Roadside object geolocation from street-level images with reduced supervision

Waqar Ahmad, Vladimir A. Krylov
ADAPT Research Center, Dublin City University, Dublin, Ireland
waqar.ahmad2@mail.dcu.ie, vladimir.krylov@dcu.ie

***Abstract*—We propose a method for automated detection and geolocation of roadside objects from street-level images by leveraging historical records of these objects. Such partial and/or noisy geo-records are often held by infrastructure owners and require frequent updating. We aim to reduce the amount of image-level supervision required for the deployment of deep learning methods to geolocation problem from segmentation masks (very costly) to binary image labels (lower cost). Our proposed method integrates an image classification deep learning pipeline with Grad-CAMs and watershed transform to identify the positions of roadside objects of interest in the images. The geolocation is performed by deploying the existing Markov Random Field-based optimization module. We analyze the robustness of the proposed low-supervision geolocation model to noisy records. We report experiments for the detection of traffic lights and public bins, with geolocation of the latter performed in central Dublin.**

***Index Terms*—Object geolocation, street-level images, historic geo-records, Grad-CAM activation map, low supervision.**

I. Introduction

The last decade has witnessed a rapid expansion of the use of Deep Learning (DL) methods in a wide variety of applications, including vision-related tasks. In this work, we consider the application of these powerful techniques to the problem of detection and geolocation of roadside assets from street-level imagery. The records of roadside assets, including utility poles, traffic lights, public bins, etc., are highly valuable and require proper digital maintenance and regular updating. Such records inform public policies, facilitate the accessibility analysis and ultimately enable autonomous driving systems and advance navigation. Maintaining the high quality of asset records is a costly and time-consuming procedure which is often outsourced to 3rd party companies that update the inventory record manually. Automated methods leveraging DL algorithms can be used to dramatically reduce these intervention costs. In order to train a DL algorithm there is an inherent need for large amounts of training data. Specialized street-level repositories including Google Street View and Apple Maps own billions of images, covering hundreds of thousands of kilometers. Popular social media platforms, e.g., Twitter, Instagram, and Facebook, encourage their users to upload street-level images along with the corresponding location information (i.e., coordinates). Alongside these massive privately operated image data repositories, public crowd-sourced datasets, like Mapillary, OpenStreetCam, and 36cities contain geo-tagged images with various amounts of metadata that can be used to detect roadside objects.

A number of techniques have been previously proposed to automate the task of object geolocation solely from street-level imagery. Two fully convolutional neural networks were used in [1]: segmentation CNN to detect the traffic lights from the Google Street View (GSV) images, and monocular depth estimator to assess the distance of the detected object from the camera. A custom-grid Markov Random Field (MRF) model is then proposed to optimize the triangulation process for object geolocation. A simplified version is tested in [2] without MRF-optimization for street trees detection. Street-level panorama images and the corresponding street addresses are used to geolocate the tress in [3]. Traffic signs are detected and classified using RetinaNet in [4]. Using the GSV imagery, the authors of [5] used RetinaNet to detect the utility poles with cross arms. A modified version of RetineNet (GPS-RetineNet) has been proposed in [6]. An end-to-end method for traffic lights geolocation in videos is proposed in [7].

Fusion of multiple types of imagery can be beneficial when addressing the roadside object geolocation. In [8] LiDAR and street-level imagery were used together for the geolocation of traffic lights and utility poles, which resulted in improved performance compared to [1]. In [9] a multi-stage methodology is used to combine Airborne Laser Scanning data with aerial orthophotographs and street-level images for geolocation of urban trees. A Siamese architecture is proposed in [10] to match GSV images of the same objects from multiple views, which is then extended fusion with aerial imagery in [11].

The first step in geolocating any roadside object is to detect the object in the images. This can be done by leveraging DL models [1], including Mask-RCNN, SSD, etc. These methods require large amounts of properly labeled datasets, which poses a substantial challenge for their deployment. The purpose of this study is to replace the supervision-expensive DL segmentation pipeline with lower cost classification DL, which requires a single label per image. This is possible due to an ample amount of such imagery and / or positional information typically available from historic records. We further investigate the tolerance of such an approach to noisy / corrupted records, which may occur due to these being partially out of date. To enable object position detection inside the classified images we deploy Grad-CAM activation method and complement it with watershed transform to identify exact positions. We rely on the rest of the pipeline from [1] to complete the geolocation process, see Fig. 1.

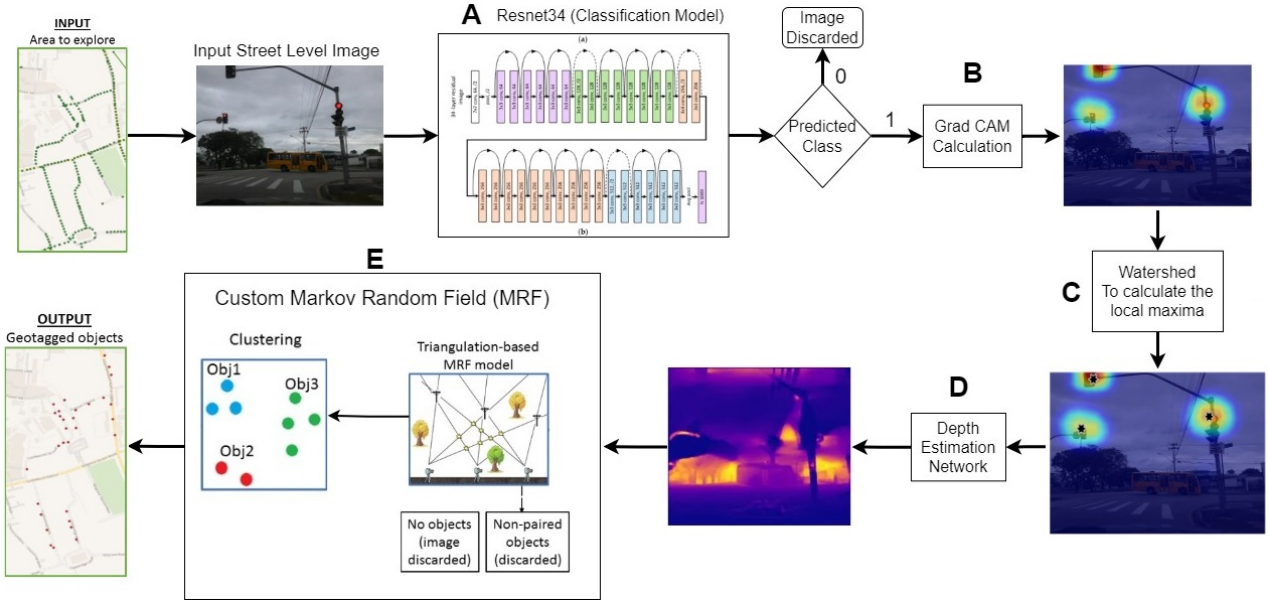

Fig. 1. Proposed geolocation pipeline. Step A: Image classification by Resnet34. Step B: The Grad-CAMs generation of the classified Image. Step C: Calculation of local maxima using Watershed. Step D: Depth estimation using pre-trained DL algorithm. Step E: Triangulation using Markov Random Fields.

II. Proposed Geolocation Pipeline

In this work, we propose an effective two step method to detect and geolocate roadside objects. This includes, object detection using DL algorithms, and object geolocation by resorting to triangulation. The overall proposed architecture pipeline is shown in Fig. 1.

In our study, we consider the training data having only two labels, that are, 0 (no object) and 1 (object presence). There is a single label per image, and the binary mask or the bounding boxes are not assumed available. To achieve the object detection and localization we followed a five-step pipeline. In step A, an image is classified based on the presence or absence of the object using Resnet34 classification model. In step B, the Grad-CAM is calculated for the image that contains the object to roughly position the activations in the image. Once the area is highlighted, the watershed algorithm is applied to find the local maxima in the highlighted area constituting step C. In step D, the approximate distance from the camera position to the local maxima is determined using a pre-trained depth model [12]. The final configuration of the objects is then reported by performing step E, i.e. custom-grid MRF optimization for object triangulation [1] based on detection from steps A-C and distances from step D.

Grad-CAM [13] is a method developed for activations maps analysis in DL models. The last convolutional layer of the model contains detailed information. Grad-CAM uses this information and generates a heatmap, localizing the area that is relevant for the prediction of a specific class. Specifically, Grad-CAM is calculated as the gradient of the class-specific output y^c w.r.t. the feature map activations A^k of a convolutional layer. Then the average of all the gradients over width i and height j dimensions is taken:

$$\alpha_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^c}.$$

Finally, ReLU function is applied to the weighted combination of feature map activation A^k:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k).$$

After performing Grad-CAM activation analysis on the classified images we then deploy the watershed algorithm to identify unique image positions for each activation blob, see Fig. 2. The watershed algorithm thresholds the image, followed by applying a distance transform that identifies the foreground regions. A local maximum is then calculated for each foreground region.

The estimated depth paired with camera coordinates and image bearing are input to triangulation. Triangulation is performed using the custom-grid Markov Random Field (MRF) [1]. This algorithm performs stochastic optimization to solve the numerical problem of finding optimal object positions as intersections of multiple view rays. Object information is contributing via energy terms that penalize discrepancies between depth estimation and intersection positions, detection false positives, and multiple objects simultaneously occupying the same position. For more details refer to [1]. The final geolocations returned by the MRF are fed through a local distance-based clustering to avoid double detections. The mean position of each cluster is the final output of our pipeline.

III. Experiments

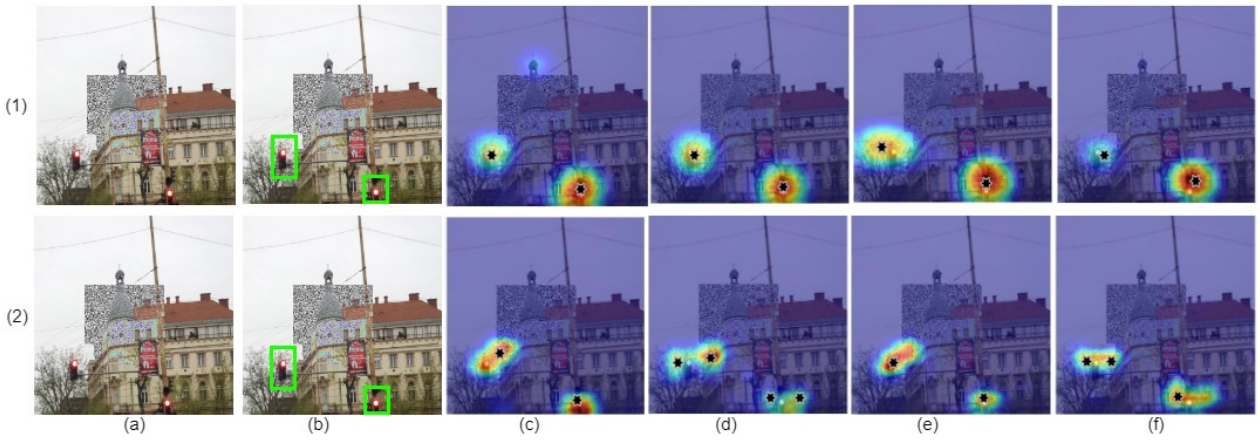

Fig. 2. Grad-CAMs of traffic light dataset. Row 1: Resnet34 Results. Row 2: Resnet50 Results. Column (a): Input Image, (b): True Traffic Light Position (c): No label noise, (d)-(f): 1%, 5%, and 10% label noise respectively. Black dots are the local maxima identified using watershed transform.

A. Datasets and setup

In this work we consider two roadside objects: traffic lights and public bins. The images of traffic lights are extracted from the Mapillary Vistas dataset. Based on the presence of traffic light, 9,000 images images, were extracted. The selection is done using the segmentation annotations provided. A total of 72,000 image crops of size 1200 x 1200 were generated, scaled 1/2 is each dimension for training. We thus obtained 24,000 images both in class 0 (no traffic light) and 1 (traffic light).

To build the public bins datasets we have extracted the images from the Mapillary API. We have extracted only those images that are flagged as the images containing the public bins. We also use the publicly available dataset of street public bins provided by the Dublin City Council (DCC). To enforce consistency with the DCC dataset we have selected only those Mapillary images that are within 15 meters distance of the locations of the DCC bins. A total of 9,827 images were obtained for class 1. Class 0 images are obtained by extracting the images that are 30-100 meters away from DCC bins. Each image is re-scaled to 1500 x 900 resolution for processing.

Importantly, due to the lack of high quality segmentation ground truth available for traffic lights (from Mapillary Vistas dataset), we cannot assess the segmentation quality obtained using the proposed method for the public bins data. We therefore rely on these two datasets for different purposes: traffic lights for assessing the segmentation performance, and public bins to assess the geolocation quality.

B. Traffic lights detection

We employ the traffic lights dataset to compare the performance of the proposed object detection method with reduced supervision to the use of full segmentation pipeline. We have deployed Resnet34 and Resnet50 models, with the corresponding classification scores reported in Table I. We observe that both models are performing comparably. Due to the model complexity, Resnet34 is preferred over Resnet50.

For the traffic lights dataset we then experiment with adding noisy labels. We consider three corrupted datasets, based on the class labels noise (1%, 5%, and 10%). Specifically, for 1% noisy dataset, the labels of random 1% of the total number of images in both classes were swapped: 0.05% images from class 1 were swapped with 0.05% images from class 0. The similar process is done for the 5% and 10% noisy datasets. The same amount of image labels were also manipulated for Mask-RCNN analysis. For example, for the 10% noisy dataset, 5% of true binary masks were replaced by complete black masks, and 5% images having no desired object were added to the dataset, and superpixel-based random binary masks were generated and added as the true labels for these images. In Table I we can see that both models demonstrate reasonable tolerance to noising, with a decrease in the performance by several percentage points when 10% of the labels are corrupted. These results confirm that Resnet34 can be applied to noisy realistic datasets (like public bins), without dramatic loss in terms of classification performance.

Post image classification, we calculated the Grad-CAMs and binarized the highlighted areas. We have calculated the IOU-based precision, recall, and detection rate for the test dataset, and compared the performance to the Mask-RCNN model. From Table II, we can see that Resnet34 is lagging in detection rate performance by only around 2.5 percent, but reports higher precision compared to Mask-RCNN. Importantly, this result is achieved with significantly reduced requirement on image supervision (for object segmentation task) while Mask-RCNN uses full supervision (bounding boxes, binary mask, and class labels, etc). Compared to the Resnet50, Resnet34 is performing better in terms of IOU score. This is an additional reason to select Resnet34 as our main DL classifier.

The qualitative results of object detection and localization are shown in Fig. 2. The rows represent the results from Resnet34 and Resnet50 and the columns show the input image and true position of the traffic lights followed by the corresponding Grad-CAMs for no label noise, 1%, 5%

TABLE I
CLASSIFICATION PERFORMANCE ON TRAFFIC LIGHTS DATASET

Model	Label Flipped	Classification Score (%)			
		Precision	Recall	F1-score	Accuracy
Resnet34	None	86.18	87.33	86.75	91.59
	1%	84.61	88	86.27	91.00
	5%	85.03	83.33	84.17	90.12
	10%	79.26	86.66	82.80	88.65
Resnet50	None	85.80	88.66	87.21	91.80
	1%	90	84	86.89	92.01
	5%	84.86	86.00	85.43	90.75
	10%	90.29	80.6	85.21	91.17

TABLE II
INSTANCE-BASED IOU SCORE FOR TRAFFIC LIGHTS DATASET

Model	Labels Flipped	Instance Based IOU Score (%)		
		Precision	Recall	Detection Rate
Resnet34	None	92.35	78.00	79.39
	1%	95.30	75.96	75.04
	5%	92.33	77.63	74.29
	10%	96.78	74.01	75.51
Resnet50	None	85.53	65.20	62.24
	1%	81.63	61.02	58.35
	5%	73.82	58.80	55.75
	10%	73.51	63.41	62.07
Mask RCNN	None	83.03	83.72	81.40
	1%	83.11	81.85	79.45
	5%	83.85	80.90	78.79
	10%	80.63	81.99	79.45

and 10% label noise, respectively. The black dots on each image represent the local maximum returned by the watershed algorithm. In row 2 columns (c)-(e), we can see two distant point-locations identified a single object. This is due to the inaccurate gradient shape calculated from the Resnet50 model.

Sample segmentation results for the traffic lights with Resnet34 and Mask-RCNN are shown in Fig. 3. Rows (2)-(3) show the results for the original dataset (no label noise), and 10% label noise respectively. We observe that the shapes returned by the Resnet34 are less accurate. However, since we target object geolocation in this study, we are solely interested in the local maxima of the Grad-CAMs and not the full delineation of the object. In Fig. 3(a), false areas are identified by Mask-RCNN. In Fig. 3 column (b), the traffic lights are near to each other, and thus both models are not performing very well.

C. Public bins detection and geolocation

In this task we consider two test sets: extracted from Mapillary API and DCC images. The latter has been captured for DCC in central Dublin (Ireland) in 2020 by Murphy Geospatial (https://murphygs.com). This dataset covers approx. 2.7kms of roads, with the images sampled approx. every 3 meters. Each sample location has 6 views captured with field of view of 68.77 degrees, covering the entire panorama. Based on the above experimental analysis Resnet34 is trained on the Mapillary data and tested on both test sets. The classification scores obtained are reported in Table III. From the table, we can see that the model is performing well on both test sets.

For object detection part, instance-based IOU score is calculated only on the Mapillary test set due to the availability

Fig. 3. Segmentation from Resnet34 Grad-CAMs (blue ellipses), and Mask-RCNN (green rectangles). (1): Image, (2-3) Segmentations obtained by using: no label noise and 10% label noise datasets, respectively.

TABLE III
INSTANCE BASED IOU SCORE FOR THE PUBLIC BINS DATASET

Model	Test Set	Classification Score (%)			
		Precision	Recall	F1-Score	Accuracy
Resnet34	Mapillary	92.32	86	89.04	90.5
	DCC	95.69	89	92.22	92.5

of the ground truth segmentation masks. We achieved 65% precision, 92% recall, and 83% detection rate. The reason for this low precision is a substantial presence of inaccurate labels in the dataset. The qualitative results of the Grad-CAMs are shown in Fig. 4. Column 1 shows the input images followed by corresponding Grad-CAMs and watershed-identified local maxima. A false positive is highlighted in the center of the image in the last figure 4(b). Such false positives have been filtered out by the depth estimation model by considering only the objects within 25 meters of the camera position.

We now proceed to the testing of the full geolocation pipeline. There are a total of 82 DCC stationary public bins in the area covered by DCC images. The results from our procedure (MRF) are plotted along with the DCC bins over the OpenStreetMap layer in Fig. 5. Out of the 82 DCC bins, the presence of 63 bins is also identified by MRF. The average distance between the positions is 5.45 meters. The following comparative analysis is performed by manual inspection using

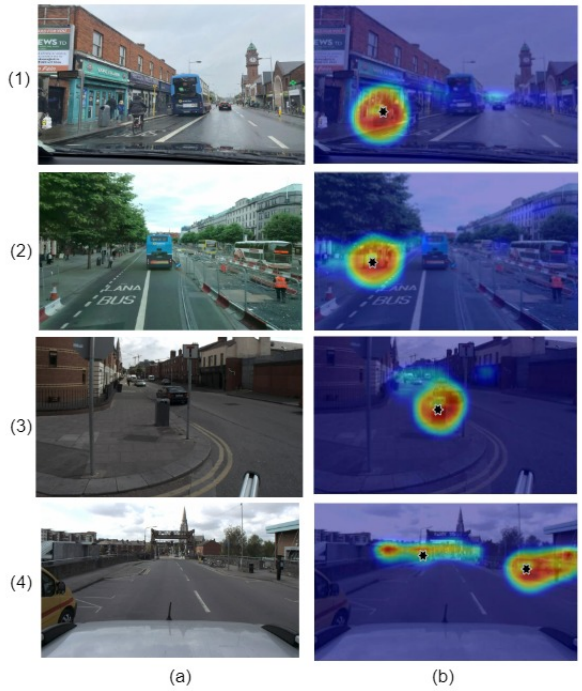

Fig. 4. Grad-CAMs and local maxima identified via watershed (black dot) on public bins. Images from Mapillary (1-2), and the DCC dataset (3-4).

the available Google Street View data in the area. The MRF-estimated locations of 40 bins are more accurate compared to the DCC bins, whereas 22 DCC dataset bin geolocations are more accurate. In addition, 43 genuine public bins are geolocated in this area which are absent from the DCC record. MRF has resulted in 46 false positives. This last number is severely impacted by the size of the training dataset, i.e. several types of objects were systematically confused for public bins (leading to over 2/3 of all the false positives), including bollards, bases of street lights, and cabinets. It is expected that these will be mitigated if the training data of class 'no-bin' includes more examples of these objects. Furthermore, the same types of training dataset deficiency have resulted in 19 false negatives (i.e., public bins in the DCC record but not found by MRF).

These results are encouraging in suggesting the proposed geolocation pipeline has a strong capacity to validate and improve the existing geo-record by confirming the existing objects and identifying missing or erroneous records through automated processing of the most recent imagery available. The current limitations are due to the limited availability of imagery that has been available for this experimental study.

IV. Conclusions

In this work, we have proposed an efficient method to achieve object detection using reduced supervision. The images are classified based on the binary labels, i.e. presence/absence of the roadside object of interest. Grad-CAMs followed by watershed transform have been employed to highlight the location of objects inside the images. We have also

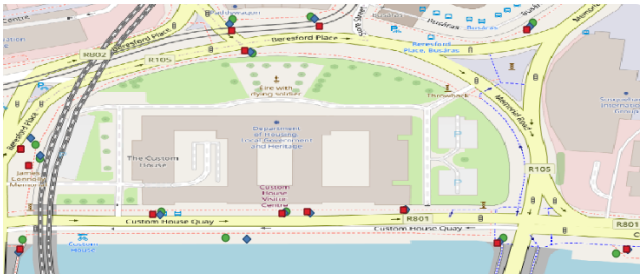

Fig. 5. Public bins in central Dublin: DCC public records (green circles) vs. MRF (red squares) vs. manually identified ground truth (blue diamonds).

used monocular depth estimation and MRF-based triangulation procedure to geolocate the objects. We have experimentally observed that the reduced supervision procedure is capable of delivering object detection performance with minimal loss of accuracy and with reasonable tolerance to noise on traffic lights detection. The performance of the proposed geolocation pipeline has been validated on public bins detection in central Dublin. These results show a strong potential in automating the records updating using the proposed pipeline, however further experimentation at larger scale is required.

References

[1] Krylov V., Kenny E., Dahyot R. Automatic discovery and geotagging of objects from street view imagery. Remote Sensing 10, 5, 2018.

[2] Lumnitz S., Devisscher T., Mayaud J., Radic V., Coops N., Griess V. Mapping trees along urban street networks with deep learning and street-level imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 175, 144-157. 2021.

[3] Laumer D., Lang L., van Doorn N., Mac Aodha O., Perona P., Wegner J. Geocoding of trees from street addresses and street-level images. ISPRS Journal of Photogrammetry and Remote Sensing, 162, 125-136, 2020.

[4] Pedersen K., Torp K. Geolocating Traffic Signs using Crowd-Sourced Imagery. Proc. of International Conf. Advances in Geographic Information Systems, New York, USA, 199–202, 2020.

[5] Zhang W., Witharana C., Li W., Zhang C., Li X., Parent J. Using Deep Learning to Identify Utility Poles with Crossarms and Estimate Their Locations from Google Street View Images. Sensors. 18(8):2484, 2018.

[6] Wilson D.; Alshaabi T.; Van Oort C.; Zhang X.; Nelson J.; Wshah S. Object Tracking and Geo-Localization from Street Images. Remote Sens. 14, 25-75, 2022.

[7] Chaabane M., Gueguen L., Trabelsi A., Beveridge R., O'Hara S. End-to-end learning improves static object geolocalization from video. Proc. of the IEEE Winter Conference on Applications of Computer Vision, pp. 2063–2072, 2021.

[8] Krylov, V., Dahyot, R. Object Geolocation Using MRF Based Multi-Sensor Fusion. Proc. of IEEE International Conf. Image Proc. 2745-2749, 2018.

[9] Rodríguez-Puerta F., Barrera C., García B., Pérez-Rodríguez F., García-Pedrero A.M. Mapping Tree Canopy in Urban Environments Using Point Clouds from Airborne Laser Scanning and Street Level Imagery. Sensors, 22(9):3269, 2022.

[10] Nassar A.S., Lefèvre S., Wegner J. Multi-View Instance Matching with Learned Geometric Soft-Constraints. ISPRS International Journal of Geo-Information, 9(11):687, 2020.

[11] Nassar, A.S., D'Aronco, S., Lefèvre, S., Wegner, J.D. (2020). GeoGraph: Graph-Based Multi-view Object Detection with Geometric Cues End-to-End. European Conf. on Computer Vision 2020.

[12] Farooq Bhat S., Alhashim I., Wonka P. "AdaBins: Depth Estimation Using Adaptive Bins," IEEE/CVF Conf on Computer Vision and Pattern Recognition, pp. 4008-4017, 2021.

[13] Selvaraju R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," IEEE International Conf on Computer Vision, Venice, Italy, pp. 618-626, 2017.