OBJECT GEOLOCATION USING MRF BASED MULTI-SENSOR FUSION

Vladimir A. Krylov and Rozenn Dahyot

ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

ABSTRACT

Abundant image and sensory data collected over the last decades represents an invaluable source of information for cataloging and monitoring of the environment. Fusion of heterogeneous data sources is a challenging but promising tool to efficiently leverage such information. In this work we propose a pipeline for automatic detection and geolocation of recurring stationary objects deployed on fusion scenario of street level imagery and LiDAR point cloud data. The objects are geolocated coherently using a fusion procedure formalized as a Markov random field problem. This allows us to efficiently combine information from object segmentation, triangulation, monocular depth estimation and position matching with LiDAR data. The proposed fusion approach produces object mappings robust to scenes reporting multiple object instances. We introduce a new challenging dataset of over 200 traffic lights in Dublin city centre and demonstrate high performance of the proposed methodology and its capacity to perform multi-sensor data fusion.

Index Terms— Object geolocation, street level imagery, LiDAR data, Markov random fields, traffic lights

1. INTRODUCTION

The last decade has witnessed unprecedented developments in computer vision largely due to the availability of immense image datasets accumulated by companies and individual users all around the world. Georeferenced imagery is a unique source of information for monitoring, cataloging and mapping tasks laying at the heart of various navigation, management and planning problems. Such imagery includes street level collections, like Google Street View (GSV) and Bing Streetside, as well as other sources of information such as satellite imagery, ground and airborne 3d point clouds. In this work we address geolocation of objects such as road-side furniture, facade elements, and street vegetation. Inventory and geolocation of such objects is a highly relevant task which OpenStreetMap and Mapillary address by encouraging their users to perform manually thus enriching their databases.

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No.713567 as well as by the ADAPT Centre for Digital Content Technology funded by the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106).

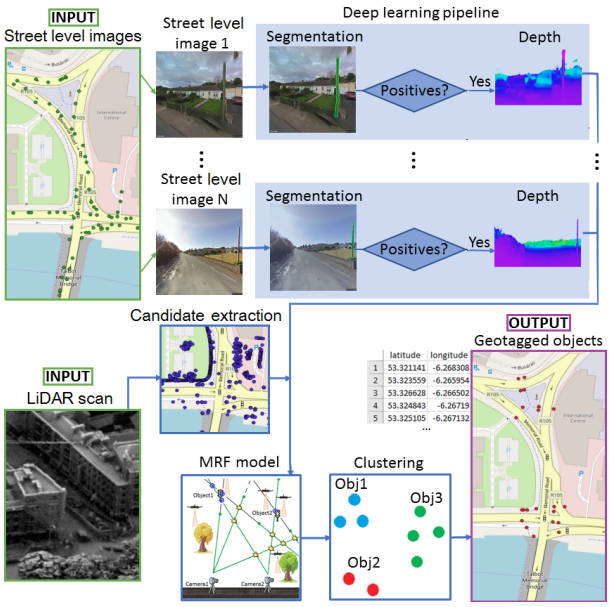

Fig. 1. Multi-sensor fusion pipeline: from street level images and LiDAR scan to object geolocation map.

A considerable effort has been dedicated to leveraging street level imagery for detection of certain types of road assets, like manholes [1], road signs [2], telecom assets [3], etc. The geolocation of traffic lights has been addressed in [4] by relying on fixed lens diameter and in [5, 6] by performing tracking or template matching in video sequences. These methods rely on specific geometric shapes of objects and perform visual matching of objects. Street level imagery has also been employed in combination with other data sources: remotely sensed optical imagery for road segmentation [7] and tree detection [8], or airborne LiDAR for land-use segmentation [9]. In all methods the objects are assumed to be identified in all the involved image modalities.

Mobile (ground) LiDAR data has been often used for road scene analysis [10, 11, 12]. Airborne LiDAR data has been employed for mapping of trees [13], trees and buildings [14] and cars [15]. Only the mobile scans have been previously employed to explore smaller road-side object that are on the edge of the geometric resolution available for airborne scans. To the best of our knowledge our's is the first work to exploit airborne LiDAR scans to such purpose.

We propose a novel model for multi-sensor fusion based

on Markov random fields (MRF) formulation. Our approach has the capacity to perform information fusion from multiple sources: multi-sensor imagery, heat maps of object density, multi-temporal imagery, etc. The proposed technique allows automatic processing of multi-object scenes and may be easily adjusted to detect custom objects thanks to its modular structure. In this study we explore a particular fusion scenario of street level imagery and 3d point cloud (LiDAR) data. Performance of the proposed method is validated on traffic lights (TL) detection. To this end we introduce a new Dublin TLs dataset. Our earlier work [16] covers a reduced version of the pipeline for mapping from street level imagery only which is extended here to multi-sensor fusion with airborne LiDAR.

Our fusion methodology is presented in Sec. 2 and validated experimentally in Sec. 3. Sec. 4 concludes this study.

2. FUSION PROCEDURE

Our fusion procedure receives as input the detection performed separately on the input data modalities. In this study we focus on a particular fusion scenario of street level imagery and LiDAR point clouds. We propose a complete fusion pipeline with the following components, see Fig. 1: street level imagery processing module, LiDAR candidate point extraction module and MRF-based information fusion module.

Assumptions To allow automatic processing of multi-object scenes we impose a mild assumption of object sparsity: instances should be at least 1m apart to be uniquely identified. This assumption may be critical only for objects that are likely to cluster such as traffic lights and poles. The street level imagery is considered the primary source of information and potential object locations are generated from this data source. This is due to the resolution limitations of the airborne 3d point cloud dataset: the geometric resolution may not allow one to identify smaller street furniture. For instance, traffic lights may not be reliably distinguished from utility poles or lampposts. Furthermore, road-side objects such as traffic lights and road signs are necessarily visible in street level imagery due to both higher geometric resolution and road regulations. Finally, whereas point cloud data is a more natural source of information for object localization, airborne scans typically suffer from blind spots in the dense urban environment: shadows from trees or buildings may result in invisible instances of street furniture.

Street level imagery processing Two state-of-the-art fully convolutional neural networks (FCNNs) for semantic segmentation [17] and monocular depth estimation [18] are used for processing the street level imagery. The object segmentation module has been prepared using datasets [19, 20] on TLs images whereas the depth estimation module is employed with no modifications, see [16] for more information.

LiDAR candidate point extraction The detection of potential object locations is addressed as a template matching problem. Depending on the type of objects various templates may be employed. Here we employ a template characterizing TLs and, more generally, pole-like objects. We assume that such objects are free-standing with height of $h \in [2, 6]$ meters above the ground level. Since the latter is not known a priori, we instead detect the height above the median elevation level of all points in 1m radius in (x, y)-plane. This may result in false positives corresponding to roof-top objects like antennas or chimneys. The thresholded points are clustered in $0.2m$ areas to produce a list of LiDAR candidate points for the locations of pole-like objects and TLs.

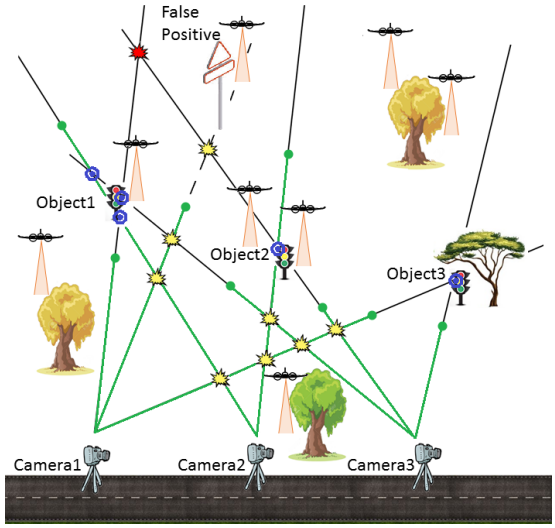

Fig. 2. Example object geolocation problem based on intersections of street level view-rays. Three objects are observed from three camera positions. Monocular depth estimates in green and LiDAR matches depicted by a drone icon.

MRF information fusion All objects are assumed to be located in a subset of intersections of view-rays cast in the direction of the traffic lights segmented in street level imagery. Formally, we explore the space $\mathcal{X}$ of all *pairwise* intersections from camera locations (see Fig. 2). Binary labels $z \in \{0, 1\}$ are associated to each node in $\mathcal{X}$: one indicates presence and zero absence of objects at the intersection. Space $\mathcal{Z}$ is considered a binary MRF [21]. Each site x_i is characterized by:

(i) Distances d_{i1} and d_{i2} from cameras obtained through triangulation of camera positions and rays. Distant intersection ($d_{i.} > 25$m) are discarded, see red intersection in Fig. 2).

(ii) Monocular depth estimates Δ_{i1} and Δ_{i2} of distances between camera positions and the detected object at x_i.

(iii) Distance L_i to the closest LiDAR candidate point.

The configuration of objects is found by relying on the distance information estimated from street level imagery and 3d point cloud data. The neighborhood of node x_i is defined as the set of all other locations x_k in $\mathcal{X}$ on rays r_1 and r_2 that generate it. Note that the number of neighbors (i.e. neighborhood size) for each node x_i in $\mathcal{X}$ is not constant and depends on the density of objects (rays) in the area.

We define coherency of configuration $\mathcal{Z}$ as follows: Any

ray may have at most one intersection with $z = 1$ with rays from any particular camera location, but several positive intersections with rays generated from different cameras are allowed, e.g. multiple intersections for Object1 in Fig. 2.

MRF energy MRF configuration is defined by $\{(x_i, z_i)\}$. For each site x_i with state z_i the MRF energy [21] is composed of the following terms:

• Unary term to enforce consistency with depth estimates:

$$u_D(z_i|\mathcal{X}, \mathcal{Z}) = z_i \sum_{j=1,2} \|\Delta_{ij} - d_{ij}\|^2 \quad (1)$$

• Unary term to penalize consistency with LiDAR matches:

$$u_L(z_i|\mathcal{X}, \mathcal{Z}) = z_i L_i^2 \quad (2)$$

• Pairwise term that enforces coherency of the configuration. Specifically, along each view ray it penalizes multiple objects of interest occluding each other, and excessive spread in case an object is characterized as several intersections. This term allows us to admit several positive intersections on the same ray only when they are in close proximity. This may occur in multi-view scenario due to segmentation inaccuracies and noise in camera geotag, see in Fig. 2 Object1 detected as a triangle of intersections with $z = 1$. The term is defined as penalty proportional to the distance to any other intersections x_k with $z_k = 1$ on rays r_1 and r_2:

$$u_C(R_i|\mathcal{X}, \mathcal{Z}) = \sum_{x_m, x_n \in R_i} z_m z_n \, \|x_m - x_n\|^2, \quad (3)$$

where R_i is a subset of $\mathcal{X}$ that belongs to the same ray.

• High-order term to penalize rays that have no intersections with $z = 1$. This corresponds to false positives or objects discovered from a single camera position (see Fig. 2):

$$u_0(R_i|\mathcal{X}, \mathcal{Z}) = \prod_{x_n \in R_i} (1 - z_n) \quad (4)$$

The full energy of configuration $\mathbf{z}$ in $\mathcal{Z}$ is defined as sum of energy contributions over all sites in $\mathcal{Z}$:

$$\begin{aligned} \mathcal{U}(\mathbf{z}) = &\sum_{\forall x_i \in \mathcal{X}} \Big[c_D \, u_D(z_i) + c_L \, u_L(z_i) \Big] + \\ &\sum_{\forall \text{ rays } R_j} \Big[c_C \, u_C(R_j) + c_0 \, u_0(R_j) \Big], \end{aligned} \quad (5)$$

with parameter vector $C = (c_D, c_L, c_C, c_0)$ with non-negative components subject to $c_D + c_L + c_C + c_0 = 1$. Distance-based contributions (1)-(3) in the energy are squared to non-linearly increase the penalty of position errors. Optimal configuration is reached at the global minimum of $\mathcal{U}(\mathbf{z})$.

MRF optimization Energy minimization is achieved with Iterative Conditional Modes [21] starting from an empty configuration: $z_i^0 = 0, \forall i$. The local optimization is driven by a random node-revisiting schedule until local minimum is reached. Use of more complex optimization method, e.g. graph-cuts, poses difficulties due to the irregular MRF grid.

Post-processing To obtain the final object configuration we perform clustering of MRF output in order to merge groups of object instances that describe the same physical object. This is relevant since we consider the space $\mathcal{X}$ of only pairwise intersections, whereas some objects are observed from three or more camera positions and result in multiple detected object instances, see Object1 in Fig. 2. We employ agglomerative hierarchical *clustering* with an intra-cluster distance of 1m which corresponds to our object sparsity assumption. Object coordinates in each cluster are averaged.

Snapping to the closest LiDAR candidate point may also be used which results in improved precision but lower recall.

3. EXPERIMENTAL VALIDATION

Dublin TL Dataset. To evaluate the performance of the proposed pipeline numerically we introduce a traffic light dataset in 0.75 km^2 area in central Dublin, Ireland, available at `github.com/vlkryl/streetview_objectmapping`, see Fig. 3. The dataset consists of GPS-coordinates of *all* 192 supported (pole-mounted) TLs in 2015 and 209 in 2017 in the specified area. Dataset contains various types of standard and multi-section TLs for pedestrians, cars and trams. Any TLs mounted on the *same* pole are considered as one object. Several suspended poles (above the road) present in the area are excluded from the set. TLs are clustered around 26 junctions of different complexity: from 2 to 16 TLs per junction.

Experimental setup. We employ GSV as source of street level imagery and high resolution airborne LiDAR scan [22] collected in March 2015. In the area covered by our dataset, the 3d point cloud contains approximatively 0.4 billion points. The analysis has revealed about 12300 locations that match the pole template and 668 locations are in 10m vicinity of the 192 TLs in our dataset. About 10% of TLs in the ground truth can not be seen in the LiDAR scan due to blind spots and object proximity. To achieve maximal consistency between data sources we use GSV imagery recorder in 2014-2015 harvested automatically through the Google API. This dataset includes 1307 panoramic images covering all roads in the area.

The following parameters are set by trial-and-error: depth weight $c_D = 0.05$, LiDAR weight $c_L = 0.1$, coherency weight $c_C = 0.1$, and non-paired object penalty $c_0 = 0.75$. ICM optimization was run for 25 iterations, final clustering performed with radius of 1m. 847 individual instances (single-views) of TLs were detected in GSV images, see examples in [16]. Fig. 3 presents detection results reported by the proposed fusion technique from depth+LiDAR and depth only ($c_L = 0$). As can be seen (zoom Fig. 3), the precision of LiDAR+depth estimates is higher than that of the depth-based estimates. Depth+LiDAR detection reported 206 objects: 79.5% (88.3%) recall and 81% (96.1%) precision in 3m (5m);

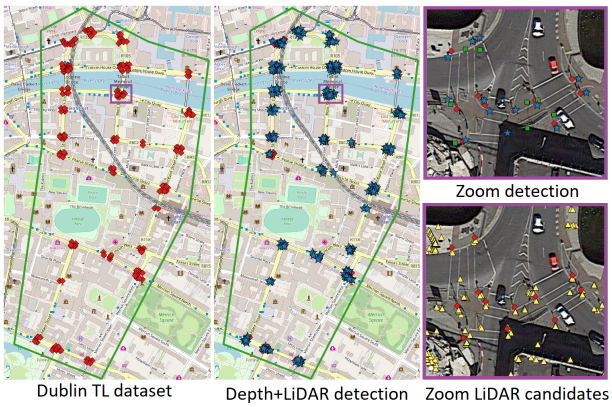

Fig. 3. Dublin TL dataset (♦) in 0.75 km^2 area inside green polygon, and depth+LiDAR detection results (★). Zooms include depth-based detection (■) and LiDAR candidates (▲).

95% empirical confidence interval is of 4.4m. The MRF module is computed in 6 seconds in a Python implementation on Ubuntu 16.04, i7-6700K CPU machine with 64 GB RAM.[1]

In Fig. 4 we analyze the average recall and precision reported by the proposed sensor fusion approach. These are reported as function of distance l to allow definition of true positives: An estimated location is considered true positive if it within l meters of a ground truth point. We plot averages from 100 reruns of the method to compensate for the stochastic impact of ICM. The top plot shows recall-precision curves for $l \in [2.5, 7.5]$ and $c_L \in [0, 0.2]$: for each colored curve c_L grows from bottom-right ($c_L = 0$ on dotted line) to top-left ($c_L = 0.2$), $c_L = 0.1$ on dashed line. The dashed line corresponds to the parametric setting of Fig. 3 LiDAR+depth experiment, and dotted line to that of the depth-based. For each l the highest recall corresponds to depth-based detection, i.e. $c_L = 0$ (dotted line). Detection precision increases dramatically with stronger LiDAR contribution, whereas recall somewhat drops. The latter is due to objects in LiDAR blind spots.

The bottom graph in Fig. 4 shows precision for different data fusion scenarios with pale colored areas showing the interval within one standard deviation of the mean (due to ICM). "LiDAR" and "depth" models are obtained by setting $c_D = 0$ and $c_L = 0$, respectively. Snapping in "depth+LiDAR" fusion scenario allows higher low-distance precision but with lower recall. The relative position of the curves clearly demonstrates the contribution of data sources: LiDAR alone gives higher precision than depths alone, and the fusion outperforms both single data-source scenarios. The "depth linear"-plot corresponds to the variant of our approach proposed in [16] where the distance-based energy terms u_D, u_L and u_C are linear w.r.t. distances. Quadratic weights improve the performance as clearly seen from the plots. Our method outperforms [3, 4] due to its capacity to geolocate multiple visually identical objects from the same scene.

[1]We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for image processing in this work.

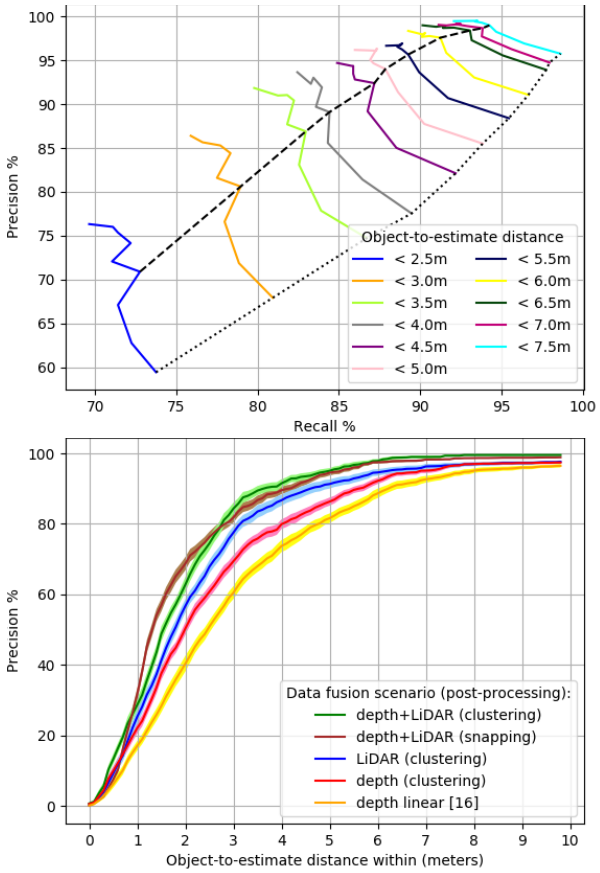

Fig. 4. Average recall-precision of TL detection with distance-based definition of positives: (top) for LiDAR weight $c_L \in [0, 0.2]$, and (bottom) as function of distance.

4. CONCLUSIONS

We have proposed an automatic object geolocation technique that is capable of fusing information from multi-sensor data. This is achieved by a novel MRF information fusion approach defined over irregular grid. This approach allows us to automatically handle complex multi-object scenes with sparse image input. Specifically, we have explored the fusion scenario of street level imagery and LiDAR data for geolocation of recurring stationary street-side objects. To evaluate the performance of the fusion methodology we introduce a challenging traffic light geolocation dataset of Dublin, in an area fully covered by publicly available GSV imagery and high resolution airborne LiDAR scan [22]. Our experiments demonstrate a clear gain in detection precision associated with the fusion with LiDAR data with the street level imagery.

As future work we will consider strategies to employ machine learning for extraction of LiDAR matches which will also allow one to address geolocation of more complex objects. Another interesting avenue of investigation is the fusion with oblique drone imagery, e.g. [22, 23], in order to reduce the dependency on street level imagery-based triangulation which is sensitive to camera-positioning noise.

5. REFERENCES

[1] R. Timofte and L. Van Gool, "Multi-view manhole detection, recognition, and 3d localisation," in *Proc IEEE ICCV Workshops*, 2011, pp. 188–195.

[2] B. Soheilian, N. Paparoditis, and B. Vallet, "Detection and 3D reconstruction of traffic signs from multiple view color images," *ISPRS J Photogram Rem Sens*, vol. 77, pp. 1–20, 2013.

[3] R. Hebbalaguppe, G. Garg, E. Hassan, H. Ghosh, and A. Verma, "Telecom inventory management via object recognition and localisation on Google street view images," in *Proc IEEE WACV*, 2017, pp. 725–733.

[4] N. Fairfield and C. Urmson, "Traffic light mapping and detection," in *Proc IEEE Int Conf Robotics Automation (ICRA)*, 2011, pp. 5421–5426.

[5] J. Levinson, J. Askeland, J. Dolson, and S. Thrun, "Traffic light mapping, localization, and state detection for autonomous vehicles," in *Proc IEEE Int Conf Robotics Automation (ICRA)*, 2011, pp. 5784–5791.

[6] G. Trehard, E. Pollard, B. Bradai, and F. Nashashibi, "Tracking both pose and status of a traffic light via an interacting multiple model filter," in *Int Conf Information Fusion (FUSION)*, 2014, pp. 1–7.

[7] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun, "HD maps: Fine-grained road segmentation by parsing ground and aerial images," in *Proc IEEE Conf CVPR*, 2016, pp. 3611–3619.

[8] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona, "Cataloging public objects using aerial and street-level images-urban trees," in *Proc IEEE Conf CVPR*, 2016, pp. 6014–6023.

[9] W. Zhang, W. Li, C. Zhang, D. M. Hanink, X. Li, and W. Wang, "Parcel-based urban land use classification in megacity using airborne lidar, high resolution orthoimagery, and google street view," *Comput Environ Urban Syst*, vol. 64, pp. 215–228, 2017.

[10] Y. Yu, J. Li, H. Guan, C. Wang, and J. Yu, "Semiautomated extraction of street light poles from mobile lidar point-clouds," *IEEE Trans Geosci Remote Sens*, vol. 53, no. 3, pp. 1374–1386, 2015.

[11] B. Yang, Z. Dong, G. Zhao, and W. Dai, "Hierarchical extraction of urban objects from mobile laser scanning data," *ISPRS J Photogram Rem Sens*, vol. 99, pp. 45–57, 2015.

[12] M. Lehtomäki, A. Jaakkola, J. Hyyppä, J. Lampinen, H. Kaartinen, A. Kukko, E. Puttonen, and H. Hyyppä, "Object classification and recognition from mobile laser scanning point clouds in a road environment," *IEEE Trans Geosci Remote Sens*, vol. 54, no. 2, pp. 1226–1239, 2016.

[13] M. Weinmann, M. Weinmann, C. Mallet, and M. Brédif, "A classification-segmentation framework for the detection of individual trees in dense MMS point cloud data acquired in urban areas," *Remote Sens*, vol. 9, no. 3, pp. 277–305, 2017.

[14] F. Rottensteiner, G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, "Results of the ISPRS benchmark on urban object detection and 3d building reconstruction," *ISPRS J Photogram Rem Sens*, vol. 93, pp. 256–271, 2014.

[15] J. Zhang, M. Duan, Q. Yan, and X. Lin, "Automatic vehicle extraction from airborne lidar data using an object-based point cloud analysis method," *Remote Sens*, vol. 6, no. 9, pp. 8405–8423, 2014.

[16] V. A. Krylov, E. Kenny, and R. Dahyot, "Automatic discovery and geotagging of objects from street view imagery," *Remote Sens.*, vol. 10, no. 5, 2018.

[17] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE T-PAMI*, vol. 39, no. 4, pp. 640–651, 2017.

[18] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc IEEE Conf CVPR*, 2017.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc IEEE Conf CVPR*, 2016, pp. 213–223, `www.cityscapes-dataset.com`.

[20] "Mapillary Vistas Dataset," `https://www.mapillary.com/dataset/vistas`.

[21] Z. Kato and J. Zerubia, "Markov random fields in image segmentation," *Foundations and Trends in Signal Processing*, vol. 5, no. 1–2, pp. 1–155, 2012.

[22] D. F. Laefer, S. Abuwarda, A.-V. Vo, L. Truong-Hong, and H. Gharibi, "2015 aerial laser and photogrammetry survey of Dublin city collection record," doi:10.17609/N8MQ0N, LiDAR dataset, 2015.

[23] J. Byrne, J. Connelly, J. Su, V. Krylov, M. Bourke, D. Moloney, and R. Dahyot, "Trinity College Dublin Drone Survey Dataset," hdl.handle.net/2262/81836, LiDAR dataset, 2017.