9.64	4.17
FLOW-H	**3.78**	**7.38**	**3.99**
FLOW-A	5.37	15.76	6.03
FLOW-HA	4.79	16.57	5.38

Table 4: Results **without** test-time rotations for ATL-SN4. Similar to Table 2, we see that the model trained without rotation augmentations, but with height supervision, performs best when the test set contains orientation bias.

Method	mag	angle	EPE
FLOW	6.04	77.31	8.79
FLOW-H	6.30	81.34	9.04
FLOW-A	4.81	**15.77**	5.39
FLOW-HA	**4.22**	23.19	**5.15**

Table 5: Results **with** test-time rotations for ATL-SN4. Similar to Table 3, we see that train-time rotations and height supervision are important when test-time rotations are applied.

IoU for DFC19 and ATL-SN4, respectively. Unrectified is the comparison between the building annotations and the footprints without warping. Ours is the comparison between warped versions of the original mask and target mask using the predicted flow vectors. GT follows the same process as Ours, but with the ground truth flow vectors.

As seen from Table 6 and Table 7, our results better capture the footprints in these datasets than the original building annotations. Note that occluded pixels prevent GT from reaching an IoU score of 1. GT represents an upper bound on what can be achieved with perfect flow vector prediction.

	Building to Footprint	**Footprint to Building**
Unrectified	0.78 (92.9%)	0.78 (90.7%)
Ours	0.83 (98.8%)	0.82 (95.3%)
GT	0.84	0.86

Table 6: IoU and percentage of GT for transforming building annotations to footprints and vice versa for DFC19.

	Building to Footprint	**Footprint to Building**
Unrectified	0.74 (89.2%)	0.74 (86.0%)
Ours	0.76 (91.6%)	0.77 (89.5%)
GT	0.83	0.86

Table 7: IoU and percentage of GT for transforming building annotations to footprints and vice versa for ATL-SN4.

Map Alignment Rectifying semantic labels to ground level simplifies the task of aligning maps and oblique im-

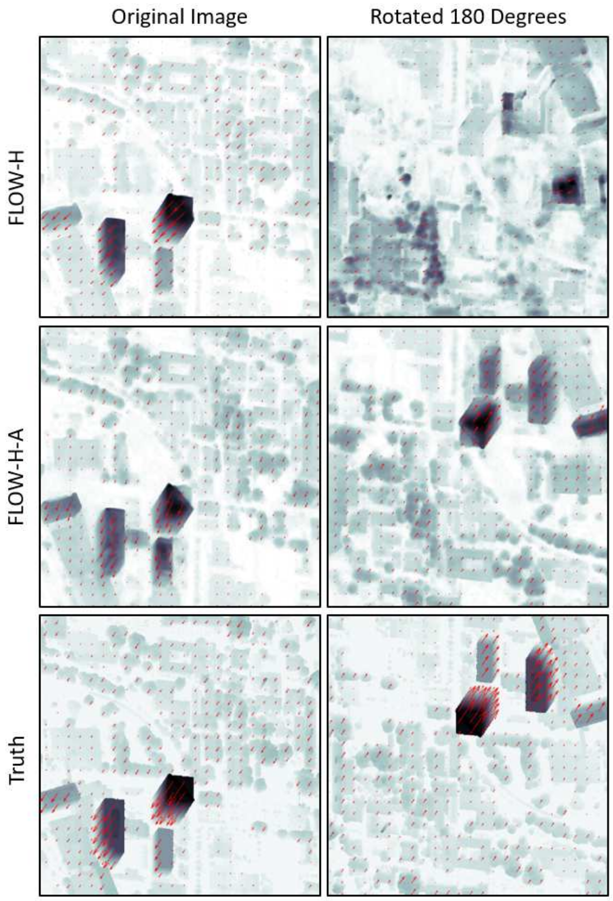

Figure 8: Orientation augmentations in training our model help to reduce bias in the satellite viewing angles. Height and flow vector ground truth and predictions from models trained with and without augmentations are shown for an example from ATL-SN4.

ages as shown in Figure 1. To demonstrate this, we apply the MATLAB imregdemons function, an efficient implementation of non-parametric image registration [27], to estimate dense displacement fields between pairs of images in the DFC19 test set. We do this for aligning RGB images as a baseline and then for rectified height images to demonstrate improved alignment. Table 8 shows mean IoU scores for reference building segmentation labels rectified to ground level and compared with the reference footprints after alignment. Mean IoU is significantly improved, and the fraction of images with IoU greater than 0.5 is significantly improved.

5. Discussion

In this paper, we have introduced the novel task of learning geocentric pose, defined as height above ground and orientation with respect to gravity, for above-ground objects in oblique monocular images. While we have shown the value of this representation for rectifying above-ground features in oblique satellite images, we believe that with minor modifications our method can also be successfully applied to airborne cameras and even ground-based cameras to address a broad range of outdoor mapping, change detection, and vision-aided navigation tasks for which a single ground plane cannot be assumed.

	Mean	**IoU > 0.5**
Unaligned	0.46	0.40
RGB aligned	0.66	0.85
FLOW-HA	0.69	0.93
FLOW-HA fixed angle	**0.69**	**0.94**

Table 8: IoU values for transforming per-pixel building annotations to footprints in other overlapping images.

Much of the prior work on geocentric pose has focused on its exploitation as hand-crafted features for semantic segmentation. In this work, we have focused on its exploitation to rectify building segmentations to ground level, enabling geospatially accurate mapping with oblique images. Similar to much prior work with the HHA representation, we expect that our representation will also provide an effective prior for regularizing semantic segmentation predictions.

While our current results clearly indicate the efficacy of the proposed method, much remains unexplored. We expect that more explicitly employing intuitive cues such as shadows and building facades will help reduce prediction error for the height variation observed in urban scenes. Further, while our rotation augmentations help account for orientation bias in satellite images, we expect that more fully accounting for true geometry and appearance variation will help address current observed failure cases. We plan to explore these ideas in future work, and we will publicly release all of our code and data.

Acknowledgments

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) contract no. 2017-17032700004. This work was further supported by the National Geospatial-Intelligence Agency (NGA) and approved for public release, 20-316, with distribution statement A – approved for public release; distribution is unlimited. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, NGA, or the U.S. Government.

References

[1] Hamed Amini Amirkolaee and Hossein Arefi. CNN-based estimation of pre-and post-earthquake height models from single optical images for identification of collapsed buildings. *Remote Sensing Letters*, 2019. 2, 3

[2] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic Stereo for Incidental Satellite Images. In *WACV*, 2019. 2, 3, 4

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. AutoCorrect: Deep Inductive Alignment of Noisy Geometric Annotations. *BMVC*, 2019. 1

[4] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-Sensitive Deconvolution Networks with Gated Fusionfor RGB-D Indoor Semantic Segmentation. In *CVPR*, 2017. 2

[5] Carlo de Franchis, Enric Meinhardt-Llopis, Julien Michel, J-M Morel, and Gabriele Facciolo. On stereo-rectification of pushbroom images. In *ICIP*, 2014. 3

[6] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In *CVPRW*, 2018. 1

[7] Jigar Doshi, Saikat Basu, and Guan Pang. From Satellite Imagery to Disaster Insights. *NeurIPS Workshops*, 2018. 1

[8] Liuyun Duan, Mathieu Desbrun, Anne Giraud, Frédéric Trastour, and Lionel Laurore. Large-Scale DTM Generation From Satellite Data. In *CVPRW*, 2019. 4

[9] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2Flow: Motion Hallucination from Static Images for Action Recognition. In *CVPR*, 2018. 2, 3

[10] Pedram Ghamisi and Naoto Yokoya. IMG2DSM: Height Simulation From Single ImageryUsing Conditional Generative Adversarial Net. *IEEE Geoscience and Remote Sensing Letters*, 2018. 2

[11] Hunter Goforth and Simon Lucey. GPS-Denied UAV Localization using Pre-existing Satellite Imagery. In *ICRA*, 2019. 1

[12] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In *CVPR*, 2013. 1, 2

[13] Saurabh Gupta, Ross Girshick, Pablo Arbelaez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, 2014. 2

[14] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross Modal Distillation for Supervision Transfer. In *CVPR*, 2016. 2

[15] Saket Kunwar. U-Net Ensemble for Semantic and Height Estimation Using Coarse-Map Initialization. In *IGARSS*, 2019. 2, 3, 6

[16] Bertrand Le Saux, Naoto Yokoya, Ronny Hansch, Myron Brown, and Greg Hager. 2019 Data Fusion Contest [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 2019. 2, 3, 4, 6

[17] Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *CVPR*, 2018. 2

[18] Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng-Ann Heng, and Hui Huang. Cascaded Feature Network for Semantic Segmentation of RGB-D Images. In *ICCV*, 2017. 2

[19] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and Think: Disentangling Semantic Scene Completion. In *NeurIPS*, 2018. 2

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*, 2015. 2

[21] Lichao Mou and Xiao Xiang Zhu. IM2HEIGHT: Height Estimation from Single Monocular Imagery via Fully Residual Convolutional-Deconvolutional Network. *arXiv:1802.10249*, 2018. 2

[22] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. RDFNet: RGB-D Multi-level Residual Feature Fusion forIndoor Semantic Segmentation. In *ICCV*, 2017. 2

[23] Silvia L Pintea, Jan C van Gemert, and Arnold WM Smeulders. Déja vu. In *ECCV*, 2014. 2, 3

[24] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3D Graph Neural Networks for RGBD Semantic Segmentation. In *ICCV*, 2017. 2

[25] Max Schwarz, Anton Milan, Arul Selvam Periyasamy, and Sven Behnke. RGB-D Object Detection and Semantic Segmentation for Autonomous Manipulation in Clutter. *IJRR*, 2018. 2

[26] Shivangi Srivastava, Michele Volpi, and Devis Tuia. Joint height estimation and semantic labeling of monocular aerial images with cnns. In *IGARSS*, 2017. 2

[27] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 2009. 8

[28] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An Uncertain Future: Forecasting from Static Images using Variational Autoencoders. In *ECCV*, 2016. 2, 3

[29] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense Optical Flow Prediction from a Static Image. In *ICCV*, 2015. 2, 3

[30] Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-D Segmentation. In *ECCV*, 2018. 2

[31] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar, and Hanlin Tang. SpaceNet MVOI: a Multi-View Overhead Imagery Dataset. In *ICCV*, 2019. 2, 3, 4

[32] Armand Zampieri, Guillaume Charpiat, Nicolas Girard, and Yuliya Tarabalka. Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In *ECCV*, 2018. 1

[33] Zhuo Zheng, Yanfei Zhong, and Junjue Wang. Pop-Net: Encoder-Dual Decoder for Semantic Segmentation and Single-View Height Estimation. In *IGARSS*, 2019. 2, 3, 6