SpaceNet MVOI: a Multi-View Overhead Imagery Dataset

Nicholas Weir[1], David Lindenbaum[2], Alexei Bastidas[3], Adam Van Etten[1], Sean McPherson[3], Jacob Shermeyer[1], Varun Kumar[3], and Hanlin Tang[3]

[1]In-Q-Tel CosmiQ Works, [nweir, avanetten, jshermeyer]@iqt.org
[2]Accenture Federal Services, david.lindenbaum@accenturefederal.com
[3]Intel AI Lab, [alexei.a.bastidas, sean.mcpherson, varun.v.kumar, hanlin.tang]@intel.com

Abstract

Detection and segmentation of objects in overheard 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* km^2 *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-

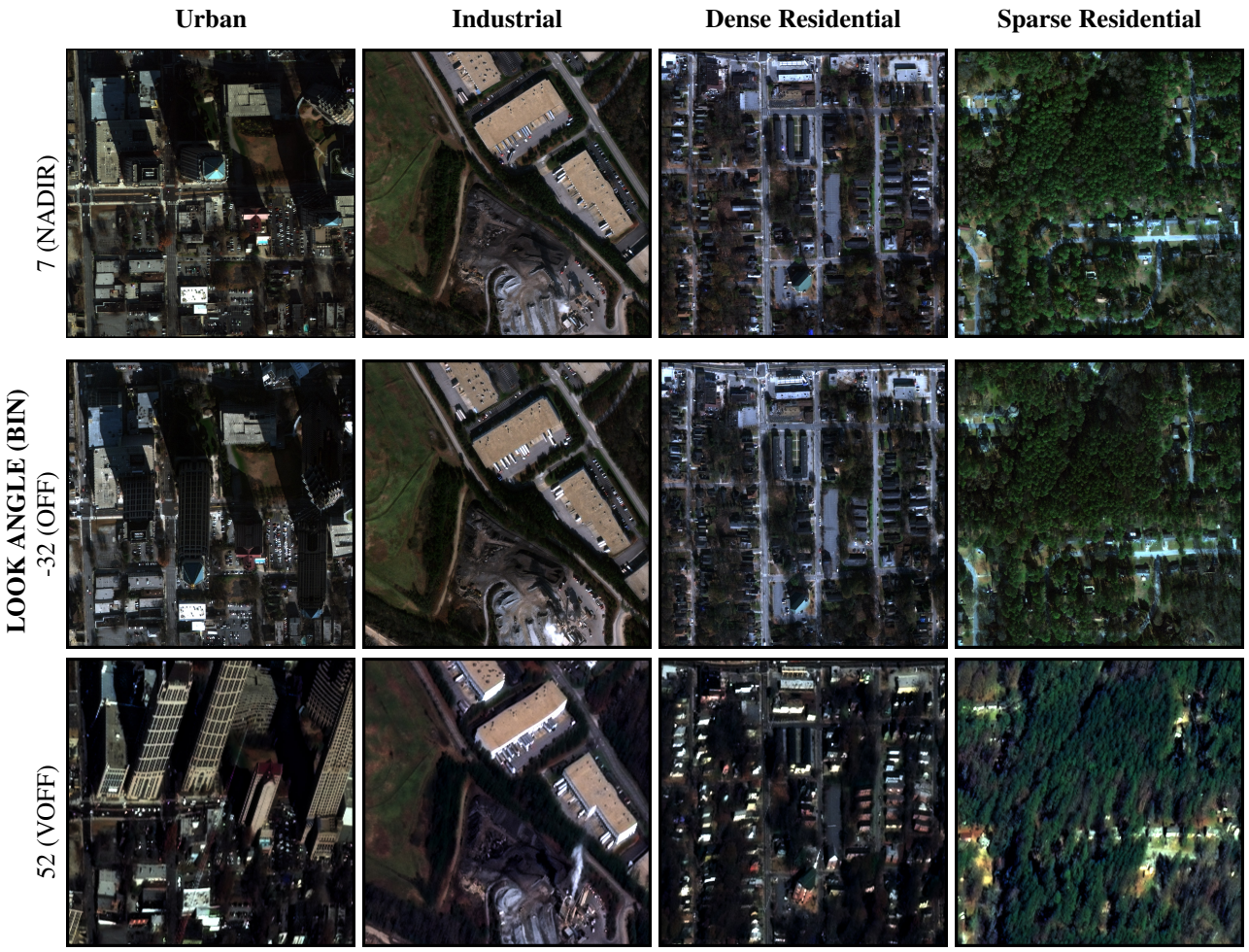

Figure 1: **Sample imagery from SpaceNet MVOI.** Four of the 2222 geographically unique image chips in the dataset are shown (columns), with three of the 27 views of that chip (rows), one from each angle bin. Negative look angle corresponds to South-facing views, whereas positive look angles correspond to North-facing views (Figure 2). Chips are down-sampled from 900×900 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` under a CC-BY SA 4.0 License.

ing evaluation of performance. These limitations extend to tasks outside of the geospatial domain: for example, convolutional neural nets perform inconsistently in many natural scene video frame classification tasks despite minimal pixel-level variation [1], and Xiao et al. showed that spatial transformation of images, effectively altering view, represents an effective adversarial attack against computer vision models [36]. Addressing generalization across views both within and outside of the geospatial domain requires two advancements: 1. A large multi-view dataset with diversity in land usage, population density, and views, and 2. A metric to assess model generalization.

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° off-nadir) to up to 54° off-nadir, with the same $665\ km^2$ 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 out-perform 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 photo-realistic 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 [34, 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 [1]. 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.

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

Table 1: **Comparison with other computer vision and overhead imagery datasets.** Our dataset has a similar scale as modern computer vision datasets, but to our knowledge is the first multi-view overhead imagery dataset designed for segmentation and object detection tasks. *Google Earth imagery is a mosaic from a variety of aerial and satellite sources and ranges from 15 cm to 12 m resolution [15].

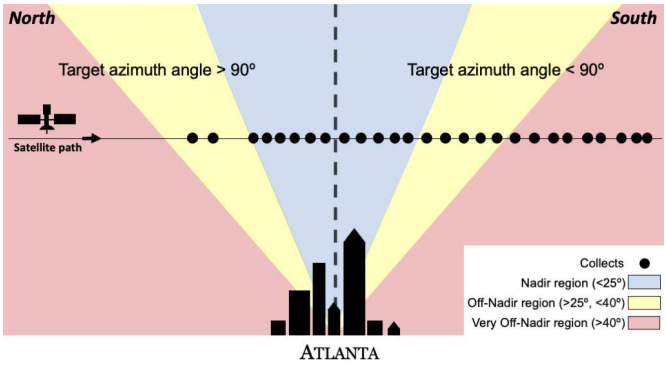

Figure 2: **Collect views.** Location of collection points during the WorldView-2 satellite pass over Atlanta, GA USA.

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 km^2 ground area during a single pass over a 5 minute span. This produced 27 views with look angles (angular distance 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).

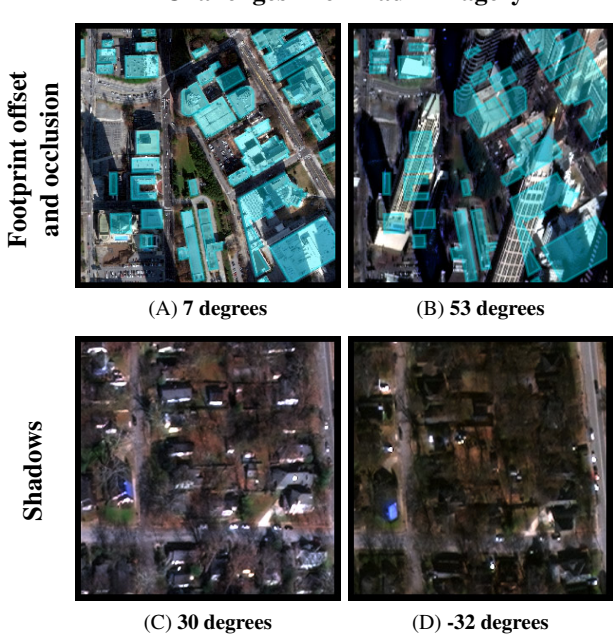

Figure 3: **Challenges with off-nadir look angles.** Though geospatially accurate building footprints (blue) perfectly match building roofs at nadir ***(A)***, this is not the case off-nadir ***(B)***, and many buildings are obscured by skyscrapers. ***(C-D):*** Visibility of some buildings changes at different look angles due to variation in reflected sunlight.

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

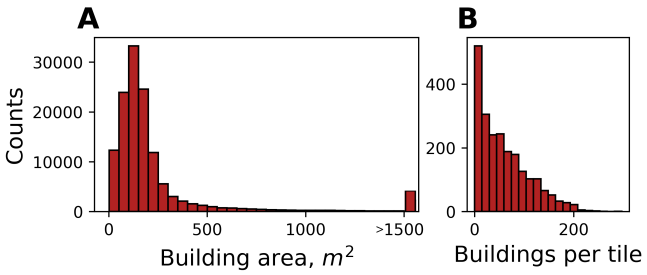

Figure 4: **Dataset statistics.** Distribution of **(A)** building footprint areas and **(B)** number of objects per $450m \times 450m$ geographic tile in the dataset.

is required to project the image onto a flat grid (Figure 1). Second, each view contains images with multiple spectral bands. For the purposes our baselines, we used 3-channel images (RGB: red, green, blue), but also examined the contributions of the near-infrared (NIR) channel (see Supplementary Material). These images were enhanced with a separate, higher resolution panchromatic (grayscale) channel to double the original resolution of the multispectral imagery (*i.e.*, "pan-sharpened"). The entire dataset was tiled into $900px \times 900px$ tiles and resampled to simulate a consistent resolution across all viewing angles of $0.5m \times 0.5m$ ground sample distance. The dataset also includes lower-resolution 8-band multispectral imagery with additional color channels, as well as panchromatic images, both of which are common overhead imagery data types.

The 16-bit pan-sharpened RGB-NIR pixel intensities were truncated at 3000 and then rescaled to an 8-bit range before normalizing to $[0, 1]$. We also trained models directly using Z-score normalized 16 bit images with no appreciable difference in the results.

3.2. Annotations

We undertook professional labeling to produce high-quality annotations. An expert geospatial team exhaustively labeled building footprints across the imaged area using the most on-nadir image (7.8° off-nadir). Importantly, the building footprint polygons represent geospatially accurate ground truth, and therefore are shared across all views. For structures occluded by trees, only the visible portion was labeled. Finally, one independent validator and one remote sensing expert evaluated the quality of each label.

3.3. Dataset statistics

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 ~ 25 pixels across. By contrast, objects in the PASCALVOC [13] or MSCOCO [20] datasets usually comprise 50-300 pixels along the major axis [34].

Task	Baseline models
Semantic Segmentation	TernausNet [17] , U-NET [27]
Instance Segmentation	Mask R-CNN [16]
Object Detection	Mask R-CNN [16], YOLT [11]

Table 2: Benchmark model selections for dataset baselines. TernausNet and YOLT are overhead imagery-specific models, whereas Mask R-CNN and U-Net are popular natural scene analysis models.

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), $\theta \leq 25°$; Off-nadir (OFF), $25° < \theta < 40°$; and Very off-nadir (VOFF), $\theta \geq 40°$. 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-

Task	Model	F_1 NADIR	OFF	VOFF	Avg.
Seg	TernausNet	**0.62**	**0.43**	**0.22**	**0.43**
Seg	U-Net	0.39	0.27	0.08	0.24
Seg	Mask R-CNN	0.47	0.34	0.07	0.29
Det	Mask R-CNN	0.40	0.30	0.07	0.25
Det	YOLT	0.49	0.37	0.20	0.36

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.

fore utilized a hybrid loss function that combines the binary cross entropy loss and intersection over union (IoU) loss with a weight factor α [31]:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{BCE}} + (1 - \alpha)\mathcal{L}_{\text{IoU}} \quad (1)$$

The details of model training and evaluation, including augmentation, optimizers, and evaluation schemes can be found in the Supplementary Material.

4.4. Metrics

We measured performance using the building IoU-F_1 score defined in Van Etten et al. [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

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (2)$$

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 [12, 8]. 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.

Test Angles	Training Resolution: Original (0.46-1.67 m)	Equalized 1.67 m
NADIR	0.62	0.59
OFF	0.43	0.41
VOFF	0.22	0.22
Summary	0.43	0.41

Table 4: **TernausNet model trained on different resolution imagery.** Building footprint extraction performance for a TernausNet model trained on ALL original-resolution imagery (0.46 m ground sample distance (GSD) for 7.8° to 1.67 m GSD at 54°), left, compared to the same model trained and tested on ALL imagery where every view is down-sampled to 1.67 m GSD (right). Rows display performance (F_1 score) on different angle bins. The original resolution imagery represents the same data as in Table 3. Training set imagery resolution had only negligible impact on model performance.

Directional asymmetry. Figure 6 illustrates performance per angle for both segmentation and object detection models. Note that models trained on positive (north-facing) angles, such as Positive OFF (Red), fair particularly poorly when tested on negative (south-facing) angles. This may be due to the smaller dataset size, but we hypothesize that the very different lighting conditions and shadows make some directions intrinsically more difficult (Figure 3C-D). This observation reinforces that developing models and datasets that can handle the diversity of conditions seen in overhead imagery in the wild remains an important challenge.

Model architectures. Interestingly, segmentation models designed specifically for overhead imagery (TernausNet and YOLT) significantly outperform general-purpose segmentation models for computer vision (U-Net, Mask R-CNN). These experiments demonstrate the value of specializing computer vision models to the target domain of overhead imagery, which has different visual, object density, size, and orientation characteristics.

Effects of resolution. OFF and VOFF images have lower base resolutions, potentially confounding analyses of effects due exclusively to look angle. To test whether resolution might explain the observed performance drop, we ran a control study with normalized resolution. We trained TernausNet on images from all look angles artificially reduced to the same resolution of $1.67m$, the lowest base resolution from the dataset. This model showed negligible change in performance versus the model trained on original resolution data (original resolution: $F_1 = 0.43$, resolution equalized: $F_1 = 0.41$) (Table 4). This experiment indicates that viewing angle-specific effects, not resolution, drive the decline in segmentation performance as viewing angle changes.

Generalization to unseen angles. Beyond exploring

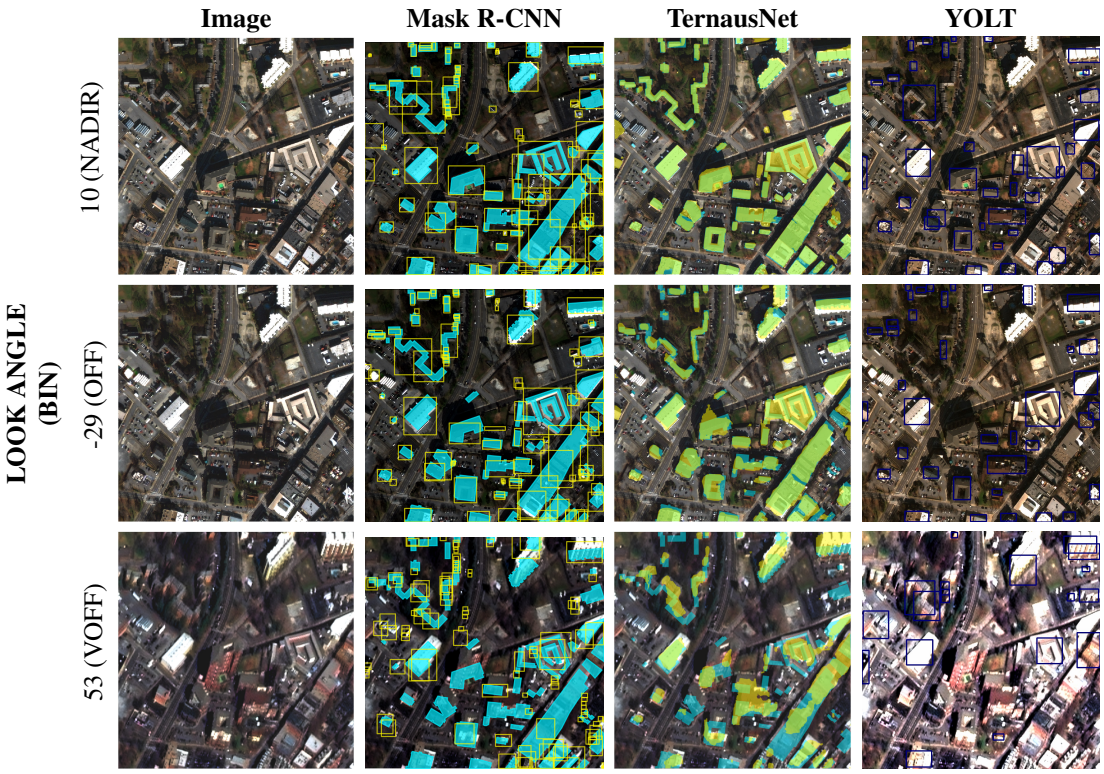

Figure 5: **Sample imagery (left) with ground truth building footprints and Mask R-CNN bounding boxes (middle left), TernausNet segmentation masks (middle right), and YOLT bounding boxes (right).** Ground truth masks (light blue) are shown under Mask R-CNN and TernausNet predictions (yellow). YOLT bounding 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).

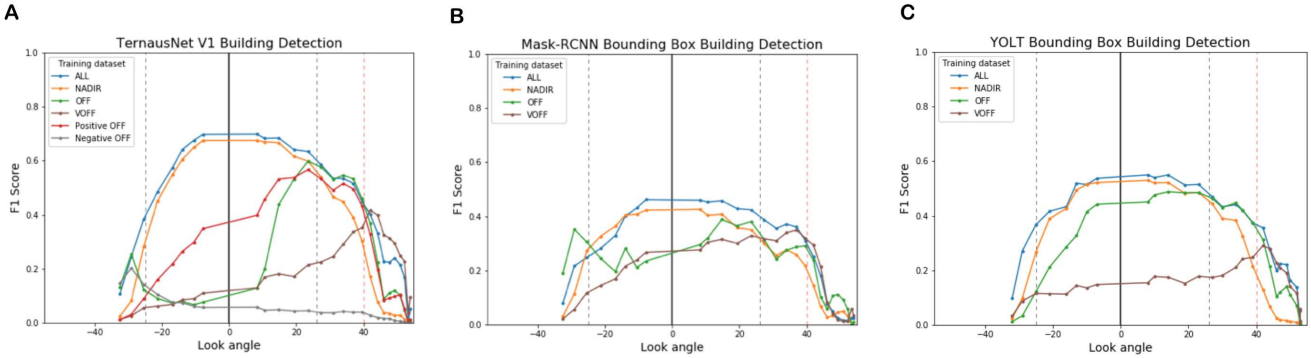

Figure 6: **Performance by look angle for various training subsets.** TernausNet (left), Mask R-CNN (middle), and YOLT (right) models, trained on ALL, NADIR, OFF, or VOFF, were evaluated in the building detection task and F_1 scores are displayed for each evaluation look angle. Imagery acquired facing 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.

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). 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

	Training Angles			
Test Angles	All	NADIR	OFF	VOFF
NADIR	0.62	0.59	0.23	0.13
OFF	0.43	0.32	0.44	0.23
VOFF	0.22	0.04	0.13	0.27
Summary	0.43	0.32	0.26	0.21

Table 5: **TernausNet model tested on unseen angles.** Performance (F_1 score) of the TernausNet model when trained on one angle bin (columns), and then tested on each of the three bins (rows). The model trained on NADIR performs worse on unseen OFF and VOFF views compared to models trained directly on imagery from those views.

performance of a model trained on Y and tested on Y:

$$G_Y = \frac{1}{N} \sum_X \frac{F_1(\text{train} = X, \text{test} = Y)}{F_1(\text{train} = Y, \text{test} = Y)} \quad (3)$$

This metric measures relative performance across viewing angles, normalized by the task difficulty of the test set. We measured G for all our model/dataset combinations, as reported in Table 6. Even though 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 cities[12] showed almost no generalization (Table S4). Additional datasets with more diverse geographies are needed.

5. Conclusion

We present a new dataset that is critical for extending object detection to real-world applications, but also presents challenges to existing computer vision algorithms. Our benchmark found that segmenting building footprints from very off-nadir views was exceedingly difficult, even for state-of-the-art segmentation and object detection models tuned specifically for overhead imagery (Table 3). The relatively low F_1 scores for these tasks (maximum VOFF F_1 score of 0.22) emphasize the amount of improvement that further research could enable in this realm.

Furthermore, on all benchmark tasks we concluded that model generalization to unseen views represents a significant challenge. We quantify the performance degradation from nadir (F_1 = 0.62) to very off-nadir (F_1 = 0.22), and

		Generalization Score G		
Task	**Model**	NADIR	OFF	VOFF
Segmentation	TernausNet	0.45	0.43	0.37
Segmentation	U-Net	0.64	0.40	0.37
Segmentation	Mask R-CNN	0.60	0.90	0.84
Detection	Mask R-CNN	0.64	0.92	0.76
Detection	YOLT	0.57	0.68	0.44

Table 6: **Generalization scores.** To measure segmentation model performance on unseen views, we compute a generalization score G (Equation 3), which quantifies performance on unseen views normalized by task difficulty. Each column corresponds to a model trained on one angle bin.

note an asymmetry between performance on well-lit north-facing imagery and south-facing imagery cloaked in shadows (Figure 3C-D and Figure 6). 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

[1] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *CoRR*, abs/1805.12177, 2018.

[2] Marc Bosch, Zachary Kurtz, Shea Hagstrom, and Myron Brown. A multiple view stereo benchmark for satellite imagery. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9, Oct 2016.

[3] Yushi Chen, Xing Zhao, and Xiuping Jia. Spectral–Spatial Classification of Hyperspectral Data Based on Deep Belief Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2381–2392, July 2015.

[4] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54:7405–7415, 2016.

[5] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A large dataset of object scans. *CoRR*, abs/1602.02481, 2016.

[6] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2018.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *The 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A Challenge to Parse the Earth Through Satellite Images. In *The 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, , Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *The 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] DigitalGlobe. Digitalglobe search and discovery. "`https://discover.digitalglobe.com`". Accessed: 2019-03-19.

[11] Adam Van Etten. You only look twice: Rapid multi-scale object detection in satellite imagery. *CoRR*, abs/1805.09512, 2018.

[12] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: A Remote Sensing Dataset and Challenge Series. *CoRR*, abs/1807.01232, 2018.

[13] Marc Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[15] Google. Google maps data help. `https://support.google.com/mapsdata`. Accessed: 2019-3-19.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *The 2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[17] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation. *CoRR*, abs/1801.05746, 2018.

[18] F.M. Lacar, Megan Lewis, and Iain Grierson. Use of hyperspectral imagery for mapping grape varieties in the Barossa Valley, South Australia. In *IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217)*, pages 2875–2877 vol.6, 2001.

[19] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xView: Objects in context in overhead imagery. *CoRR*, abs/1802.07856, 2018.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *2014 European Conference on Computer Vision (ECCV)*, Zurich, 2014. Oral.

[21] Kang Liu and Gellért Máttyus. Fast multiclass vehicle detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, 12:1938–1942, 2015.

[22] Nathan Longbotham, Chuck Chaapel, Laurence Bleiler, Chris Padwick, William J. Emery, and Fabio Pacifici. Very High Resolution Multiangle Urban Classification Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 50(4):1155–1170, April 2012.

[23] William Lotter, Gabriel Kreiman, and David D. Cox. Unsupervised learning of visual structure using predictive generative networks. *CoRR*, abs/1511.06380, 2015.

[24] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016.

[25] T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. *ECCV*, abs/1609.04453, 2016.

[26] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *The 2016 European Conference on Computer Vision (ECCV)*, 2016.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net - Convolutional Networks for Biomedical Image Segmentation. *MICCAI*, 9351(Chapter 28):234–241, 2015.

[28] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. *CoRR*, abs/1812.04098, 2018.

[29] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[30] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *2013 IEEE International Conference on Computer Vision (ICCV)*, pages 2456–2463, 2013.

[31] Tao Sun, Zehui Chen, Wenxiang Yang, and Yin Wang. Stacked u-nets with multi-output for road extraction. In *The 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[32] Burak Uzkent, Aneesh Rangnekar, and M.J. Hoffman. Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 233–242, July 2017.

[33] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2411–2418, 06 2013.

[34] Gui-Song Xia, Xiang Bai, Zhen Zhu Jian Ding, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Nov. 2017.

[35] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[36] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

[37] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015.

[38] Kevan Yuen and Mohan Manubhai Trivedi. An occluded stacked hourglass approach to facial landmark localization and occlusion estimation. *IEEE Transactions on Intelligent Vehicles*, 2:321–331, 2017.

[39] Peter W.T. Yuen and Mark A. Canton Richardson. An introduction to hyperspectral imaging and its application for security, surveillance and target acquisition. *The Imaging Science Journal*, 58(5):241–253, 2010.

[40] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *The 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[41] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3735–3739, 2015.