Abstract

Detection and segmentation of objects in overhead imagery is a challenging task. The variable density, random orientation, small size, and instance-to-instance heterogeneity of objects in overhead imagery calls for approaches distinct from existing models designed for natural scene datasets. Though new overhead imagery datasets are being developed, they almost universally comprise a single view taken from directly overhead (“at nadir”), failing to address a critical variable: look angle. By contrast, views vary in real-world overhead imagery, particularly in dynamic scenarios such as natural disasters where first looks are often over 40° off-nadir. This represents an important challenge to computer vision methods, as changing view angle adds distortions, alters resolution, and changes lighting. At present, the impact of these perturbations for algorithmic detection and segmentation of objects is untested. To address this problem, we present an open source Multi-View Overhead Imagery dataset, termed SpaceNet MVOI, with 27 unique looks from a broad range of viewing angles (−32.5° to 54.0°). Each of these images cover the same 665 km² geographic extent and are annotated with 126,747 building footprint labels, enabling direct assessment of the impact of viewpoint perturbation on model performance. We benchmark multiple leading segmentation and object detection models on: (1) building detection, (2) generalization to unseen viewing angles and resolutions, and (3) sensitivity of building footprint extraction to changes in resolution. We find that state of the art segmentation and object detection models struggle to identify buildings in off-nadir imagery and generalize poorly to unseen views, presenting an important benchmark to explore the broadly relevant challenge of detecting small, heterogeneous target objects in visually dynamic contexts.

1. Introduction

Recent years have seen increasing use of convolutional neural networks to analyze overhead imagery collected by aerial vehicles or space-based sensors, for applications ranging from agriculture [18] to surveillance [39, 32] to land type classification [3]. Segmentation and object detection of overhead imagery data requires identifying small, visually heterogeneous objects (e.g. cars and buildings) with varying orientation and density in images, a task ill-addressed by existing models developed for identification of comparatively larger and lower-abundance objects in natural scene images. The density and visual appearance of target objects change dramatically as look angle, geographic location, time of day, and seasonality vary, further complicating the problem. Addressing these challenges will provide broadly useful insights for the computer vision community as a whole: for example, how to build segmentation models to identify low-information objects in dense contexts.

Though public overhead imagery datasets explore geographic and sensor homogeneity [8, 12, 22, 34, 19], they generally comprise a single view of the imaged location(s) taken nearly directly overhead (“at nadir”). Nadir imagery is not representative of collections during disaster response or other urgent situations: for example, the first public high-resolution cloud-free image of San Juan, Puerto Rico following Hurricane Maria was taken at 51.9° “off-nadir”, i.e., a 51.9° angle between the nadir point directly underneath the satellite and the center of the imaged scene [10]. The disparity between looks in public training data and relevant use cases hinders development of models applicable to real-world problems. More generally, satellite and drone images rarely capture identical looks at objects in different contexts, or even when repeatedly imaging the same geography. Furthermore, no existing datasets or metrics permit assessment of model robustness to different looks, prohibit-
Figure 1: Sample imagery from SpaceNet MVOI. Four of the 2222 geographically unique image chips in the dataset are shown (columns), with three of the 27 views of that chip (rows), one from each angle bin. Negative look angle corresponds to South-facing views, whereas positive look angles correspond to North-facing views (Figure 2). Chips are down-sampled from $900 \times 900$ pixel high-resolution images. In addition to the RGB images shown, the dataset comprises a high-resolution pan-chromatic (grayscale) band, a high-resolution near-infrared band, and a lower-resolution 8-band multispectral image for each geographic location/view combination. The dataset is available at [https://spacenet.ai](https://spacenet.ai) under a CC-BY SA 4.0 License.

To address the limitations detailed above, we introduce the SpaceNet Multi-View Overhead Imagery (MVOI) dataset, which includes 62,000 overhead images collected over Atlanta, Georgia USA and the surrounding areas. The dataset comprises 27 distinct looks, including both North- and South-facing views, taken during a single pass of a Maxar WorldView-2 satellite. The looks range from almost directly overhead ($7.8^\circ$ off-nadir) to up to $54^\circ$ off-nadir, with the same $665 \text{ km}^2$ geographic area covered by each. Alongside the imagery we open sourced an attendant 126,747 building footprints created by expert labelers. To our knowledge, this is the first multi-viewpoint dataset for overhead imagery with dense object annotations. The dataset covers heterogeneous geographies, including highly treed rural areas, suburbs, industrial areas, and high-density urban environments, resulting in heterogeneous building size, density, context and appearance (Figure 1). At the same time, the dataset abstracts away many other time-
sensitive variables (e.g. seasonality), enabling careful assessment of the impact of look angle on model training and inference. The training imagery and labels and public test images are available at https://spacenet.ai under the CC-BY SA 4.0 International License.

Though an ideal overhead imagery dataset would cover all the variables present in overhead imagery, i.e. look angle, seasonality, geography, weather condition, sensor, and light conditions, creating such a dataset is impossible with existing imagery. To our knowledge, the 27 unique looks in SpaceNet MVOI represent one of only two such imagery collections available in the commercial realm, even behind imagery acquisition company paywalls. We thus chose to focus SpaceNet MVOI on providing a diverse set of views with varying look angle and direction, a variable that is not represented in any existing overhead imagery dataset. SpaceNet MVOI could potentially be combined with existing datasets to train models which generalize across more variables.

We benchmark state-of-the art models on three tasks:

1. Building segmentation and detection.
2. Generalization of segmentation and object detection models to previously unseen angles.
3. Consequences of changes in resolution for segmentation and object detection models.

Our benchmarking reveals that state-of-the-art detectors are challenged by SpaceNet MVOI, particularly in views left out during model training. Segmentation and object detection models struggled to account for displacement of building footprints, occlusion, shadows, and distortion in highly off-nadir looks (Figure 3). The challenge of addressing footprint displacement is of particular interest, as it requires models not only to learn visual features, but to adjust footprint localization dependent upon the view context. Addressing these challenges is relevant to a number of applications outside of overhead imagery analysis, e.g. autonomous vehicle vision.

To assess model generalization to new looks we developed a generalization metric $G$, which reports the relative performance of models when they are applied to previously unseen looks. While specialized models designed for overhead imagery outperform general baseline models in building footprint detection, we found that models developed for natural image computer vision tasks have better $G$ scores on views absent during training. These observations highlight the challenges associated with developing robust models for multi-view object detection and semantic segmentation tasks. We therefore expect that developments in computer vision models for multi-view analysis made using SpaceNet MVOI, as well as analysis using our metric $G$, will be broadly relevant for many computer vision tasks.

The dataset is available at www.spacenet.ai

2. Related Work

Object detection and segmentation is a well-studied problem for natural scene images, but those objects are generally much larger and suffer minimally from distortions exacerbated in overhead imagery. Natural scene research is driven by datasets such as MSCOCO [20] and PASCALVOC [13], but those datasets lack multiple views of each object. PASCAL3D [35], autonomous driving datasets such as KITTI [14], CityScapes [7], existing multi-view datasets [29, 30], and tracking datasets such as MOT2017 [24] or OBT [33] contains different views but are confined to a narrow range of angles, lack sufficient heterogeneity to test generalization between views, and are restricted to natural scene images. Multiple viewpoints are found in 3D model datasets [5, 23], but those are not photorealistic and lack the occlusion and visual distortion properties encountered with real imagery.

Previous datasets for overhead imagery focus on classification [6], bounding box object detection [64, 19, 25], instance-based segmentation [12], and object tracking [26] tasks. None of these datasets comprise multiple images of the same field of view from substantially different look angles, making it difficult to assess model robustness to new views. Within segmentation datasets, SpaceNet [12] represents the closest work, with dense building and road annotations created by the same methodology. We summarize the key characteristics of each dataset in Table 1. Our dataset matches or exceeds existing datasets in terms of imagery size and annotation density, but critically includes varying look direction and angle to better reflect the visual heterogeneity of real-world imagery.

The effect of different views on segmentation or object detection in natural scenes has not been thoroughly studied, as feature characteristics are relatively preserved even under rotation of the object in that context. Nonetheless, preliminary studies of classification model performance on video frames suggests that minimal pixel-level changes can impact performance [11]. By contrast, substantial occlusion and distortion occurs in off-nadir overhead imagery, complicating segmentation and placement of geospatially accurate object footprints, as shown in Figure 3A-B. Furthermore, due to the comparatively small size of target objects (e.g. buildings) in overhead imagery, changing view substantially alters their appearance (Figure 3C-D). We expect similar challenges to occur when detecting objects in natural scene images at a distance or in crowded views. Existing solutions to occlusion are often domain specific [37] or rely on attention mechanisms to identify common elements [40] or landmarks [38]. The heterogeneity in building appearance in overhead imagery, and the absence of landmark features to identify them, makes their detection an ideal research task for developing domain-agnostic models that are robust to occlusion.
### Table 1: Comparison with other computer vision and overhead imagery datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gigapixels</th>
<th># Images</th>
<th>Resolution (m)</th>
<th>Nadir Angles</th>
<th># Objects</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpaceNet [12, 8]</td>
<td>10.3</td>
<td>24586</td>
<td>0.31</td>
<td>On-Nadir</td>
<td>302701</td>
<td>Polygons</td>
</tr>
<tr>
<td>DOTA [34]</td>
<td>44.9</td>
<td>2806</td>
<td>Google Earth*</td>
<td>On-Nadir</td>
<td>188282</td>
<td>Oriented Bbox</td>
</tr>
<tr>
<td>3K Vehicle Detection [21]</td>
<td>N/A</td>
<td>20</td>
<td>0.20</td>
<td>Aerial</td>
<td>14235</td>
<td>Oriented Bbox</td>
</tr>
<tr>
<td>UCAS-AOD [41]</td>
<td>N/A</td>
<td>1510</td>
<td>Google Earth*</td>
<td>On-Nadir</td>
<td>3651</td>
<td>Oriented Bbox</td>
</tr>
<tr>
<td>NWPU VHR-10 [4]</td>
<td>N/A</td>
<td>800</td>
<td>Google Earth*</td>
<td>On-Nadir</td>
<td>3651</td>
<td>Bbox</td>
</tr>
<tr>
<td>MVS [2]</td>
<td>111</td>
<td>50</td>
<td>0.31-0.58</td>
<td>[5.3, 43.3]</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td>FMoW [6]</td>
<td>1,084.0</td>
<td>523846</td>
<td>0.31-1.60</td>
<td>[0.22, 57.5]</td>
<td>132716</td>
<td>Classification</td>
</tr>
<tr>
<td>xView [19]</td>
<td>56.0</td>
<td>1400</td>
<td>0.31</td>
<td>On-Nadir</td>
<td>1000000</td>
<td>Bbox</td>
</tr>
<tr>
<td>SpaceNet MVOI (Ours)</td>
<td>50.2</td>
<td>60000</td>
<td>0.46-1.67</td>
<td>[-32.5, +54.0]</td>
<td>126747</td>
<td>Polygons</td>
</tr>
<tr>
<td>PascalVOC [13]</td>
<td>-</td>
<td>21503</td>
<td>-</td>
<td>-</td>
<td>62199</td>
<td>Bbox</td>
</tr>
<tr>
<td>MSCOCO [20]</td>
<td>-</td>
<td>123287</td>
<td>-</td>
<td>-</td>
<td>886266</td>
<td>Bbox</td>
</tr>
<tr>
<td>ImageNet [9]</td>
<td>-</td>
<td>349319</td>
<td>-</td>
<td>-</td>
<td>478806</td>
<td>Bbox</td>
</tr>
</tbody>
</table>

*Google Earth imagery is a mosaic from a variety of aerial and satellite sources and ranges from 15 cm to 12 m resolution [15].

---

**3. Dataset Creation**

SpaceNet MVOI contains images of Atlanta, GA USA and surrounding geography collected by Maxar’s WorldView-2 Satellite on December 22, 2009 [22]. The satellite collected 27 distinct views of the same 665 km² ground area during a single pass over a 5 minute span. This produced 27 views with look angles (angular distances between the nadir point directly underneath the satellite and the center of the scene) from 7.8° to 54° off-nadir and with a target azimuth angle (compass direction of image acquisition) of 17° to 182.8° from true North (see Figure 2). See the Supplementary Material and Tables S1 and S2 for further details regarding the collections. The 27 views in a narrow temporal band provide a dense set of visually distinct perspectives of static objects (buildings, roads, trees, utilities, etc.) while limiting complicating factors common to remote sensing datasets such as changes in cloud cover, sun angle, or land-use change. The imaged area is geographically diverse, including urban areas, industrial zones, forested suburbs, and undeveloped areas (Figure 1).

**Challenges in off-nadir imagery**

![Figure 2: Collect views. Location of collection points during the WorldView-2 satellite pass over Atlanta, GA USA.](image)

![Figure 3: Challenges with off-nadir look angles. Though geospatially accurate building footprints (blue) perfectly match building roofs at nadir (A), this is not the case off-nadir (B), and many buildings are obscured by skyscrapers.](image)

### 3.1 Preprocessing

Multi-view satellite imagery datasets are distinct from related natural image datasets in several interesting ways. First, as look angle increases in satellite imagery, the native resolution of the image decreases because greater distortion...
Our dataset labels comprise a broad distribution of building sizes, as shown in Figure 4A. Compared to natural image datasets, our dataset more heavily emphasizes small objects, with the majority of objects less than 700 pixels in area, or \( \sim \) 25 pixels across. By contrast, objects in the PASCALVOC [13] or MSCOCO [20] datasets usually comprise 50-300 pixels along the major axis [24].

### Table 2: Benchmark model selections for dataset baselines.

<table>
<thead>
<tr>
<th>Task</th>
<th>Baseline models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic Segmentation</td>
<td>TernausNet [17], U-Net [27]</td>
</tr>
<tr>
<td>Instance Segmentation</td>
<td>Mask R-CNN [16]</td>
</tr>
</tbody>
</table>

An additional challenge presented by this dataset, consistent with many real-world computer vision tasks, is the heterogeneity in target object density (Figure 4B). Images contained between zero and 300 footprints, with substantial coverage throughout that range. This variability presents a challenge to object detection algorithms, which often require estimation of the number of features per image [16]. Segmentation and object detection of dense or variable density objects is challenging, making this an ideal dataset to test the limits of algorithms’ performance.

### 4. Building Detection Experiments

#### 4.1. Dataset preparation for analysis

We split the training and test sets 80/20 by randomly selecting geographic locations and including all views for that location in one split, ensuring that each type of geography was represented in both splits. We group each angle into one of three categories: Nadir (NADIR), \( \theta \leq 25^\circ \); Off-nadir (OFF), \( 25^\circ < \theta < 40^\circ \); and Very off-nadir (VOFF), \( \theta \geq 40^\circ \). In all experiments, we trained baselines using all viewing angles (ALL) or one of the three subsets. These trained models were then evaluated on the test set of each of the 27 viewing angles individually.

#### 4.2. Models

We measured several state of the art baselines for semantic or instance segmentation and object detection (Table 2). Where possible, we selected overhead imagery-specific models as well as models for natural scenes to compare their performance. Object detection baselines were trained using rectangular boundaries extracted from the building footprints. To fairly compare with semantic segmentation studies, the resulting bounding boxes were compared against the ground truth building polygons for scoring (see Metrics).

#### 4.3. Segmentation Loss

Due to the class imbalance of the training data – only 9.5% of the pixels in the training set correspond to buildings – segmentation models trained with binary cross-entropy (BCE) loss failed to identify building pixels, a problem observed previously for overhead imagery segmentation models [31]. For the semantic segmentation models, we there-

Figure 4: **Dataset statistics.** Distribution of (A) building footprint areas and (B) number of objects per \( 450m \times 450m \) geographic tile in the dataset.
In aggregate, we then report the\( P \)lated the precision false positives. Using these statistics and the number of un-detected ground truth polygons (False Negatives), we calculated precision \( P \) and recall \( R \) of the model predictions in aggregate. We then report the \( F_1 \) score as
\[
F_1 = \frac{2 \times P \times R}{P + R}
\]
\( F_1 \) score was calculated within each angle bin (NADIR, OFF, or VOFF) and then averaged for an aggregate score.

### 4.4. Metrics

We measured performance using the building IoU-\( F_1 \) score defined in Van Etten et al. \cite{12}. Briefly, building footprint polygons were extracted from segmentation masks (or taken directly from object detection bounding box outputs) and compared to ground truth polygons. Predictions were labeled True Positive if they had an IoU with a ground truth polygon above 0.5 and all other predictions were deemed False Positives. Using these statistics and the number of undetected ground truth polygons (False Negatives), we calculated the precision \( P \) and recall \( R \) of the model predictions in aggregate. We then report the \( F_1 \) score as
\[
F_1 = \frac{2 \times P \times R}{P + R}
\]
\( F_1 \) score was calculated within each angle bin (NADIR, OFF, or VOFF) and then averaged for an aggregate score.

### 4.5. Results

The state-of-the-art segmentation and object detection models we measured were challenged by this task. As shown in Table 3, TernausNet trained on all angles achieves \( F_1 = 0.62 \) on the nadir angles, which is on par with previous building segmentation results and competitions \cite{12}. However, performance drops significantly for off-nadir \( (F_1 = 0.43) \) and very off-nadir \( (F_1 = 0.22) \) images. Other models display a similar degradation in performance. Example results are shown in Figure 5.

### Table 3: Overall task difficulty

As a measure of overall task difficulty, the performance (\( F_1 \) score) is assessed for the baseline models trained on all angles, and tested on the three different viewing angle bins: nadir (NADIR), off-nadir (OFF), and very off-nadir (VOFF). Avg. is the linear mean of the three bins. Seg, segmentation; Det, object detection.

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>NADIR</th>
<th>OFF</th>
<th>VOFF</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seg</td>
<td>TernausNet</td>
<td>0.62</td>
<td>0.43</td>
<td>0.22</td>
<td>0.43</td>
</tr>
<tr>
<td>Seg</td>
<td>U-Net</td>
<td>0.39</td>
<td>0.27</td>
<td>0.08</td>
<td>0.24</td>
</tr>
<tr>
<td>Seg</td>
<td>Mask R-CNN</td>
<td>0.47</td>
<td>0.34</td>
<td>0.07</td>
<td>0.29</td>
</tr>
<tr>
<td>Det</td>
<td>Mask R-CNN</td>
<td>0.40</td>
<td>0.30</td>
<td>0.07</td>
<td>0.25</td>
</tr>
<tr>
<td>Det</td>
<td>YOLT</td>
<td>0.49</td>
<td>0.37</td>
<td>0.20</td>
<td>0.36</td>
</tr>
</tbody>
</table>

### Table 4: Summary of results.

<table>
<thead>
<tr>
<th>Test Angles</th>
<th>Original</th>
<th>Equalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>NADIR</td>
<td>0.62</td>
<td>0.59</td>
</tr>
<tr>
<td>OFF</td>
<td>0.43</td>
<td>0.41</td>
</tr>
<tr>
<td>VOFF</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>Summary</td>
<td>0.43</td>
<td>0.41</td>
</tr>
</tbody>
</table>
Figure 5: Sample imagery (left) with ground truth building footprints and Mask R-CNN bounding boxes (middle left), TernausNet segmentation masks (middle right), and YOLT bounding boxes (right). Ground truth masks (light blue) are shown under Mask R-CNN and TernausNet predictions (yellow). YOLT bounding boxes shown in blue. Sign of the look angle represents look direction (negative=south-facing, positive=north-facing). Predictions from models trained on on all angles (see Table 3).

Figure 6: Performance by look angle for various training subsets. TernausNet (left), Mask R-CNN (middle), and YOLT (right) models, trained on ALL, NADIR, OFF, or VOFF, were evaluated in the building detection task and $F_1$ scores are displayed for each evaluation look angle. Imagery acquired facing South is represented as a negative number, whereas looks facing North are represented by a positive angle value. Additionally, TernausNet models trained only on North-facing OFF imagery (positive OFF) and South-facing OFF imagery (negative OFF) were evaluated on each angle to explore the importance of look direction.

To measure performance on unseen angles, we introduce a generalization score $G$, which measures the performance of a model trained on $X$ and tested on $Y$, normalized by the performance of models trained with many views, we also explored how effectively models could identify building footprints on look angles absent during training. We found that the TernausNet model trained only on NADIR performed worse on evaluation images from OFF (0.32) than models trained directly on OFF (0.44), as shown in Table 5. Similar trends are observed for object detection (Figure 6).
We quantify the performance degradation model generalization to unseen views represents a significant challenge. We measure relative performance across viewing angles, normalized by the task difficulty of the test set. We computed the generalization score \( G \) for all our model/dataset combinations, as reported in Table 6. The Mask R-CNN model has worse overall performance, the model achieved a higher generalization score \( G = 0.78 \) compared to TernausNet \( G = 0.42 \) as its performance did not decline as rapidly when look angle increased. Overall however, generalization scores to unseen angles were low, highlighting the importance of future study in this challenging task.

4.6. Effects of geography

We broke down geographic tiles into Industrial, Sparse Residential, Dense Residential, and Urban bins, and examined how look angle influenced performance in each. We observed greater effects on residential areas than other types (Table S3). Testing models trained on MVOI with unseen imagery (due to simple geometry), which theoretically compiles building extraction for high off-nadir angles. However, by experimenting with imagery degraded to the same resolution, we show that resolution has an insignificant impact on performance (Table 4). Rather, variations in illumination and viewing angle are the dominant factors. This runs contrary to recent observations [28], which found that object detection models identify small cars and other vehicles better in super-resolved imagery.

The off-nadir imagery has a lower resolution than nadir imagery (due to simple geometry), which theoretically complicates building extraction for high off-nadir angles. However, by experimenting with imagery degraded to the same low 1.67m resolution, we show that resolution has an insignificant impact on performance (Figure 3). We speculate that distortions in objects, occlusion, and variable lighting in off-nadir imagery (Figure 3), as well as the small size of buildings in general (Figure 4), pose an unusual challenge for segmentation and object detection of overhead imagery.

The off-nadir imagery has a lower resolution than nadir imagery (due to simple geometry), which theoretically complicates building extraction for high off-nadir angles. However, by experimenting with imagery degraded to the same low 1.67m resolution, we show that resolution has an insignificant impact on performance (Table 4). Rather, variations in illumination and viewing angle are the dominant factors. This runs contrary to recent observations [28], which found that object detection models identify small cars and other vehicles better in super-resolved imagery.

The generalization score \( G \) is low for the highest-performing, overhead imagery-specific models in these tasks (Table 6), suggesting that these models may be overfitting to view-specific properties. This challenge is not specific to overhead imagery: for example, accounting for distortion of objects due to imagery perspective is an essential component of 3-dimensional scene modeling, or rotation prediction tasks [23]. Taken together, this dataset and the \( G \) metric provide an exciting opportunity for future research on algorithmic generalization to unseen views.

Our aim for future work is to expose problems of interest to the larger computer vision community with the help of overhead imagery datasets. While only one specific application, advances in enabling analysis of overhead imagery in the wild can concurrently solve broader tasks. For example, we had anecdotally observed that image translation and domain transfer models failed to convert off-nadir images to nadir images, potentially due to the spatial shifts in the image. Exploring these tasks as well as other novel research avenues will enable advancement of a variety of current computer vision challenges.
References


SpaceNet MVOI: a Multi-View Overhead Imagery Dataset
Supplementary Material

A. Dataset

A.1. Imagery details

The images from our dataset were obtained from DigitalGlobe, with 27 different viewing angles collected over the same geographical region of Atlanta, GA. Each viewing angle is characterized as both an off-nadir angle and a target azimuth. We binned each angle into one of three categories (Nadir, Off-Nadir, and Very Off-Nadir) based on the angle (see Table 8). Collects were also separated into South- or North-facing based on the target azimuth angle.

The imagery dataset comprises Panchromatic, Multi-Spectral, and Pan-Sharpened Red-Green-Blue-near IR (RGB-NIR) images. The ground resolution of image varied depending on the viewing angle and the type of image (Panchromatic, Multi-spectral, Pan-sharpened). See Table 7 for more details. All experiments in this study were performed using the Pan-Sharpened RGB-NIR image (with the NIR band removed, except for the U-Net model).

The imagery was uploaded into the spacenet-dataset AWS S3 bucket, which is publicly readable with no cost to download. Download instructions can be found at www.spacenet.ai/off-nadir-building-detection/.

A.2. Dataset breakdown

The imagery described above was split into three folds: 50% in a training set, 25% in a validation set, and 25% in a final test set. 900 × 900-pixel geographic tiles were randomly placed in one of the three categories, with all of the look angles for a given geography assigned to the same subset to avoid geographic leakage. The full training set and building footprint labels as well as the validation set imagery were open sourced, and the validation set labels and final test imagery and labels were withheld as scoring sets for public coding challenges.

B. Model Training

B.1. TernausNet

The TernausNet model was trained without pre-trained weights roughly as described previously [17], with modifications. Firstly, only the Pan-sharpened RGB channels were used for training, and were re-scaled to 8-bit. 90° rotations, X and Y flips, imagery zooming of up to 25%, and linear brightness adjustments of up to 50% were applied randomly to training images. After augmentations, a 512 × 512 crop was randomly selected from within each 900 × 900 training chip, with one crop used per chip per training epoch. Secondly, as described in the Models section of the main text, a combination loss function was used with a weight parameter α = 0.8. Secondly, a variant of Adam incorporating Nesterov momentum [?] with default parameters was used as the optimizer. The model was trained for 25-40 epochs, and learning rate was decreased 5-fold when validation loss failed to improve for 5 epochs. Model training was halted when validation loss failed to improve for 10 epochs.

B.2. U-Net

The original U-Net [27] architecture was trained for 30 epochs with Pan-Sharpened RGB+NIR 16-bit imagery, on a binary segmentation mask with a combination loss as described in the main text with α = 0.5. Dropout and batch normalization were used at each layer, with dropout with p = 0.33. The same augmentation pipeline was used as with TernausNet. An Adam Optimizer [?] was used with learning rate of 0.0001 was used for training.

B.3. YOLT

The You Only Look Twice (YOLT) model was trained as described previously [11]. Bounding box training targets were generated by converting polygon building footprints into the minimal un-oriented bounding box that enclosed each polygon.

B.4. Mask R-CNN

The Mask R-CNN model with the ResNet50-C4 backbone was trained as described previously [16] using the same augmentation pipeline as TernausNet. Bounding boxes were created as described above for YOLT.

C. Geography-specific performance

C.1. Distinct geographies within SpaceNet MVOI

We asked how well the TernausNet model trained on SpaceNet MVOI performed both within and outside of the dataset. First, we broke down the test dataset into

<table>
<thead>
<tr>
<th>Image</th>
<th>Resolution at 7.8°</th>
<th>Resolution at 54°</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panchromatic</td>
<td>0.46m/px</td>
<td>1.67m/px</td>
</tr>
<tr>
<td>Multi-spectral</td>
<td>1.8m/px</td>
<td>7.0m/px</td>
</tr>
<tr>
<td>Pan-sharpened</td>
<td>0.46m/px</td>
<td>1.67m/px</td>
</tr>
</tbody>
</table>

Table 7: Resolution across different image types for two nadir angles.
<table>
<thead>
<tr>
<th>Catalog ID</th>
<th>Pan-sharpened Resolution</th>
<th>Look Angle</th>
<th>Target Azimuth Angle</th>
<th>Angle Bin</th>
<th>Look Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1030010003D22F00</td>
<td>0.48</td>
<td>7.8</td>
<td>118.4</td>
<td>Nadir</td>
<td>South</td>
</tr>
<tr>
<td>10300100023BC100</td>
<td>0.49</td>
<td>8.3</td>
<td>78.4</td>
<td>Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003993E00</td>
<td>0.49</td>
<td>10.5</td>
<td>148.6</td>
<td>Nadir</td>
<td>South</td>
</tr>
<tr>
<td>1030010003CAF100</td>
<td>0.48</td>
<td>10.6</td>
<td>57.6</td>
<td>Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010002B7D800</td>
<td>0.49</td>
<td>13.9</td>
<td>162</td>
<td>Nadir</td>
<td>South</td>
</tr>
<tr>
<td>10300100039AB000</td>
<td>0.49</td>
<td>14.8</td>
<td>35.1</td>
<td>Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010002649200</td>
<td>0.52</td>
<td>16.9</td>
<td>168.7</td>
<td>Nadir</td>
<td>South</td>
</tr>
<tr>
<td>1030010003993E00</td>
<td>0.52</td>
<td>19.3</td>
<td>30.7</td>
<td>Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003C92000</td>
<td>0.57</td>
<td>25.4</td>
<td>178.4</td>
<td>Nadir</td>
<td>South</td>
</tr>
<tr>
<td>1030010003472200</td>
<td>0.58</td>
<td>27.4</td>
<td>27.7</td>
<td>Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003315300</td>
<td>0.61</td>
<td>29.1</td>
<td>181</td>
<td>Off-Nadir</td>
<td>South</td>
</tr>
<tr>
<td>10300100036D5200</td>
<td>0.62</td>
<td>31</td>
<td>25.5</td>
<td>Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>103001000392F600</td>
<td>0.65</td>
<td>32.5</td>
<td>182.8</td>
<td>Off-Nadir</td>
<td>South</td>
</tr>
<tr>
<td>1030010003697400</td>
<td>0.68</td>
<td>34</td>
<td>23.8</td>
<td>Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003895500</td>
<td>0.74</td>
<td>37</td>
<td>22.6</td>
<td>Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003832800</td>
<td>0.8</td>
<td>39.6</td>
<td>21.5</td>
<td>Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>10300100035D1B00</td>
<td>0.87</td>
<td>42</td>
<td>20.7</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>10300100033CD700</td>
<td>0.95</td>
<td>44.2</td>
<td>20</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003713C00</td>
<td>1.03</td>
<td>46.1</td>
<td>19.5</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>10300100033C5200</td>
<td>1.13</td>
<td>47.8</td>
<td>19</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003492700</td>
<td>1.23</td>
<td>49.3</td>
<td>18.5</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>10300100039E6200</td>
<td>1.36</td>
<td>50.9</td>
<td>18</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003BDDC00</td>
<td>1.48</td>
<td>52.2</td>
<td>17.7</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003193D00</td>
<td>1.63</td>
<td>53.4</td>
<td>17.4</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
<tr>
<td>1030010003CD4300</td>
<td>1.67</td>
<td>54</td>
<td>17.4</td>
<td>Very Off-Nadir</td>
<td>North</td>
</tr>
</tbody>
</table>

Table 8: DigitalGlobe Catalog IDs and the resolution of each image based upon off-nadir angle and target azimuth angle.

the four bins represented in main text Figure 1: Industrial, Sparse Residential, Dense Residential, and Urban, and scored models within those bins (Table 9). We observed slightly worse performance in Industrials areas than elsewhere at nadir, but markedly stronger drops in performance in residential areas as look angle increased.

### C.2. Generalization to unseen geographies

We also explored how models trained on SpaceNet MVOI performed on building footprint extraction from imagery from other geographies, in this case, the Las Vegas imagery from SpaceNet [12]. After normalizing the Las Vegas (LV) imagery for consistent pixel intensities and channel order with SpaceNet MVOI, we predicted building footprints in LV imagery and scored prediction quality as described in Metrics. We also re-trained TernausNet on the LV imagery and examined building footprint extraction quality on the SpaceNet MVOI test set. Strikingly, neither model was able to identify building footprints in the unseen geographies, highlighting that adding novel looks angles does not necessarily enable generalization to new geographic areas.

<table>
<thead>
<tr>
<th>Type</th>
<th>NADIR</th>
<th>OFF - NADIR</th>
<th>VOFF - NADIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industrial</td>
<td>0.51</td>
<td>−0.13</td>
<td>−0.28</td>
</tr>
<tr>
<td>Sparse Res</td>
<td>0.57</td>
<td>−0.19</td>
<td>−0.37</td>
</tr>
<tr>
<td>Dense Res</td>
<td>0.66</td>
<td>−0.21</td>
<td>−0.41</td>
</tr>
<tr>
<td>Urban</td>
<td>0.64</td>
<td>−0.13</td>
<td>−0.30</td>
</tr>
</tbody>
</table>

Table 9: $F_1$ score for the model trained on all angles and evaluated evaluated on the nadir bins (NADIR), then the relative decrease in $F_1$ for the off-nadir and very off-nadir bins.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Training Set</th>
<th>MVOI ALL</th>
<th>SN LV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.68</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.62</td>
<td></td>
</tr>
</tbody>
</table>

Table 10: Cross-dataset $F_1$. Models trained on MVOI or SpaceNet Las Vegas [12] were inferenced on held out imagery from one of those two geographies, and building footprint quality was assessed as described in Metrics.
References


