Semantic Localisation via Globally Unique Instance Segmentation

Ignas Budvytis[1]
ib255@cam.ac.uk
Patrick Sauer[2]
patrick.sauer@toyota-europe.com
Roberto Cipolla[1]
rc10001@eng.cam.ac.uk

[1] Department of Engineering
University of Cambridge
Cambridge, UK

[2] Toyota Motor Europe,
Belgium

Abstract

In this work we propose a novel approach to semantic localisation. Our work is motivated by the need for environment perception techniques which not only perform self-localisation within a map but also simultaneously recognise surrounding objects. Such capabilities are crucial for computer vision applications which interact with the environment: autonomous driving, augmented reality or robotics.

In order to achieve this goal we propose a solution which consists of three key steps. Firstly, a database of panoramic RGB images and corresponding *globally unique, per-pixel object instance labels* is built for the desired environment where we typically consider objects from static categories such as "building" or "tree". Secondly, a semantic segmentation network capable of predicting more than 3000 labels is trained on the collected data. Finally, for a given panoramic query image, the corresponding instance label image predicted by the network is used for semantic matching within the database. The matching is performed in two stages: (i) a fast retrieval of a small subset of database images (~100) with highly overlapping instance label histograms, followed by (ii) an explicit approximate 3 DoF (yaw, pitch, roll) alignment of the selected subset of images and the query image. We evaluate our approach in challenging indoor and outdoor navigation scenarios, achieving better or similar performance when compared to state-of-the-art image retrieval-based localisation approaches using key-point matching [29, 63] and image level embedding [3].

Our contribution includes: (i) a description of a novel semantic localisation approach using *globally unique instance segmentation*, (ii) corresponding quantitative and qualitative analysis and (iii) a novel CamVid-360 dataset containing 986 labelled instances of buildings, trees, road signs and poles.

1 Introduction

As applications of Computer Vision algorithms transition from passive perception such as face recognition [48] to active decision making such as autonomous driving [7], augmented reality [20] and robotics [49], standard frameworks for object recognition [48], semantic segmentation [61] and localisation [3, 26, 63] are becoming too limiting. For example, the output of a typical object recognition [48] framework comes in the form of a labelled

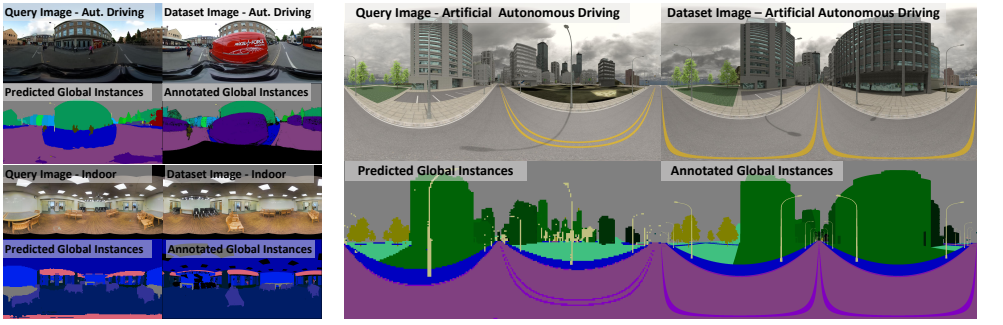


Figure 1: This figure shows *globally unique instance segmentation* and corresponding localisation results for outdoor (top left quadrant), indoor (bottom left quadrant) and artificial data (right quadrant). The left two images of each quadrant correspond to the query image and the predicted labels while the image pair on the right corresponds to the best matching image in the database and the corresponding labels. Note the high quality of the segmentation and localisation under changes in illumination (outdoor images), viewpoint, in the presence of smooth surfaces with little texture (indoor images) and in the case of significant changes to the map due to missing buildings (artificial images).

bounding box which does not contain enough information for most robot interaction tasks such as grasping. Conversely, the non object-specific output of typical semantic class segmentation [61] (or detection [50]) methods severely limits the interaction possibilities of an autonomous agent. Similarly, object instance segmentation [23] frameworks often produce inconsistent labels accross queries, thus preventing decision making accross time. Finally, standard localisation approaches [3, 26, 63] indicate only a location in a map, either as a 6DoF camera pose prediction [26] or by specifying the most similar image in a database [3, 63] without providing semantic information about the surrounding environment.

In this work, we propose to address the aforementioned problems by utilizing *globally unique object instance segmentation* - a sub-task of semantic segmentation. For such a task, each object of interest within the operating environment is assigned a label mask with a globally unique instance id, as shown in Figure 1. Using this data for training a semantic segmentation network, we obtain a model which may be used for simultaneous localisation, surrounding object recognition and segmentation. Localisation is performed in a two-step procedure. First, fast histogram matching is performed between the predicted labels of the query image and the database label images, resulting in a small number (~100) of well matching candidate frames. Then, a further refinement using a slower but more accurate label image alignment in 3 degrees of freedom (yaw, pitch, roll) is carried out.

We evaluated the proposed algorithm on multiple real and artificial datasets. We created an artificial autonomous driving dataset using the SceneCity [2] tool and collected real world data by tracing the original path of the CamVid [9] dataset using a panoramic Ricoh Theta S camera. We used the Stanford 2D-3D-S [6] dataset for indoor experiments. We obtained high localisation accuracy (above 98%) and high segmentation accuracy (above 94% global and 52% class average for 837 object labels) on autonomous driving scenarios. While we achieved relatively lower segmentation accuracy (61% global and 38% class average for 3138 object labels) for challenging, low-texture indoor experiments, we still obtained an 11% higher localisation accuracy than classical approaches of localisation based on keypoint matching [29, 63] or image embedding [3]. Our contribution includes: (i) a description of a novel semantic localisation approach using *globally unique instance segmentation*, (ii) corresponding quantitative and qualitative analysis and (iii) a novel CamVid-360 dataset with

986 labelled instances of buildings, trees, road signs and poles.

The rest of this work is divided as follows. Section 2 discusses relevant work in semantic segmentation and localisation. Section 3 provides details of our proposed localisation approach. Sections 4 and 5 describe the experiment setup and corresponding results.

2 Related Work

In this section we provide a discussion of various approaches to semantic segmentation and localisation which are related to our work.

Semantic segmentation. In the field of Computer Vision, semantic segmentation encompasses various approaches to grouping pixels in images or videos. For example, class segmentation [9, 52] is concerned with grouping pixels of the same object class (e.g. buildings, trees or tables), whereas instance segmentation [22] aims to group pixels belonging to the same object within the target image. Panoptic segmentation [28] is concerned with combining class segmentation output and instance segmentation output. In contrast to the aforementioned semantic segmentation approaches, *globally unique instance segmentation* aims to group pixels of the same object across the whole operating environment, thus providing both segmentation and recognition simultaneously.

Image retrieval based localisation. A large subset of localisation approaches formulate localisation as an image retrieval problem. They identify the most similar looking image in a database primarily in two ways by employing either (i) a pipeline of keypoint detection and matching [29, 63] or (ii) fast-to-compare image level encoding [9, 55]. Approaches of the first type often have large storage requirements and relatively slow matching procedures. Methods of the latter type are significantly faster, yet as suggested in [37] are more likely to be sensitive to large occlusions and scene changes[1]. Both types of approach provide a location for the whole scene and not of individual objects.

3D geometry based localisation. Another large group of localisation approaches work by explicitly recovering 3D geometry of the environment using 3D sensors such as structured light [24, 46], time of flight [60] cameras as well as RGB based structure from motion [54] or mixed [6] approaches. Such works perform 3D matching between locally reconstructed 3D scenes and pre-built 3D maps either by using depth information only [24], or by using a combination of depth and appearance information [18]. A 6DoF camera pose is recovered as result. The majority of such approaches are computationally expensive and require relatively large amounts of memory for storage of the environment map. Some recent techniques [26] attempt to improve query time performance by recovering 6DoF camera pose using deep neural networks without an explicit 3D geometry estimation. In contrast to this, and as with image retrieval-based localisation, our proposed method not only provides a camera location estimate but also indicates surrounding objects. It also may be used to augment the aforementioned approaches, or use them to perform a refinement of the outputs.

Semantic localisation. Works attempting to incorporate semantic information into localisation often take one of three approaches. Methods in the first group perform either explict filtering [35] or use feature reweighting [27, 47] in order to filter out uninformative classes of objects (e.g. cars, people) when performing image matching. Methods in the second group attempt to fit 3D models (e.g. CAD) of a single room [44] or building [13, 14] or of detailed maps [12, 38] as well as a combination of less precise information such as a single

[1] It is important to note that in our experiments described in Section 5 with NetVlad [9], a high tolerance for changes in the images was observed. We leave a deeper investigation of this observation for future work.

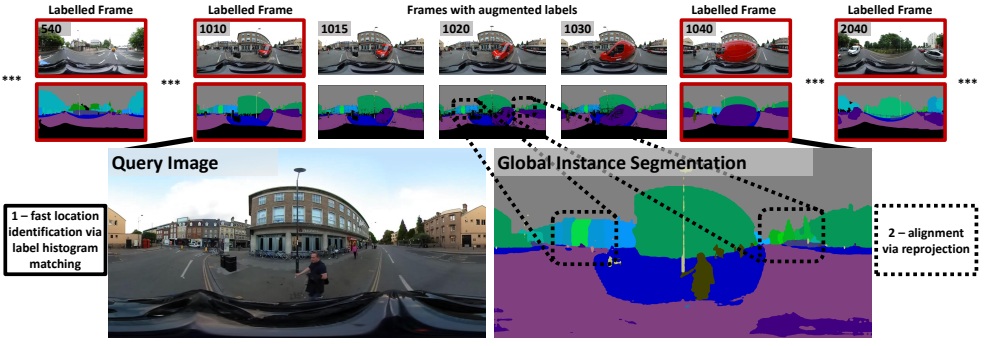


Figure 2: This figure illustrates the key steps of our proposed localisation framework. Firstly, a database of panoramic RGB images and corresponding *globally unique, per-pixel object instance labels* is built for the desired environment. A multi-class label propagation [10] method is used to propagate labels between key labelled frames (marked with red borders). Secondly, a semantic segmentation network capable of predicting more than 3000 labels is trained on the collected data. Finally, for a given panoramic query image, the corresponding instance label image predicted by the network is used for semantic matching within the database. The matching is performed in two stages. During the first step, a fast label histogram matching is performed in order to predict an approximate location. During the second step, the predicted location is further refined by finding the label image in the database which best agrees with the predicted instances. See Section 3 for more details.

large building floor plan [32, 57] or of multiple building 2D layout and corresponding height information [4, 5, 23]. The final group of methods employ direct learning of classifiers to output a GPS location [21, 59], or a bounding box of a specific building [64]. The first family of approaches aims to improve localisation accuracy, yet does not provide semantic information (e.g. identification of surrounding objects) about the environment. Methods in the second group require expensive-to-build maps and complex fitting techniques. Note that while [57] uses easy-to-obtain existing building plans it heavily relies on locations having informative textual information. The examples from the final group demonstrate only coarse localisation capability. Our method can be viewed as a member of this group.

Datasets. While the advent of deep learning has seen a proliferation of various datasets, obtaining a dataset which would contain both (i) panoramic videos[2] (or large sets of images densely sampled at different views) and (ii) large amount globally unique object instance labels for indoor or autonomous driving scenarious is not easily available. For example, many of the popular real-world datasets for autonomous driving [9, 15, 19, 30, 36] or localisation [12, 37, 56] lack one or the other. Even artificial datasets [17, 25, 40, 42], or simulators [16, 34, 39, 41, 51] often do not provide the ability to sample equirectangular images, obtain instance labels of static objects or provide only very limited urban landscapes. In order to evaluate our localisation method, we collected our own real world and virtual datasets for autonomous driving. For real indoor environments, we were able to identify only the Stanford 2D-3D-S [6] dataset to contain instance segmentations of equirectangular images, which we use in this work. Note that the Matterport 3D dataset [11] provides labelled meshes, however labels for equirectangular images are not provided. Artificial highly realistic indoor datasets [31, 39, 45] suffer from similar issues as autonomous driving, preventing from convenient sampling of rich, large scale, photo-realistic, equirectangular videos with various object instance labels.

[2]We chose panoramic as opposed standard small field of view images for our experiments in order to emphasise the potential of semantic localisation which is more sensitive to limited views than RGB texture based methods.

3 Method

Our proposed semantic localisation framework comprises three key steps: (i) collecting a database of labelled images from the environment, (ii) training a globally unique instance segmentation network and (iii) matching the prediction on the a given query image with the best labelled image in the collected database. A detailed description of each step is provided bellow. See Figure 2 for an illustration.

Collecting labelled database. The database collection for the desired environment consists of three steps. Firstly, a set of panoramic images of the environment is obtained by capturing panoramic videos (real data) or by dense sampling of still images (artificial data). Secondly, a subset of those images (e.g. every 30th frame) are labeled with globally unique instance labels. Finally, an original implementation of label propagation method [10] is used to extend the labels for the rest of the image set at full resolution with parameters of patch size $p = 7$, cross-correlation window dimensions $W \times H = 200 \times 100$ and label similarity constant $\delta = 0.001$ set empirically by following recommendations of [10]. Note that for experiments described in this work, key frames were hand labelled, however semi-automatic methods making use of instance segmentation such as Mask-RCNN [22] or 3D map alignment [38] could be employed in the future.

Training a segmentation network. Once a dataset of labelled images is collected, a semantic segmentation network is trained. We use the *Model A* variant of wide shallow residual networks [61]. It is trained for 200 epochs using four standard Titan X GPUs and batch size of 12 samples. Weights are initialized from a network pretrained on ImageNet [43] dataset. Initial learning rate of 0.0001 is chosen and gradually reduced using linear schedule [33]. Note that we amended the original implementation of [61] by removing the upsampling layer. The resulting network works well up to 4000 class labels per pixel at a moderate computational cost. This is enough to cover significant areas of interest, especially when only a subset of objects of a class of interest receives globally unique instance labels. Also note that multiple local networks can be learned to cover even larger areas, and techniques based on hierarchical classification [53] and embedding [8] can be used to further increase scalability.

Localisation. For any query image, a two step procedure is performed. Firstly, a small set (~100) of the highest-matching database images in the label histogram domain is found using the distance function M_{hist}. Then, an explicit alignment of query and database label images is performed in order to find the best match under any possible rotation (distance function M_{alig}). The label histogram matching score M_{hist} ranks pairs of label images by the number of different labels appearing in both images ($|L_{com}|$), followed by a ranking using the average (per-label) upper bound of intersection over union score ($M_{IoUHist}$). It is computed as follows:

$$M_{hist} = \alpha M_{IoUHist} + \beta |L_{com}|, \tag{1}$$

where

$$M_{IoUHist}(Q,D) = \frac{\sum_{\forall l \in L_{com}} IoU_{upper}(Q,D,l)}{|L_{com}|}. \tag{2}$$

Here,

$$IoU_{upper}(Q,D,l) = \frac{\min(|Q(l)|,|D(l)|)}{\max(|Q(l)|,|D(l)|)} \tag{3}$$

is an upper bound of the intersection over union (IoU) score[3] between the query (Q) and database (D) images for label l. $Q(l)$ and $D(l)$ are the sets of pixels having class label l

[3]Note that the intersection over union for label l can be defined as $IoU = \frac{|Q(l) \cap D(l)|}{|Q(l) \cup D(l)|}$.

for images Q and D respectively. $L_{com} = (L_Q \cup L_D) - L_{Q_{filt}}$, where L_Q is a set of labels in the query image, L_D is a set of labels in the database image and $L_{Q_{filt}}$ is a set of labels which have less than d pixels (we use $d = 20$) in the query image. $L_{Q_{filt}}$ labels are excluded in order to gain robustness to miss-prediction of small objects. Also note that unlabelled pixels in database images are assigned a "void" label and are excluded from the matching calculation, as are dynamic classes (e.g. cars, pedestrians for autonomous driving scenarios). We empirically set $\alpha = 1.0$ and $\beta = 1.0$. In order to rank images during the second step, the upper bound of the intersection over union ($M_{IoUHist}$) is replaced with an estimation of the IoU score which utilises full label images (M_{IoUEst}), resulting in the following definition of

$$M_{alig} = \alpha M_{IoUEst} + \beta |L_{com}|. \tag{4}$$

Here

$$M_{IoUEst}(Q,D) = min_{\forall R \in ROT} \frac{\sum_{\forall l \in L_{com}} IoU_{est}(R(Q),D,l)}{|L_{com}|}, \tag{5}$$

where ROT is the set of 3D rotations (yaw, pitch, roll) considered, $R(Q)$ is a label image rotated by R and

$$IoU_{est} = \frac{|R(Q)(l)| * \frac{|RQ^{'}(l) \cap D(l)|}{|RQ^{'}(l)|}}{|R(Q)(l) \cup D(l)|} \tag{6}$$

is an estimation of an IoU score for a class label l for images $R(Q)$ and D. Here $RQ^{'}(l)$ is a small subset (40 samples, set empirically) of randomly sampled pixels from $R(Q)(l)$. Note that for the experiments described in this work ROT consists of angles sampled at every 1° in interval ($\pm 20^\circ$, $\pm 20^\circ$, $\pm 180^\circ$) for (yaw, pitch, roll). It is important to note that label histogram matching algorithm is very efficient, taking less than 1 millisecond to evaluate more than 10000 matches even with a simple implementation. Dedicated data structures designed for fast item retrieval could be used for further improvement. While label image alignment matching is slower, it provides more accurate image level matches and can output the 3DoF camera view details. Our simple implementation achieved 60 FPS for 100 matches on a Titan X GPU and can also be significantly sped up by further optimising the code, which is out of the scope of this work.

4 Experimental Setup

In this section we describe the datasets used for indoor and autonomous driving scenarios and explain the protocol for evaluation.

SceneCity dataset. SceneCity [2] is a Blender [1] plugin for creating artifically generated 3D urban landscapes. The generation process can be controlled via multiple environmental parameters such as the amount of water and mountains, the road network density, the building size distribution and other parameters. This tool was used to generate three city maps for our experiments, examples of which can be seen in Figures 1 and 3 as well as in supplementary material. The first city map was borrowed from [65]. This city contained 102 buildings and 156 road segments[4]. To create the map for the second city, 20% of buildings were removed from the first city at random. The final city map was generated from scratch using the SceneCity [2] tool, resulting in 827 buildings and 966 road segments. Semantic

[4]Note that each road segment is a unit square (1×1) in original SceneCity coordinates correspond to the size of typical road segment in the real world.

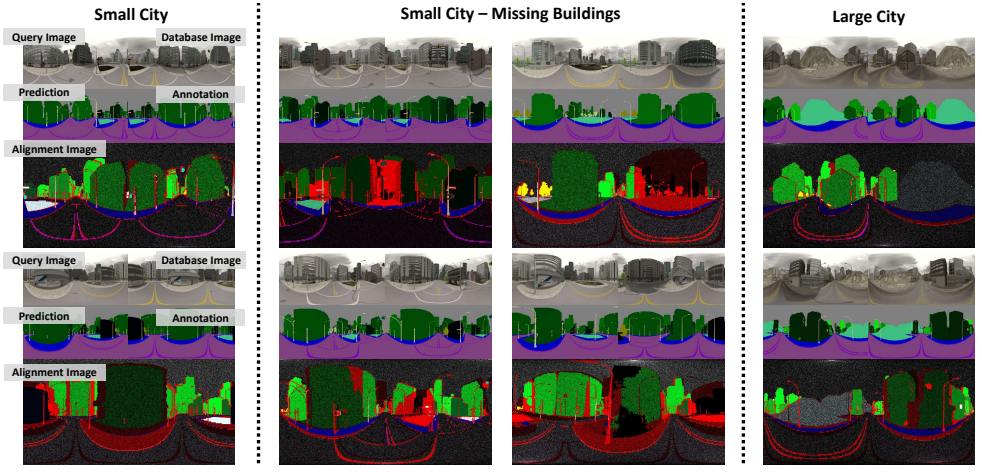


Figure 3: This figure illustrates qualitative results for globally unique instance segmentation and localisation in three artificial SceneCity datasets corresponding to a small city, a small city with missing buildings and a large city model. Alignment images are built by finding the best alignment between the query label image and the database label image, followed by subsequent colouring of pixels of the wrongly predicted label in red. Pixel colours are enhanced for better visibility. Note the high quality of segmentation and alignment for all images. Also note how missing buildings can be identified by large regions of red. Zoom in for better view. See supplementary material for more results.

labelling of 11 class labels (e.g. road, pole, etc.) were used for all three cities. Corresponding image databases were created by sampling images (2048×1024) at the centers of road segments and at 6 intermediate point coordinates along the centers (with an elevation of 0.064 units above the road) of two neighbouring road segments, resulting in a total of 1146 and 6774 images for the small and large cities, respectively. For all experiments, a cloudy sky lighting with randomly changing rotation of the sun was chosen. Query images for the small city were obtained by sampling 300 images from the original camera path provided in [63]. Query images for the large city consisted of 1000 samples at the center of road segments taken at an elevation of 0.064 units above the road and at a random deviation up to $(\pm 0.14, \pm 0.14, \pm 0.014)$ units in (X, Y, Z). The aforementioned dataset was generated in order to evaluate our proposed method for localisation under large, highly repetitive maps as well as under changes in the environment between data captures.

CamVid-360 dataset. CamVid-360 is a dataset of panoramic videos captured by cycling along the original path of the CamVid [9] dataset using a Ricoh Theta S camera. It provides an evaluation of proposed method for a real world autonomous driving scenario. As in the SceneCity experiments, CamVid-360 is divided into two sets of images. The database consists of 7835 images (sampled at 30 fps, 1920×960) tracing sequences 016E5, 001TP from the original CamVid dataset [9]. The query image set consists of 266 images (sampled at 1 fps) tracing sequences 016E5, 001TP and 006R0, the latter corresponding to parts of the city which are not represented in the database. The database images are labelled using the protocol explained in Section 3 with one of 11 semantic classes and multiple globally unique object instance labels of buildings and trees, resulting in 298 and 142 instances, respectively. Examples of CamVid-360 images and labels can be found in Figures 1, 2, 4 and in the supplementary material. Finally, in order to create the ground truth for localisation, each query image was manually assigned the best matching image in the database.

Stanford-2D-3D-S dataset. The Stanford 2D-3D-S [6] dataset consists of 1413 panoramic

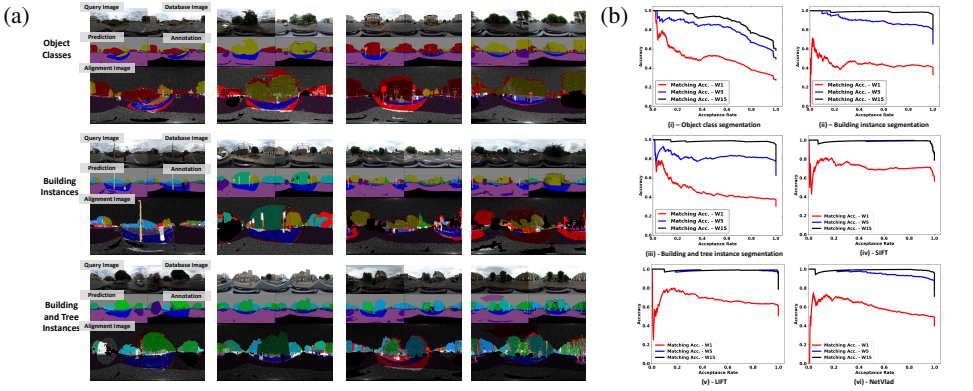


Figure 4: Figure (a) illustrates qualitative results on the CamVid-360 dataset. It follows the same visualisation format as Figure 3. Note a high accuracy overlap of class labels in different locations (row 1, column 2). Also note that errors in image alignment seem to be mainly introduced by the horizontal translation relative to the motion path in the database. Figure (b) shows matching accuracy as a function of changing the confidence threshold normalized for the percentage of images accepted (*acceptance rate*) at multiple error tolerances of ± 1, ± 5 and ± 15 frames.

images from 6 indoor areas and 270 locations. 13 class labels (e.g. ceiling, table, etc.) and 6005 corresponding object instance labels are provided. This dataset is used for evaluating our proposed localisation method for indoor robotics scenario. As with previous experiments, we randomly split the dataset into an environment image database (1180 images) and a query image set (233 images). When creating the split, it is ensured that at least one image of each location is contained in both subsets of the datasets. Also images were resized to 2048×1024. The proposed localisation method is evaluated under multiple scenarios: when using only class labels (11 in total), using class labels and instances of walls, ceilings and floors (2184 in total) as well as when using all static object instances (3138 in total). Note that we define an object as static if it is not likely to be moved in a room (e.g. bookshelf). Examples of images from Stanford-2D-3D-S and corresponding labellings can be found in Figures 1, 6, and in the supplementary material.

Evaluation protocol. We evaluated our proposed semantic localisation method by measuring the *matching accuracy* (MA), defined as $MA = \frac{CM}{CM+IM+NM}$, where CM and IM are numbers of correctly and incorrectly matched query images which have a ground truth match in the database, and NM is the number of images which were assigned a match but do not have a ground truth match in the database (only applies for CamVid-360 experiments). For the SceneCity experiments, a match was considered to be correct if query image was matched to the closed database image in Euclidean distance. For the CamVid-360 experiments, the matching accuracy was evaluated under three different settings, each considering a match to be correct if the predicted database image was within a range of ± 1, ± 5 and ± 15 video frames from the ground truth match, respectively. Finally, for indoor experiments, a match was considered to be correct if the query image and the recovered database image were taken in the same room. For all experiments (see Figures 5 and 6 (b)), accuracies were reported for the 95% *acceptance rate* (AR), which corresponds to a ratio between the number of query images which have a ground truth match in a database and pass a chosen confidence threshold with the total number of images which have a ground truth match in a database. The acceptance rate is used in order to avoid low confidence predictions affecting matching accuracies. Figure 4(b) illustrates how varying *acceptance rate* affects *matching accuracy*.

(a)

Localisation Method	Match Function	Small City					Small City – Missing Buildings (20%)					Large City				
		Lab.	MA	GA	CA	IoU	Lab.	MA	GA	CA	IoU	Lab.	MA	GA	CA	IoU
Class Only	Hist.	11	77	96	78	69	11	24	97	77	68	11	80	97	78	71
	Alig.		94					44					100			
Building Instances	Hist.	112	100	96	79	69	112	86	94	66	50	**837**	98	94	52	37
	Alig.		100					100					100			
SIFT	# of Dsc.	N/A	100	N/A			N/A	94	N/A			N/A	98	N/A		
LIFT	# of Dsc.		83					73					97			
NetVLAD	Euc. Dist.		100					99					58			

(b)

Localisation Method	Matching Function	Lab.	W +/- 1 MA	W +/- 5 MA	W +/- 15 MA
Class Only	Hist.	11	7	16	31
	Alig.		33	60	70
Building Instances	Hist.	308	18	48	81
	Alig.		43	83	98
Building and Tree Instances	Hist.	450	17	47	84
	Alig.		38	80	98
SIFT	# of Desc.	N/A	71	99	99
LIFT	# of Desc.		64	98	99
NetVLAD	Euc. Dist.		50	90	98

Figure 5: Figure (a) illustrates quantitative results of comparing various localisation approaches on three artificial cities. It reports matching accuracy (MA) for the best matches and segmentation quality (GA - global accuracy, CA - class accuracy, IoU - mean intersection over union). The aforementioned metrics are reported both for localisation via label histogram matching and more for computationally expensive matching using image alignment. Figure (b) shows the quantitative evaluation on the CamVid-360 dataset. It evaluates the proposed localisation method under three different error tolerance settings for correct matches ($\pm 1, \pm 5, \pm 15$ frames). See Figure 4(a) for qualitative results.

We also reported global accuracy, class accuracy of per-pixel segmentation as well as mean intersection over union score (IoU) for the SceneCity and Stanford-2D-3D-S datasets where the ground truth segmentation is readily available for the query images. We compared our semantic localisation method with classic localisation techniques based on hand-designed SIFT [29] features, learnt LIFT [63] features and NetVlad [3] image embedding. For SIFT and LIFT based localisation we defined matching confidence by counting the number of well matching (ratio of distances between nearest neighbours higher than 0.8) feature descriptors. For the NetVlad [3] method we used the Euclidean distance between image embedding vectors as a confidence score directly.

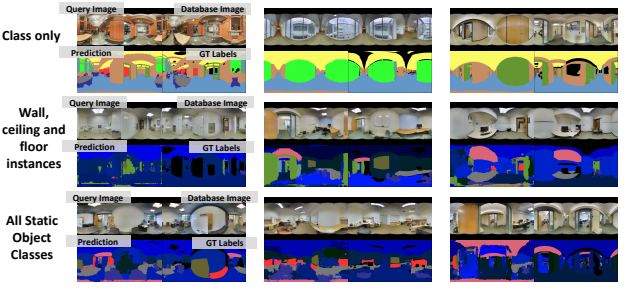
5 Results

In this section we discuss the results of the previously described experiments.

SceneCity dataset. Figures 3 and 5(a) provide qualitative and quantitative results on artificially generated cities. Our proposed localisation method attains high localization and segmentation accuracies. As expected, localisation accuracies for our method using label histogram matching are greatly improved on all three city map setups when building instances are used, the smallest increase being 18% in the case of the large city map. Similarly, using label image alignment for location refinement provides a significant improvement both when object class labels and building instances are employed, resulting in a 20% and 14% respective improvement in the small city map with missing buildings. More surprisingly, label histogram localisation using building instance labels performs at the same accuracy or better in all cities except in the case of the small city with missing buildings when compared to classical localisation approaches (LIFT [63], SIFT [29], NetVlad [3]). Semantic localisation using label image alignment and building instances outperforms all other techniques, including the difficult scenario of missing buildings, as demonstrated in Figure 3. Also note that in the case of missing buildings, the relatively good performance of NetVlad [3] localisation contrasts with previous claims [37]. While more investigation is needed, our current hypothesis is that image embedding techniques are more sensitive to increases in map size (only 58% MA) than to local changes in the map. We would however expect the changes in the map to have a larger impact as the size of the environment increases.

CamVid-360 dataset. Figures 4 and 5(b) provide quantitative and qualitative results for the CamVid-360 dataset. As in the artificial data experiments, the same trends of significant increases in accuracy as a result of (i) using instance labels instead of class labels and (ii)

(a)



(b)

Localisation Method	Match Function	Lab.	MA	GA	CA	IoU
Class only	Label Hist.	12	17	90	78	60
	Image Alig.		20			
Wall, ceiling, floor instances	Label Hist.	2184	76	65	44	18
	Image Alig.		63			
All static instances	Label Hist.	**3138**	77	61	38	15
	Image Alig.		74			
SIFT	# of Dsc.	N/A	55	N/A		
LIFT	# of Dsc.		47			
NetVLAD	Euc. Dist.		66			

Figure 6: This figure illustrates qualitative (a) and quantitative (b) results on the Stanford-2D-3D-S indoor dataset, respectively. While instance segmentation accuracy (also reflected quantitatively) is relatively lower than in the case of the artificial data experiments or the CamVid-360 dataset, the matching accuracy outperforms classical localisation techniques.

employing label image alignment over class histogram matching can be observed. While localisation accuracy is rather low in the case of the low error tolerance settings, i.e. ± 1 and ± 5 frames, it is within 1% of classical localisation algorithms for the least conservative setting (± 15), which still corresponds to localising within the hand-labelled frames. A worse relative performance compared to artificial city experiments can be primarily explained by (i) not accounting for translation in image alignment and (ii) by insuficient lighting augmentation. The former could be addressed by utilizing semantic 3D point clouds of the database images. The latter can be addressed by GAN-enabled data augmentation [58, 62]. Finally, Figure 4(b) confirms that confidence scores of both instance-based semantic localisation and classical techniques eliminate the majority of incorrect matches at an acceptance rate lower than 95-98%.

Stanford 2D-3D-S dataset. Qualitative and quantitative results for indoor experiments can be found in Figure 6. Similarly to the results for autonomous driving scenerios, an increase in the number of instances used leads to an increase in matching accuracy. However, differently than in previous experiments, simple class histogram matching outperformed label image alignment matching. This is expected, as relatively large translation is present between the database and query images. Nevertheless our technique significantly outperformed SIFT [29], LIFT [63] and NetVlad [3] which, especially in the case of key-point matching techniques, suffer in the case of lacking surface texture. Also note the high tolerance of our proposed localisation method to relatively low segmentation accuracies when compared to previous experiments. This is due to the low likelihood of miss-predicted object labels having geographically close locations.

6 Conclusions

In this work, we proposed a novel approach to semantic localisation consisting of three key steps: (i) building a database of panoramic RGB images and corresponding *globally unique object instance* masks for the desired environment, (ii) training a semantic segmentation network on the collected data and (iii) using a predicted global unique object instance mask image for semantic matching within the database. This method not only provided an ability to localise within a desired database, but also to segment and recognise objects from the surrounding environment. This is necessary for active vision applications such as autonomous driving, augmented reality and robotics. Quantitative and qualitative evaluation on indoor and autonomous driving scenerios demonstrated promising results. Our method outperformed classical localisation techniques [3, 29, 63] in two out of three datasets.

References

[1] Free and open source 3D creation suite. https://www.blender.org/. Accessed: 2018-04-29.

[2] SceneCity - Build Cities in Blender. https://www.cgchan.com/store/scenecity. Accessed: 2018-04-29.

[3] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] A. Armagan, M. Hirzer, P. M. Roth, and V. Lepetit. Learning to align semantic segmentation and 2.5D maps for geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] A. Armagan, M. Hirzer, P. M. Roth, and V. Lepetit. Accurate camera registration in urban environments using high-level feature matching. In *Proceedings of the British Machine Vision Conference (BMVC)*, September 2017.

[6] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*, February 2017.

[7] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016.

[8] B. D. Brabandere, D. Neven, and L.V. Gool. Semantic instance segmentation with a discriminative loss function. In *Deep Learning for Robotic Vision workshop in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[10] I. Budvytis, P. Sauer, T. Roddick, K. Breen, and R. Cipolla. Large scale labelled video data augmentation for semantic segmentation in driving scenarios. In *5th Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving in IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[11] A. X. Chang, A. Dai, T. A. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: learning from RGB-D data in indoor environments. In *International Conference on 3D Vision (3DV)*, October 2018.

[12] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011.

[13] H. Chu, S. Wang, R. Urtasun, and S. Fidler. Housecraft: Building houses from rental ads and street views. In *European Conference on Computer Vision (ECCV)*, October 2016.

[14] A. Cohen, J. L. Schönberger, P. Speciale, T. Sattler, J. M. Frahm, and M. Pollefeys. Indoor-outdoor 3D reconstruction alignment. In *European Conference on Computer Vision (ECCV)*, October 2016.

[15] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, November 2017.

[17] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[18] X. Gao and T. Zhang. Robust RGB-D simultaneous localization and mapping using planar point features. *Robotics and Autonomous Systems*, 72:1 – 14, 2015. ISSN 0921-8890.

[19] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[20] L. Goode and T. Warren. This is what microsoft hololens is really like. 2015. URL https://www.theverge.com/2016/4/1/11334488/microsofthololens-video-augmented-reality-ar-headset-hands-on.

[21] P. Gronát, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, October 2017.

[23] M. Hirzer, C. Arth, P. M. Roth, and V. Lepetit. Efficient 3D tracking in urban environments with semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, September 2017.

[24] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User interface Software and Technology*, October 2011.

[25] D. H. Juárez, L. Schneider, A. Espinosa, D. Vázquez, A. M. López, U. Franke, M. Pollefeys, and J. C. Moure. Slanted stixels: Representing San Francisco's steepest streets. In *Proceedings of the British Machine Vision Conference (BMVC)*, September 2017.

[26] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, December 2015.

[27] H. J. Kim, E. Dunn, and J. M. Frahm. Learned contextual feature reweighting for image geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[28] A. Kirillov, K. He, R. B. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. *ArXiv*, 2018.

[29] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, November 2004.

[30] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2016.

[31] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. SceneNet RGB-D: can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the International Conference on Computer Vision (ICCV)*, October 2017.

[32] O. Mendez, S. Hadfield, N. Pugeault, and R. Bowden. Sedar-semantic detection and ranging: Humans can localise without lidar, can robots? In *IEEE International Conference on Robotics and Automation*, May 2018.

[33] D. Mishkin, N. Sergievskiy, and J. Matas. Systematic evaluation of cnn advances on the ImageNet. *Computer Vision and Image Understanding*, 161(C):11–19, August 2017.

[34] M. Mueller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem. UE4Sim: a photo-realistic simulator for computer vision applications. *International Journal of Computer Vision (ICJV)*, 2018.

[35] T. Naseer, G. Oliveira, T. Brox, and W. Burgard. Semantics-aware visual localization under challenging perceptual conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, May 2017.

[36] G. Neuhold, T. Ollmann, S. R. Buló, and P. Kontschieder. The Mapillary Vistas dataset for semantic understanding of street scenes. In *IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[37] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *IEEE International Conference on Computer Vision, (ICCV)*, October 2017.

[38] T. Pylvänäinen, K. Roimela, R. Vedantham, R. Wang, and R. Grzeszczuk. Automatic alignment and multi-view segmentation of street view data using 3d shape priors. In *Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, December 2010.

[39] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim, Y. Wang, and A. Yuille. UnrealCV: virtual worlds for computer vision. In *Proceedings of ACM on Multimedia Conference*, October 2017.

[40] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision (ECCV)*, October 2016.

[41] S. R. Richter, Z. Hayder, and V. Koltun. Playing for benchmarks. In *International Conference on Computer Vision (ICCV)*, October 2017.

[42] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 12 2015.

[44] S. Satkin, J. Lin, and M. Hebert. Data-driven scene understanding from 3D models. In *Proceedings of British Machine Vision Conference (BMVC)*, September 2012.

[45] Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017.

[46] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2003.

[47] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler. Semantic visual localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[48] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: a unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[49] M. Schwarz, C. Lenz, G. M. Garcia, S. Koo, A. S. Periyasamy, M. Schreiber, and S. Behnke. Fast object learning and dual-arm coordination for cluttered stowing, picking, and packing. In *IEEE International Conference on Robotics and Automation (ICRA)*, May 2018.

[50] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke. Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, 37(4-5):437–451, 2018.

[51] S. Shah, D. Dey, C. Lovett, and A. Kapoor. AirSim: high-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017.

[52] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, Jan 2009.

[53] C. N. Silla and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72, Jan 2011.

[54] A. Torii, M. Havlena, and T. Pajdla. From google street view to 3d city models. In *Workshop in International Conference on Computer Vision (ICCV)*, September 2009.

[55] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[56] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual place recognition with repetitive structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11): 2346–2359, Nov 2015.

[57] S. Wang, S. Fidler, and R. Urtasun. Lost shopping! Monocular localization in large indoor spaces. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[58] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[59] T. Weyand, I. Kostrikov, and J. Philbin. PlaNet - photo geolocation with convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, October 2016.

[60] R. W. Wolcott and R. M. Eustice. Visual localization within LIDAR maps for automated urban driving. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2014.

[61] Z. Wu, C. Shen, and A. V. D. Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *CoRR*, abs/1611.10080, 2016.

[62] M. Wulfmeier, A. Bewley, and I. Posner. Addressing appearance change in outdoor robotics with adversarial domain adaptation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2017.

[63] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned Invariant Feature Transform. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016.

[64] Q. Yu, C. Szegedy, M. C. Stumpe, L. Yatziv, V. D. Shet, J. Ibarz, and C. Arnoud. Large scale business discovery from street level imagery. *ArXiv*, abs/1512.05430, 2015.

[65] Z. Zhang, H. Rebecq, S. Forster, and D. Scaramuzza. Benefit of large field-of-view cameras for visual odometry. In *IEEE International Conference on Robotics and Automation, (ICRA)*, May 2016.