PlaceNet: A Multi-Scale Semantic-Aware Model for Visual Loop Closure Detection

Hussein Osman[a,b], Nevin Darwish[a], AbdElMoniem Bayoumi[a,*]

[a] *Department of Computer Engineering, Faculty of Engineering, Cairo University, Gamaa Street, 12613 Giza, Egypt*
[b] *École Polytechnique Fédérale de Lausanne (EPFL), INN 134 (Bâtiment INN), Station 14, CH-1015 Lausanne, Switzerland*

Abstract

Loop closure detection helps simultaneous localization and mapping systems reduce map and state uncertainty via recognizing previously visited places along the path of a mobile robot. However, visual loop closure detection is susceptible to scenes with dynamic objects and changes in illumination, background, and weather conditions. This paper introduces PlaceNet, a novel plug-and-play model for visual loop closure detection. PlaceNet is a multi-scale deep autoencoder network augmented with a semantic fusion layer for scene understanding. The main idea of PlaceNet is to learn where not to look in a dynamic scene full of moving objects, i.e., avoid being distracted by dynamic objects to focus on the scene landmarks instead. We train PlaceNet to identify dynamic objects in scenes via learning a grayscale semantic map indicating the position of static and moving objects in the image. PlaceNet generates semantic-aware deep features that are robust to dynamic environments and scale invariant. We evaluated our method on different challenging indoor and outdoor benchmarks. To conclude, PlaceNet demonstrated competitive results compared to the state-of-the-art methods over various datasets used in our experiments.

Keywords:
Visual loop closure detection, deep learning for visual perception, visual SLAM, localization

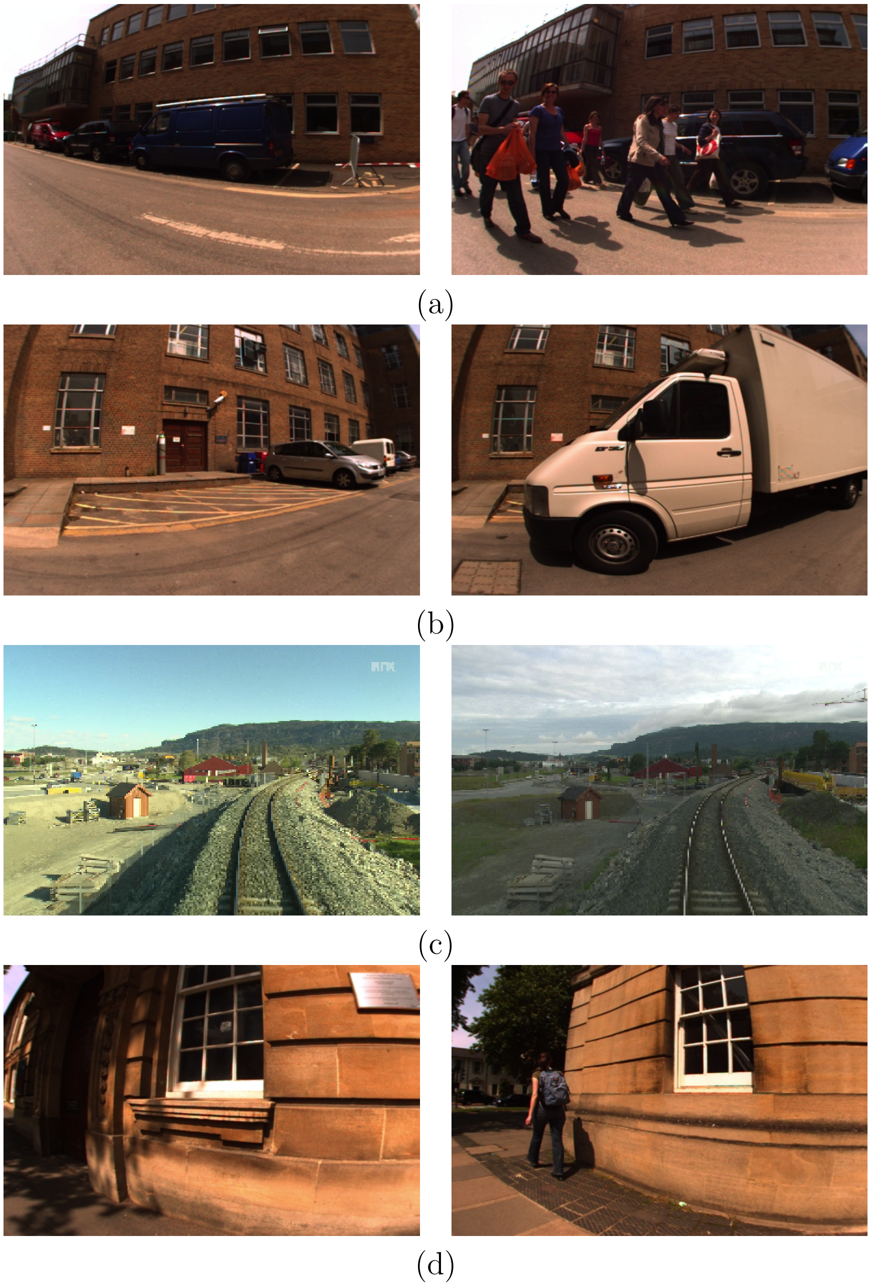
[*] Corresponding author
Email address: abayoumi@cu.edu.eg (AbdElMoniem Bayoumi)

1. Introduction

Visual SLAM has been receiving increasing attention in computer vision and robotics communities in the past few years due to its wide range of applications, including autonomous navigation, mine exploration, and reef monitoring. Visual SLAM estimates visual sensor motion while simultaneously constructing an unknown environment's map structure [1]. However, motion uncertainties lead to an accumulated error in calculating the camera pose and the map. Therefore, visual loop closure detection can help mitigate that accumulated error, increasing the accuracy and efficiency of Visual SLAM models.

Visual loop closure detection minimizes uncertainties of location and map estimates by detecting previously visited scenes based on their appearance in the captured images from the visual sensor [2]. In other words, Visual SLAM relies on such loop closures to correct its estimated camera pose and map, leading to mitigating the motion noise effect. However, current state-of-the-art visual loop closure detection modules face several difficulties in dealing with visual detection challenges. For instance, as shown in Fig. 1, two images of the same scene may look completely different due to dynamic objects causing partial occlusion of the scene, illumination variation, viewpoint changes, and seasonal variations leading to severe background changes. Additionally, indoor scene understanding suffers from perceptual aliasing due to similar and repetitive place structures such as hallways and corridors.

This paper introduces a novel plug-and-play model for visual loop closure detection entitled "PlaceNet." PlaceNet is a multi-scale architecture for deep convolutional auto-encoder networks, which yields competitive results compared to the state-of-the-art approaches on standard benchmarks. The main idea of PlaceNet is to learn where not to look at in a dynamic scene full of moving objects. Therefore, we augment the encoder network of PlaceNet with a semantic fusion layer to generate semantic-aware feature representations in an unsupervised scheme. As such, the reconstruction of these semantic maps via the decoder network urges PlaceNet to yield feature maps that are semantic-aware of the scene's objects. Simultaneously, to avoid being distracted by dynamic objects, we train PlaceNet in a supervised way to output a grayscale semantic channel that highlights such dynamic objects, e.g., vehicles, pedestrians, and cyclists. After that, we apply principal component analysis to eliminate the redundancy in extracted features caused by the appearance of these dynamic objects in many consecutive frames.



(a)

(b)

(c)

(d)

Figure 1: Some examples of correctly detected loops by our workflow (a) in a dynamic environment (b) with partial occlusion (c) with severe illumination change (d) an example of perceptual aliasing case identified by our method successfully as a non-loop.

Furthermore, PlaceNet generates feature representations that are also scale-invariant besides being robust to dynamic environments on various scenarios. Our novel multi-scale architecture yields scale-invariant feature representations by introducing input frames at different scales. Although the literature on computer vision has heavily discussed the fusion of feature maps at different scales, they generated such features from only a single input. On the other hand, our approach simultaneously fuses feature maps: (i) directly from multiple inputs and (ii) indirectly from previous convolutional layers in one global end-to-end network that shares parameters between the different scales.

We evaluated our method on different challenging benchmarks for loop closure detection. We experimented with outdoor and indoor environments under challenging conditions that include perceptual aliasing, partial occlusion due to dynamic objects, changes in scene background, and variations caused by changes in weather conditions, and illumination. We show the robustness of our method to combat these variations in our experimental results. We also performed an ablation study to show the effect of the multi-scaling and semantic fusion layers.

To conclude, the contribution of this paper is as follows:

1. Introducing PlaceNet as a novel multi-scale semantic-aware architecture for visual loop closure detection that yields competitive results and robust performance in various scenarios compared to the state-of-the-art approaches.
2. Performing loop closures in a dynamic scene full of moving objects following a ”where not to look at” strategy.
3. Generating scale-invariant feature representations by introducing input frames at different scales and performing a simultaneous fusion of feature maps directly from multiple inputs and indirectly from previous convolutional layers in one global end-to-end network that shares parameters between the different scales.

The paper outline is as follows: Section II discusses previous work related to loop closure detection in literature. Next, Section III proposes PlaceNet as a multi-scale semantic aware deep architecture for loop closure detection. After that, Section VI presents experimental results and provides ablation studies for loop closure detection using PlaceNet compared to the state-of-the-art approaches. Finally, Section V concludes this paper.

2. Related Work

Visual loop closure detection has gained an increasing attention in both robotics and computer vision fields, respectively. The research scope developed from recognizing static scenes with few changes [3, 4] to realistic highly dynamic scenes [5, 6]. Furthermore, several approaches were introduced to handle challenging conditions such as scene structure repetition [7, 8] and major changes in appearance due to viewpoint, illumination, shadows, day-night change or season change [9, 10, 11, 12, 13, 14, 15, 16]. Moreover, many solutions were introduced that rely on matching image sequences [5, 6], improving distinctiveness of features [17], using graph representations [18], and exploiting geo-tagged images [19, 8, 20]. Additionally, visual loop closure detection has benefited from the improvements achieved by image retrieval systems using handcrafted features [21, 22, 23, 24, 25, 26, 27, 28, 29], local convolutional features [30, 31, 32, 33] and global convolutional representations [34, 35, 36, 37, 38].

Global handcrafted descriptors, such as Histogram of Oriented Gradients (HOG) [6], Gist [39], and Gist-BRIEF [40], have been used for image representation in loop closure detection techniques. For example, SeqSLAM [5] relies on image difference vectors as global descriptors to find the best match of an input image within navigation sequences for the purpose of localization. Other variants of SeqSLAM such as Fast-SeqSLAM [6] reduces the time complexity of SeqSLAM without trading off accuracy via computing a histogram of oriented gradients (HOG) as a global descriptor and applying approximate nearest neighbor matching. However, global description-based methods are less robust to occlusion and illumination effect, resulting in a lower discriminative power and more perceptual aliasing effect.

Scale and rotation invariant local handcrafted features represented by keypoints and their descriptor vectors, such as SIFT [41], SURF[42], ORB [43], and FAST [44] with BRIEF [45], have been widely employed in many successful visual place recognition methods such as FAB-MAP [3] which is a probabilistic appearance-based method for visual place recognition based on the widely known Bag-of-Words (BoW) algorithm. In FAB-MAP, the BoW algorithm is trained on a large-scale image dataset to extract and cluster local features e.g. SIFT or SURF, and build a large visual vocabulary codebook. Consequently, each image is represented by a vector of visual words which is used for similarity measurement between different frames. Similar to FAB-MAP, another successful place recognition method is DBoW [46] which builds a vocabulary tree and uses FAST corners and BRIEF binary descriptors instead of SIFT and SURF leading to real time performance. This method has been successful in many real life applications and has been applied in ORB-SLAM system [47] resulting in a significant improvement in both accuracy and efficiency. Furthermore, Bampis *et al.* [48] extends DBoW [46] to describe image sequences, instead of single images, using visual-word-vectors. Moreover, Angeli *et al.* [49] presented an incremental online BoW approach that relies on Bayesian filtering to improve the generalization ability of offline BoW algorithms to new environments other than those used for training. Additionally, Garcia-Fidalgo and Ortiz [50] tackled the generalization problem as well using an incremental Bag-of-Binary-Words. Other related work in this area is the work of Tsintotas *et al.* [51] which operates online in real time via assigning local descriptors to visual words generated from earlier frames of a given sequence refraining from any pre-training procedure or vocabulary construction, then a probabilistic nearest neighbor search method detects loops. Similarly, the work of Labb and Michaud [52] is suitable for real-time large-scale and long-term operation with less memory requirements and less time complexity. However, loop closure detection algorithms based on BoWs still suffer from several limitations leading to poor performance in challenging environments, such as dynamic environments with varying conditions of lighting, shadows, and seasonal changes.

Convolutional neural networks (CNNs) [35] lead to substantial improvements when used in loop closure detection tasks [20]. Training deep CNN models on information-rich datasets results in high-level abstractions of input images, which correspond to complex features that outperform manually handcrafted features. Therefore, image representations of CNNs, such as AlexNet [53], OverFeat [54], VGG [55], Inception [56], ResNet [57], and Inception-ResNet [58], result in better scene understanding and better scene recognition.

In general, extracting features from CNNs can be performed in several ways: (1) the whole image is used as the input of the network, and the activations generated at one of its last hidden layers are extracted as the image's descriptor [59, 20] (2) specific regions of the image are input to the network, and the respective activations are aggregated to generate the image's descriptor [60, 61, 62] (3) the whole image is input to the network, and the activations of specific convolutional layers that detect distinct patterns are extracted, thus identifying the most prominent regions [63, 64] (4) the whole image is input to the network to predict its global and local descriptors simultaneously [65, 66]. These convolutional features are faster, more accurate, and less sensitive to illumination changes compared to handcrafted features [10, 67, 68, 69, 70, 71, 72, 73]. Furthermore, Zhang *et al.* [74] combine convolutional features and temporal information of sequence images in a graph-based visual recognition and apply a diffusion process leading to accuracy and time improvements. As opposed to the holistic approaches mentioned above, Li *et al.* [75] divide images into smaller patches and construct for every image pair an adaptive weighted similarity matrix between convolutional descriptors representing each patch. Similarly, Sünderhauf *et al.* [32] proposed a robust method for place recognition based on pre-trained convolutional features extracted from landmark regions in the image rather than the whole image, which renders their model invariant to viewpoint and appearance variation. Then, Chen *et al.* [76] improved the choice of these landmark and selected regions.

Instead of relying on supervised convolutional models, several approaches [77], [78], [79] have proposed learning image features using auto-encoders in an unsupervised way. For example, Gao and Zhang [77] divided raw input images into equally-sized vectorized patches and fed them to a stacked denoising auto-encoder to learn compressed representations of these raw input images. They used these learned features as descriptors to construct a similarity matrix between different frames. Merrill and Huang [78] stepped a further step and trained a denoising convolutional auto-encoder for loop closure (CALC) on HOG descriptors instead of raw images to learn more robust features to extreme variations in appearance. This method achieves outstanding results in loop detection accuracy and extraction speed, however it performs neither temporal consistency checks nor geometric checks as a post-processing step to filter false positives.

Other approaches tended to improve features extraction via supporting input images with extra information. For example, some approaches employed a multi-scale feature embedding method to generate condition- and viewpoint-invariant features [80, 81]. Additionally, the use of semantics [82, 83, 84] has recently received wide attention for place recognition tasks since similar images will have comparable semantic responses. Garg *et al.* [85] presented a local semantic descriptor using convolutional feature maps generated from a dense semantic segmentation network. They combined semantic and appearance-based global descriptors for image-pair matching. They check against the frequency map of high activation regions in higher-order convolutional layers of the network, since these regions capture visual semantics. Similarly, Camara and Přeučil [86] use pre-trained VGG-16 [55] activations of different layers as image features. These convolutional features that encode semantic and spatial information are image matching. Furthermore, Wang *et al.* [87] focused on compressing redundant information (e.g. moving objects and background change) in convolutional holistic representations to achieve more robust performance in measuring image similarity in highly dynamic environments.

Recent approaches use end-to-end Convolutional Siamese Networks (CSN) to combine feature extraction and similarity measurement steps. Liu *et al.* [88] apply a hierarchical weighted distance layer to fuse features from multiple scales of CSNs in different layers, while Garg *et al.* [11] use feature pyramid Siamese networks on RGB-D images via providing information about scale, structure and depth of the scene in order to ease capturing of object representations from different scales to improve the overall performance. Alternatively, Appalaraju and Chaoji [89] proposed a deep multi-scale CSN where each branch of the Siamese network operates on a scale of the original input image. The final layer fuses the outputs of these branches to generate image embeddings for similarity measurement.

The state-of-the-art approaches of visual loop closure detection includes LoopNet [90], NetVLAD [20], CALC2.0 [79], GeM [91], AP-GeM [92], DenseVLAD [93], Wang *et al.* [87], FILD++ [94], LiPo-LCD++ [95], Zhang et *al.* [96, 97], Xu et *al.* [98], Gehrig et *al.* [99], Tsintotas et *al.* [100, 101, 102] and Papapetros et *al.* [103]. In our previous work, LoopNet [90], we detect similarities between scenes using a multi-scale attention-based Siamese network. The attention mechanism helps in focusing on key landmarks in scenes. As for NetVLAD [20], it extracts convolutional features from an off-the-shelf CNN and pools these features into a trainable VLAD pooling layer. Unlike VLAD which relies on hard assignment of descriptors to offline-learned clusters, NetVLAD applies soft assignment of VLAD descriptors to these clusters rendering NetVLAD as a trainable end-to-end architecture for visual place recognition using a triplet ranking loss function. Furthermore, the work of Merrill and Huang [79] (CALC2.0) constructed a robust holistic-image descriptor describing both the visual appearance and semantic layout of an image by training a network comprising a semantic segmentator, variational auto-encoder and a Siamese triplet embedding network to extract such descriptors. Moreover, GeM [91] introduces generalized-mean (GeM) pooling layers that leads to better image retrieval compared to conventional global max and average pooling layers. Then, AP-GeM [92] optimizes mean average precision (mAP) directly instead of pair-wise losses, besides adopting GeM pooling layers. As for DenseVLAD [93], it uses dense VLAD descriptors to match scenes regardless of possible variations in scene appearance. Furthermore, Wang *et al.* [87] perform redundant information compression via post-processing raw holistic convolutional representations, where the compression ratio corresponds to the level of scene variations. However, they focus only on outdoor environments. FILD++ [94] extracts global and local convolutional features using different scales. Then, it constructs an incremental database using the global features to recommend potential loop closures to be evaluated using the local features. However, FILD++ faces some difficulties dealing with significant background changes due to weather conditions and perceptual aliasing, as we show later in our experimental section. Additionally, LiPo-LCD++ [95] learns lines and points for low-textured scenes; however, it suffers from scenes with perceptual aliasing, as mentioned in [95]. Furthermore, Zhang et *al.* [96] learn the motion field of local neighborhood structures based on extracted convolution features to detect loop closures. Then, they extended their approach in [97] to operate online without the need to construct a vocabulary database relying on an attention variant of NetVLAD [20]. Also, Xu et *al.* [98] rely on NetVLAD [20] and combine it with features generated from a second-order attention module to recommend matching candidates. However, their approach is sensitive to view-point change. On the other hand, Gehrig et *al.* [99] rely on probabilistic voting, while Tsintotas et *al.* [100, 101, 102] and Papapetros et *al.* [103] consider bag-of-words-based models.

These state-of-the-art methods still lack generality to perform efficiently in any outdoor or indoor environment that may feature highly repetitive structures, substantial scene variations, or changes in illumination. As opposed to these methods, we address these challenges by proposing a general and efficient solution for visual loop closure detection that can operate in any environment. Our method learns feature representations that are semantic aware of dynamic objects to avoid being distracted by them to focus on the scene landmarks instead. All in all, PlaceNet generates scale invariant feature representations that are also semantic-aware and robust to dynamic environments.

3. PlaceNet

Convolutional Neural Networks have shown their dominance in loop closure detection due to their powerful representation. As such, we propose PlaceNet as a multi-input multi-scale deep convolutional auto-encoder (CAE) network with semantic map fusion and weighted scale-wise loss function. We feed to PlaceNet multi-scale variants of input frames and their corresponding semantic information. We perform semantic segmentation as a pre-processing step using the pre-trained UPerNet network [104] on the original RGB input image to generate a 3-channel RGB pixel-level annotated semantic map. Then, we train PlaceNet in an unsupervised way to generate multi-scale outputs similar to the multi-scale inputs by learning a compact representation of the discriminative features. Additionally, we train it simultaneously in a supervised way to generate a grayscale semantic channel that highlights dynamic objects.

PlaceNet, similar to any convolutional autoencoder network, consists mainly of two sub-networks: (1) an encoder network that compresses the input into a bottleneck layer via convolutional and pooling layers, and (2) a decoder network that reconstructs the original input from the bottleneck layer via upsampling or deconvolutional layers. After the training phase, we extract a deep feature representation that captures the input image's most distinctive features from the bottleneck layer. Then, we rely on such compact deep representation to perform image similarity comparison. The rest of this section highlights the impact of multi-scale semantic information fusion and discusses the network architecture of PlaceNet. It also presents the weighted scale-wise loss function used in training our model and the similarity metrics used in detecting loop closures.

3.1. Multi-Scale Semantic Information Fusion

Scene parsing via semantic segmentation is crucial for scene understanding and enhances extracting powerful features that genuinely describe scenes for loop closure detection. Thus, we designed the auto-encoder network of PlaceNet to restore both input RGB images and their semantic information such that the extracted features of PlaceNet can describe the semantics of scene objects in addition to their structure, texture, and location. It is worth mentioning that there is no need for ground-truth semantic information to be provided. However, as a pre-processing step, we perform semantic segmentation using the pre-trained UPerNet network [104] on the original RGB input image to generate a 3-channel RGB pixel-level annotated semantic map.

Furthermore, since loop-closure detection is sensitive to dynamic scenes, it is essential to differentiate between dynamic objects and static objects that act as a scene background. Thus, we trained PlaceNet to output a grayscale semantic map categorizing objects into static objects (e.g., buildings, trees, and walls), weakly-dynamic objects (e.g., vehicles, buses, and trains) and dynamic objects (e.g., pedestrians, cyclists, and vendors) with pixel intensities 0, 0.5, and 1, respectively. As such, the feature representations learned by the network should be able to locate the regions in the input image containing dynamic objects, and then perceive these regions as redundant information that does not help in loop closure detection. The features representing dynamic objects are considered redundant since dynamic objects appear in several consecutive frames due to their dynamic nature. Accordingly, following the work of Wang *et al.* [87], we apply principal component analysis to eliminate redundancy in extracted features caused by the appearance of dynamic objects in several consecutive frames.

We fuse the RGB semantic map and the original RGB image, as shown in Fig. 2, in a 6-channel input volume and generate a 7-channel output volume. Such volume comprises the reconstructed RGB image and semantic map, learned in an unsupervised way, in addition to the grayscale dynamic map learned simultaneously in a supervised way. Therefore, PlaceNet can learn features that deeply understand scenes and focus on scenes' backgrounds.

Moreover, we introduce a novel multi-scale architecture that results in scale invariant feature representations by introducing input frames at different scales. As opposed to the literature in computer vision, which heavily discusses the fusion of feature maps at different scales generated only from a single input, we perform multi-scaling on the 6-channel input volume and feed the multi-scale variants, including the original input, to PlaceNet. Accordingly, the expected output of PlaceNet is also a 7-channel output volume per each scale.

3.2. Network Architecture

Inspired by the Unet architecture used by Ronneberger et al. [105] for medical image segmentation, we introduce PlaceNet as a CAE for extracting powerful features that are invariant to scale changes while incorporating semantic information about the scene objects and their dynamism. Fig. 3 shows that the network architecture consists of an encoder part and a decoder part with a bottleneck layer in between. Instead of having a single-

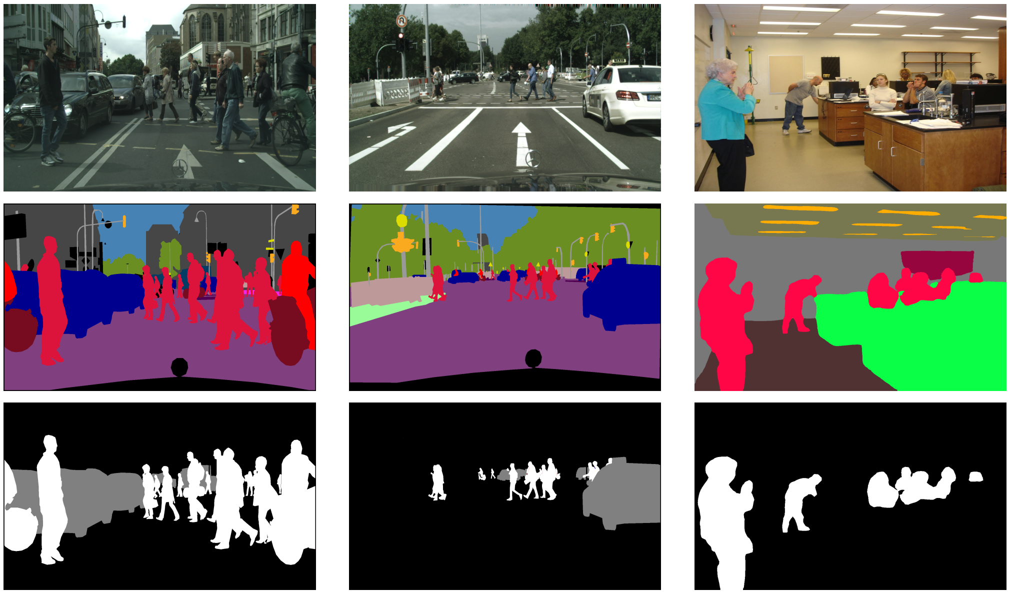


Figure 2: Examples of semantic fusion for PlaceNet with the original image, pixel-annotated semantic map, and grayscale dynamic map at the first, second, and third row, respectively. The left two columns represent two different images from CityScapes dataset captured outdoors with moving objects at different times of the day, where the right column represents an indoor image from ADE20K dataset.

input single-output network, PlaceNet is a three-input three-output network (or a three-channel-network). Each input represents one scale (octave) of the input image I and its semantic map I_s. The three network channels represent the input at full-scale, half-scale, and quarter-scale. The encoder network follows a typical architecture of a convolutional network with 7x7, 5x5, and 3x3 single convolutional layers applied to the full-scale, half-scale, and quarter-scale input volumes, respectively. We follow each convolutional layer implicitly by a batch normalization (BN) layer and a rectified linear unit (ReLU) activation layer. Next, we downsample the full-scale convolutional layer's output via a 2x2 max-pooling operation and add it to the half-scale first convolutional layer's output. Then, we pass the fused output to two 3x3 convolutional layers, further downsample and add it to the quarter-scale first convolutional layer's output. The three scales, combined in the generated feature map, are repeatedly applied to two 3x3 convolutional layers and 2x2 max-pooling for downsampling until we reach the bottleneck layer (the last two convolutional layers). Note that we double the number of feature

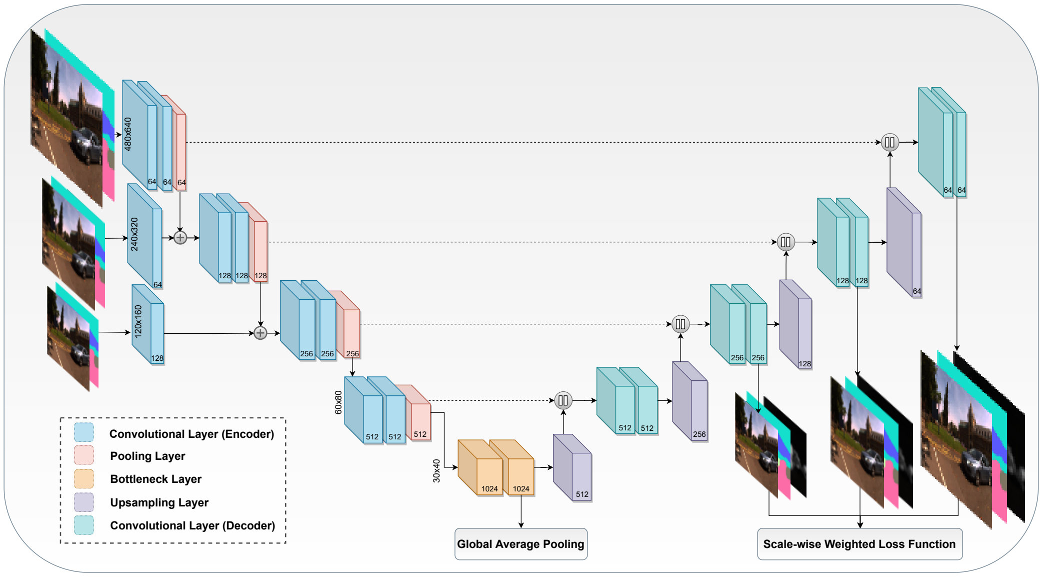


Figure 3: PlaceNet Multi-scale Architecture. The encoder network (left part of the network) is trained on three scales of the input image (full, half and quarter). The bottleneck layer (middle part of the network) is the feature extraction layer. The decoder network (right part of the network) consists of upsampling and convolutional layers with skip connections (dotted) to the encoder network to reconstruct the input image, the semantic map and the grayscale dynamic map.

maps at each scale down towards the bottleneck layer and that we reduce the dimensions of these maps to half.

On the other hand, every step along the reconstruction path consists of an upsampling layer which doubles the dimensions of feature maps and reduces the number of these maps to half, then a concatenation with the corresponding feature maps from the encoding path via a skip connection, followed by two 3x3 convolutional layers. The output layer of each of the three scales in the decoder network consists of one convolutional layer with a sigmoid activation function to reconstruct the output of the given scales.

The skip connections result in learning more powerful multi-scale features of the input image since they allow the network to learn directly from the input at the corresponding scale. Accordingly, combining two different ways of learning at each scale, i.e., directly via skip connections and in a cascaded way from the upper layers, allows the network to focus on both coarse and fine details of the image. Hence, the learned features become robust to scale variance.

3.3. A Weighted Scale-Wise Loss Function

The loss function must account for restoring different scales of input images and their corresponding semantic maps to ensure that the generated features are powerful enough to describe coarse and fine details of the image in addition to their semantic information, including objects' dynamism. We use an element-wise mean squared error loss function, as follows:

$$L = \frac{-1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2, \tag{1}$$

where $\hat{y}$ represents the predicted output value, y is the corresponding ground-truth value, and N is the total number of output volume elements. We apply this loss function to compute the reconstruction loss for both the output image, L_I, and its corresponding output semantic and dynamic maps, L_{I_S}, defined as:

$$L_I = \sum_{m}^{M} L_I^m, \tag{2}$$

and

$$L_{I_S} = \sum_{m}^{M} L_{I_S}^m, \tag{3}$$

where M is the set of scales considered in the network, while L_I^m and $L_{I_S}^m$ correspond to the reconstruction loss of a particular scale m for both an output image and its corresponding semantic and dynamic maps, respectively. After that, we aggregate both losses, as follows:

$$L_{I,I_S} = \beta_S L_I + (1 - \beta_S) L_{I_S}, \tag{4}$$

where β_S is weighting hyper-parameter, and we tune it to give more weight to the image reconstruction loss than the semantic map reconstruction loss. We favor image reconstruction since semantic maps, as input to PlaceNet, are not ground-truth maps but are rather generated from a pre-trained semantic segmentation network, i.e., UPerNet [104]. Therefore, the proximity of these generated semantic maps to the ground-truth pixel annotated semantic maps is heavily dependent on the performance of the semantic segmentation network.

3.4. Similarity Measurement

In order to measure the similarity between images to detect loop closures, we extract features from the bottleneck layer of PlaceNet and perform global average pooling (GAP) to reduce the dimensionality. Then, we concatenate the results in a 1-D feature vector for each image, perform a PCA whitening as a post-processing step to eliminate objects semantically labelled as dynamic. Finally, we can evaluate visual similarity using a cosine similarity measurement score between normalized feature vectors of any pairwise frames i and j, as follows:

$$S_v(i,j) = max(0, \frac{f_i}{||f_i||}.\frac{f_j}{||f_j||}), \tag{5}$$

where f_i and f_j are their feature vectors, respectively. We note that the PCA step is performed incrementally at each keyframe while the SLAM system is checking for loop closure. Thus, the descriptors of the captured frames up to the current step are taken into consideration when performing the PCA.

Although adjacent frames to the query frame look very similar, they do not represent an actual loop and lead to many false positives. Therefore, similar to the work presented by Zhang *et al.* [74], we add a temporal constraint on similarity score to account for adjacent frames. So, we use a temporal similarity measurement, defined as:

$$S_t(i,j) = exp(-\kappa_t * (i-j)^2), \tag{6}$$

where κ_t is a temporal similarity parameter. Thus, adjacent frames will have very high temporal similarity score and vice versa. This approach is better than having a sliding window over the query frame and searching everywhere outside this window which may not be the best option when the same sequence is collected in different speeds. Accordingly, we use this temporal similarity score to penalize the visual similarity score, as follows:

$$S(i,j) = max(0, \gamma_t * \frac{f_i}{||f_i||}.\frac{f_j}{||f_j||} - (1-\gamma_t) * S_t(i,j)), \tag{7}$$

where γ_t is a parameter used to weight the importance of visual similarity compared to temporal similarity. Thus, we can construct a similarity matrix S between all frames in a given sequence. In practice, after the model is fully trained and tested, we inspect the similarity score between two frames and detect a loop if it exceeds 0.75 indicating a positive match and return a negative match otherwise.

4. Experiments

We evaluated our approach on several publicly available benchmarking datasets. Furthermore, we validate our approach with quantitative results and qualitative analysis and compare our method with the state-of-the-art loop closure detection algorithms. It is worth mentioning that the testing datasets used in our experiments are non-overlapping and totally different from those used in training our models.

4.1. Training Phase

We trained PlaceNet on CityScapes dataset [106] and a subset of ADE20K dataset [107]. We chose to train our network on scene-centric sequences rather than object-centric images to obtain more scene-representative features. Cityscapes is a dataset for semantic urban scene understanding with 5,000 high-quality pixel-level annotated images collected from 50 cities across Germany in different seasons with 30 classes. Alternatively, ADE20K is a huge dataset with around 20K training images and up to 150 classes, and it represents a more challenging and diverse dataset that includes outdoor and indoor sequences. We unified the class labels and annotations across the two datasets.

Inspired by [108], we used a polynomial learning rate policy with learning rate at each iteration is multiplied by $(1 - \frac{iter}{max_{iter}})^{power}$ of the base learning rate. We set the initial learning rate and *power* to 0.01, and 0.9 respectively. We adopted mini-batch learning with mini-batch size of 64 for 2K epochs, using Adam optimizer with momentum parameters β_1 and β_2 set to 0.9 and 0.999 respectively and weight decay regularization parameter of 0.0001. We used the parameters β_S, κ_t, and γ_t with values 0.65, 0.3, and 0.002 respectively. We applied data augmentation on the training set in the form of random mirroring, random Gaussian noise, and random rotation between +10 and -10 degrees to reduce the effect of overfitting. Additionally, we specified a training/validation split of 10K/1K images, respectively. We used Keras [109] with TensorFlow backend for training our model (24 MB in total), and we performed our experiments on a core i7 5820K 3.3 GHz machine, with 32 GB RAM and GPU Nvidia RTX2070 with 8 GB memory.

4.2. Testing Datasets

In this section, we discuss the characteristics and challenges of each testing dataset. Furthermore, Table 1 demonstrates a summarized comparison between these test datasets.

Table 1: Datasets detailed description.

Dataset	No. of frames	Environment	Camera Position	Image Resolution	Background Change	Viewpoint Change	Dynamic Objects	Time Interval
City Center	2474	Outdoors, Urban	Lateral	640x480	Minor	Medium	Many	Short
New College	2146	Outdoors, Campus	Lateral	640x480	Minor	Medium	Few	Short
KITTI-00	4551	Outdoors, Urban	Frontal	1241x376	Minor	Medium	Many	Short
KITTI-05	2761	Outdoors, Urban	Frontal	1241x376	Minor	Medium	Many	Short
KITTI-06	1101	Outdoors, Urban	Frontal	1241x376	Minor	Medium	Many	Short
Nordland	2828	Outdoors, Railway	Frontal	1920x1080	Severe	None	None	Long
Gardens Point	400	Outdoors, Campus	Lateral	960x540	Minor	High	Many	Medium
TUM-SLAM	2585	Indoors, Office	Frontal, Handheld	640x480	Minor	High	None	Short

4.2.1. City Center and New College

Both datasets, published in [3], are widely used in Visual SLAM research and have been established as a benchmarking standard for loop closure detection algorithms. A robot collects these two datasets with two monocular cameras (left and right) mounted on it without any overlap. We perform our experiments on the left and right image sequences separately and combine their results with the ground-truth values readily available. Additionally, both sequences' images were captured during a relatively short time and thus had stable illumination conditions and very little background change. City Center (CC) dataset collected along with urban areas and roads near the city center features many dynamic objects such as cars, trucks, and pedestrians. Thus, City Center dataset faces partial occlusion and unstable shadow features in some scenes. The second dataset, New College (NC), collected around New College in Oxford, covers large regions with intense visual repetition and repeated structures, with little scene change or moving objects, including challenging identical repeating archways and a garden area surrounded by long uniform stone walls and bushes.

4.2.2. KITTI

The KITTI vision benchmark suite [110] is used mainly to evaluate visual odometry and SLAM systems. It consists of 22 sequences captured in urban outdoor environments. We evaluated our model using only three sequences, namely KITTI-00, KITTI-05, and KITTI-06, which contain several loop closures compared to the other sequences with very few loop closures. These three sequences feature a minor background change similar to New College and City Center datasets but face significant viewpoint changes caused by different trajectories. The main challenge in KITTI sequences is the significant amount of moving objects, mainly cars, trucks, and pedestrians, causing partial occlusions in loop scenes. We use the ground-truth values provided by Arroyo *et al.* [111] since KITTI sequences do not include ground-truth labels for loop closure detection.

4.2.3. Nordland

The Nordland dataset [112] is one of the longest sequences collected from the same viewpoint of a moving train on a railroad between two cities in northern Norway. The images were collected in four different seasons. Thus, this dataset corresponds to a loop that is traversed four times. It is considered one of the most challenging datasets because of the dramatic changes in the landscape between different seasons. Throughout the journey, most of the scenes are mainly natural scenery with rare occasions when the train passes by urban areas or stops at railway stations. Unlike the datasets mentioned above, there is no viewpoint change and barely any moving objects causing occlusions.

We evaluated our method on the Spring-Summer and the Spring-Winter sequences since they include significant background changes due to weather conditions such as snow, fog, cloud, and sunlight change. For the ground-truth values, the images are time-synchronized and matched one-to-one between the two sequences. We consider two scenes to represent the same place if they are separated temporally by less than ten frames relative to the speed of the train from which the sequence was captured.

4.2.4. Gardens Point

The Gardens Point (GP) dataset [10] is a two-day sequence taken at a university campus in Brisbane, using a forward-facing hand-held mobile camera. The two sequences correspond to a cycle that is traversed twice. However, one route is traversed on the left-hand side of the path while the other being on the right-hand side of the path, capturing both pose and condition change. Furthermore, the two sequences include many moving students on campus, which is a challenge in this dataset. We consider two images to represent the same place if they are separated temporally by five images or less relative to the speed of the camera capturing the sequence.

4.2.5. TUM-SLAM

Unlike previous datasets, which feature mainly outdoor scenes, TUM dataset [113] incorporates many indoor RGB-D sequences and is commonly used in visual SLAM research. We evaluated our method on the indoor office sequence ”freiburg3/long_office_household” which is the only sequence that contains loop closures. This experiment’s main purpose is to evaluate the system’s performance in indoor environments on image sequences captured with hand-held cameras.

4.3. Evaluation Metrics

We evaluate our model by comparing our workflow predictions against the ground truth values for each sequence. We use precision-recall (PR) curves to evaluate the performance of the proposed method. Accordingly, we count the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Recall is the proportion of correct loops retrieved from all actual loops $TP/(TP + FN)$) while precision is the ratio of correct loops to all detected loops by our method ($TP/(TP + FP)$). We construct the PR curve by computing the recall and precision values for a variety of thresholds above which the scene is detected as a loop.

The higher the area under the precision-recall curve indicates that a method achieves high precision, i.e., returns accurate results, and high recall, i.e., returns the majority of all positive results. Although there are many ways to interpret the precision-recall curve, we consider the maximum recall value (r) achieved at 100% precision to compare our workflow against other methods since it implies no false positives. A false positive, i.e., wrong loop closure, in mapping tasks can lead to inconsistent maps, and therefore avoiding false positives is essential to the robustness of the mapping algorithm. Furthermore, the (r) metric is more reliable than the area under the curve (AUC) score since some classifiers can have non-perfect precision for all recall values despite having a high AUC score.

4.4. Results and Evaluation

We compare, as shown in Table 2, PlaceNet with state-of-the-art CNN-based approaches for visual loop closure detection [20, 79, 91, 92, 94, 93, 90, 72, 95, 74, 96, 97, 98] beside classical methods including probabilistic models [3, 5, 99] and bag-of-words-based models [50, 102, 103, 51, 100, 101]. Our method clearly shows an improved performance in the case of KITTI-00, Gardens Points, and TUM-SLAM datasets. Furthermore, our method achieves

Table 2: Maximum recall (r) at 100% precision for different benchmark datasets

Method	**City Center**	**New College**	**KITTI 00**	**KITTI 05**	**KITTI 06**	**Nordland Spr-Win**	**Nordland Spr-Sum**	**Gardens Point**	**TUM-SLAM**	**AVG.**
PlaceNet - Ours	92.50	90.66	**98.50**	92.46	98.14	90.26	93.26	**100.0**	**96.32**	**93.83**
NetVLAD [20]	86.34	85.17	96.72	87.90	95.50	39.52	43.51	**100.0**	53.64	68.03
CALC2.0 [79]	84.47	81.32	97.25	82.24	97.54	**98.58**	**99.50**	47.00	85.23	82.68
GeM [91]	86.74	85.35	97.00	89.50	95.30	86.50	74.26	98.50	84.62	88.64
AP-GeM [92]	87.29	84.22	98.12	89.25	96.00	89.24	75.59	98.50	86.10	86.82
FILD++ [94]	90.01	82.37	94.92	95.42	98.16	81.00	80.50	95.67	90.15	89.80
DenseVLAD [93]	83.45	81.28	95.40	84.26	93.22	45.90	76.19	99.20	83.24	78.21
LoopNet [90]	89.15	84.62	$*$	$*$	$*$	82.95	88.66	**100.0**	92.45	n.a.
Wang et *al.* [72]	87.52	88.10	96.68	79.96	97.69	96.83	-	86.50	-	n.a.
LiPo-LCD++ [95]	**92.99**	-	98.08	93.68	**99.62**	-	-	-	-	n.a.
Zhang et *al.* [74]	63.19	48.79	95.37	71.01	-	-	-	-	-	n.a.
Zhang et *al.* [96]	84.75	-	94.29	91.81	99.26	-	-	-	-	n.a.
Zhang et *al.* [97]	-	89.05	94.29	91.57	-	-	-	-	-	n.a.
Xu et *al.* [98]	-	**91.02**	97.46	-	98.9	-	-	-	-	n.a.
FAB-MAP [3]	38.50	51.91	49.21	37.65	55.34	-	-	-	-	n.a.
Seq-SLAM[5]	51.91	49.39	67.04	41.37	64.68	-	-	-	-	n.a.
Gehrig et *al.* [99]	74.00	84.70	92.80	86.00	98.50	-	-	-	-	n.a.
iBOW-LCD[50]	88.25	73.10	76.50	53.00	95.53	85.23	87.23	95.00	91.27	82.79
BoTW-LCD[102]	36.00	87.00	97.70	94.00	98.10	86.20	83.00	96.50	93.43	85.57
Papapetros et *al.* [103]	-	85.8	83.4	$*$	-	-	-	-	-	n.a.
Tsintotas et *al.* [51]	-	87.97	93.18	**94.20**	-	-	-	-	-	n.a.
Tsintotas et *al.* [100]	-	-	93.5	90.0	-	-	-	-	-	n.a.
Tsintotas et *al.* [101]	-	-	97.5	-	-	-	-	-	-	n.a.

The (-) symbol means that the corresponding method did not experiment on the open dataset.
The ($*$) symbol means that the corresponding method was trained on that dataset.

a very comparable performance in City Center, New College, KITTI-05, and KITTI-06. As opposed to the state-of-the-art, our model demonstrates its ability to generalize and perform robustly in various scenarios, including dynamic objects (as in KITTI sequences), perceptual aliasing (as in New College), scene change (as in City Center), and indoor environments (as in TUM-SLAM). In this regard, it is worth mentioning that LiPo-LCD++ [95] suffers from scenes with perceptual aliasing, as mentioned in [95], which also applies to [51] based on its performance on New College dataset. Furthermore, CALC2.0 [79] suffers from view-point change, especially when combined with many dynamic obstacles, as in the Gardens Point dataset. Additionally, the sensitivity to view-point change also applies to the approach of Xu et *al.* [98], as they mentioned.

PlaceNet is outperformed by CALC2.0 in Nordland sequences due to the limited number of images containing natural scenery in our training set. Also, the grayscale semantic map for dynamic objects does not add much repre-

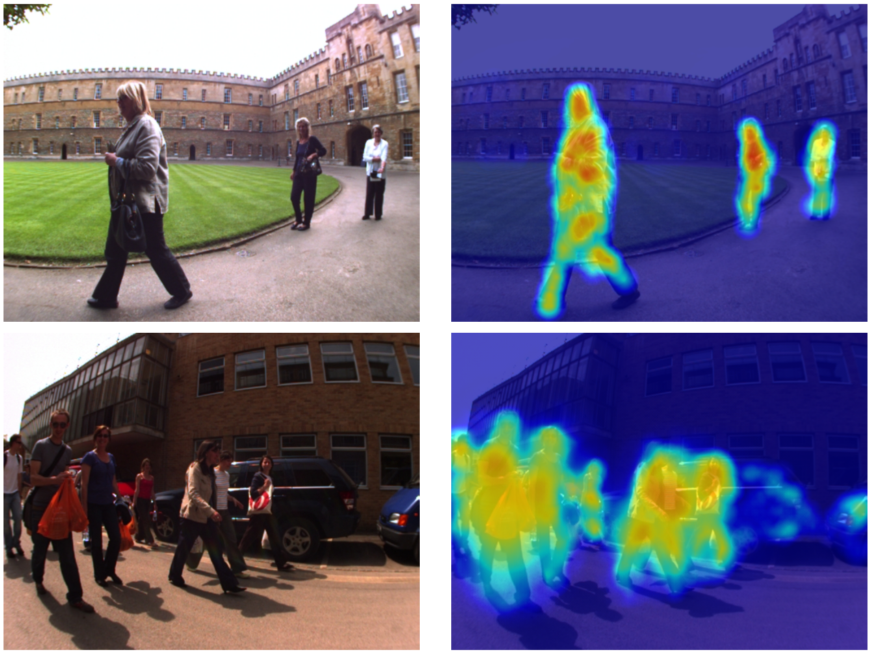


Figure 4: Examples of reconstructed grayscale dynamic maps by PlaceNet.

sentative power to this sequence since there are barely any moving objects on a train journey. Additionally, the scenes in the winter sequence are mostly covered in snow, concealing many details about object semantics and leading to less successful scene understanding with semantic fusion. However, PlaceNet is more robust and performs more consistently than CALC2.0 in various scenarios, as shown by the results on the rest of the datasets yielding an average improvement of 11.15% compared to CALC2.0.

We also show in Fig. 4 the reconstruction of the dynamic maps generated by PlaceNet in some dynamic environments. Fig. 4 demonstrates that PlaceNet learns a feature vector that embeds information about the semantics of moving objects in the scene as an encoded representation. Accordingly, the power of PlaceNet can be viewed in Fig. 5 showing loops detected by PlaceNet in dynamic scenes rich with moving objects.

Table 3 demonstrates the average computation time required to encode and match an image in a sequence. We demonstrate that PlaceNet is also computationally efficient. PlaceNet achieves a slightly improved average computation time to [20, 79, 93], and is very comparable to [90, 91, 92].

Additionally, We performed an ablation study using six variants of PlaceNet

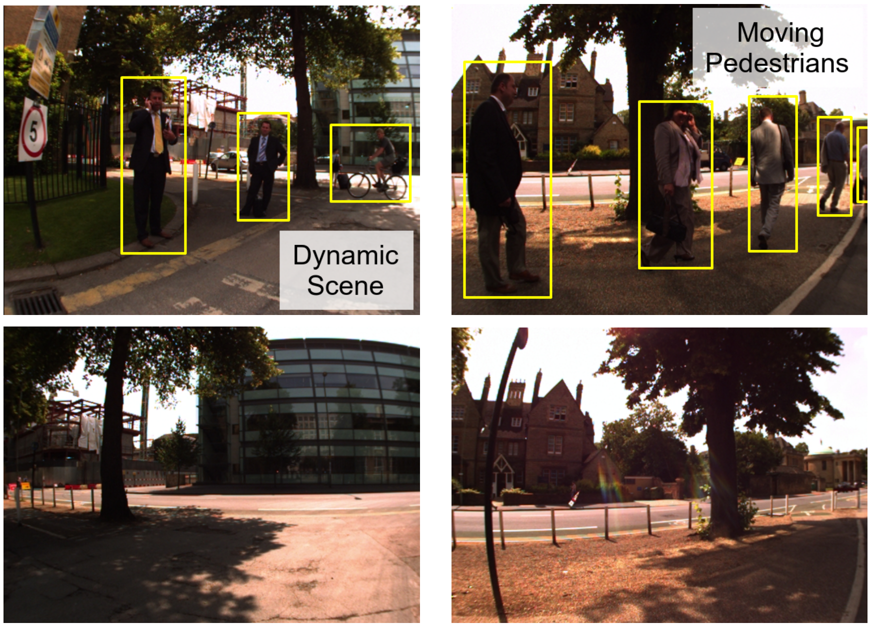


Figure 5: Examples of detected loops along the trajectory of the navigating vehicle in City Center Experiment where each column represents a loop closure instance.

Table 3: Comparison of running time, i.e., total of encoding and matching times, per image for different benchmark datasets (in msec).

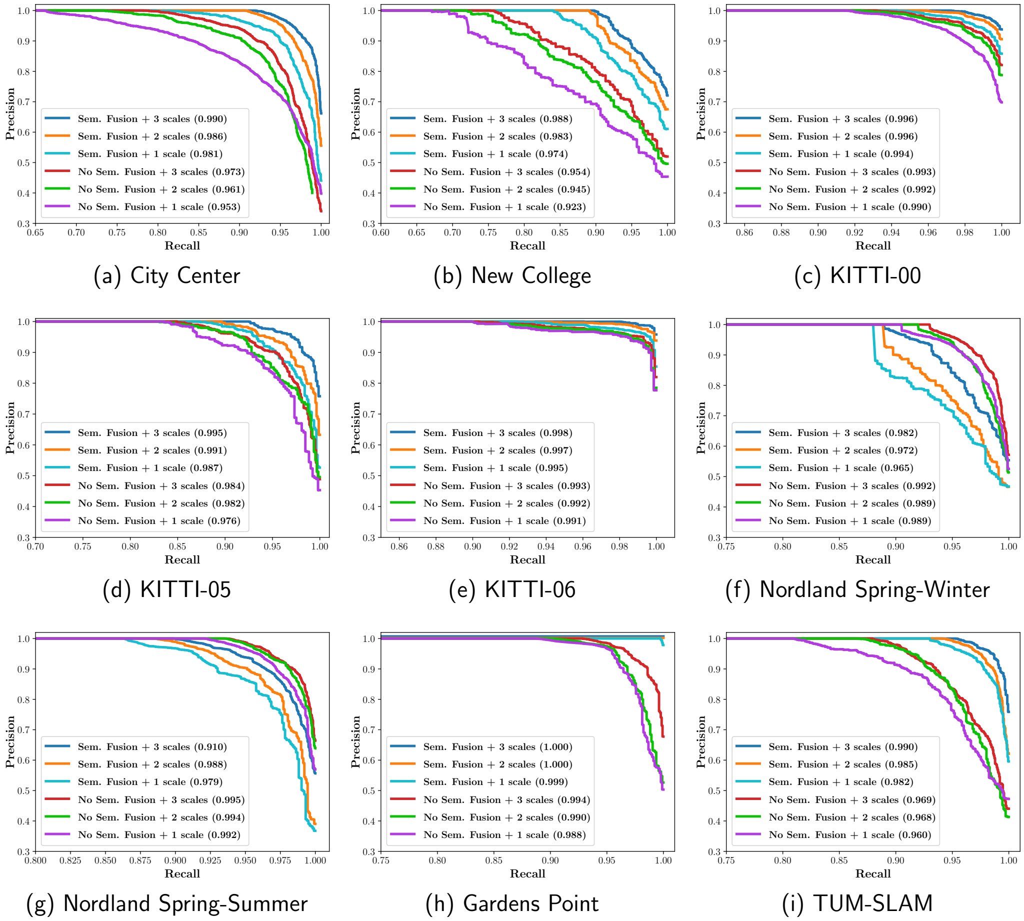
Method	City Center	New College	Gardens Point	Nordland Spr-Sum	TUM-SLAM	Average
PlaceNet - Ours	5.24	5.26	5.38	4.98	5.10	5.19
LoopNet [90]	4.85	4.70	4.68	4.71	4.80	4.58
NetVLAD [20]	8.96	8.85	8.7	8.45	8.79	8.75
CALC2.0 [79]	7.52	7.19	7.48	7.69	7.55	7.49
GeM [91]	3.29	3.25	3.88	3.74	3.69	3.57
AP-GeM [92]	3.85	3.79	3.81	3.88	3.74	3.81
DenseVLAD [93]	5.69	5.60	5.74	5.70	5.56	5.65

in order to investigate the effect of multi-scaling and semantic fusion on the performance of our method in loop closure detection. As shown in Table 4, we conducted experiments on PlaceNet with and without semantic fusion and with different combinations of input scales. We trained these models using the same settings as PlaceNet.

As shown in Fig. 6 and Table 4, PlaceNet models with semantic fusion achieve higher AUC scores and maximum recall values than those without

Table 4: Maximum recall (r) at 100% precision for different variants of PlaceNet

Semantic Fusion	**No. of Scales**	**City Center**	**New College**	**KITTI 00**	**KITTI 05**	**KITTI 06**	**Nordland Spr-Win**	**Nordland Spr-Sum**	**Gardens Point**	**TUM-SLAM**
Yes	3	**92.50**	**90.66**	**98.50**	**92.46**	**98.14**	90.26	93.14	**100.0**	**96.32**
Yes	2	90.16	89.00	97.53	90.10	96.47	89.50	92.75	**100.0**	93.76
Yes	1	86.26	83.56	96.68	87.97	94.98	87.91	88.21	99.24	92.48
No	3	78.56	74.42	94.05	85.40	93.75	**93.25**	**93.50**	93.0	86.22
No	2	77.45	73.87	93.50	83.23	93.28	92.69	92.87	92.80	85.24
No	1	72.10	68.32	93.24	82.66	93.10	89.34	90.58	91.17	81.50



(a) City Center

(b) New College

(c) KITTI-00

(d) KITTI-05

(e) KITTI-06

(f) Nordland Spring-Winter

(g) Nordland Spring-Summer

(h) Gardens Point

(i) TUM-SLAM

Figure 6: Precision-Recall results for different datasets. The scale of the x-axis (recall) in the above figures is adjusted for better view and interpretation of results. The Area Under the Curve (AUC) score is shown between parentheses for every model.

semantic fusion for almost all datasets. The most significant improvements correspond to New College, City Center, and TUM-SLAM datasets with an approximate increase of 16%, 14%, and 10%, respectively, in the maximum recall value at 100% precision. However, PlaceNet models with semantic fusion do not boost performance in Nordland sequences and are slightly outperformed by 3% in maximum recall value. Since most of the scenes throughout the train journey in the Nordland dataset are mainly natural scenery; however, most of the training set is collected in urban areas with minimal landscape scenes or country-side views. Thus, the semantic interpretation of objects in Nordland scenes could not help the model much as expected.

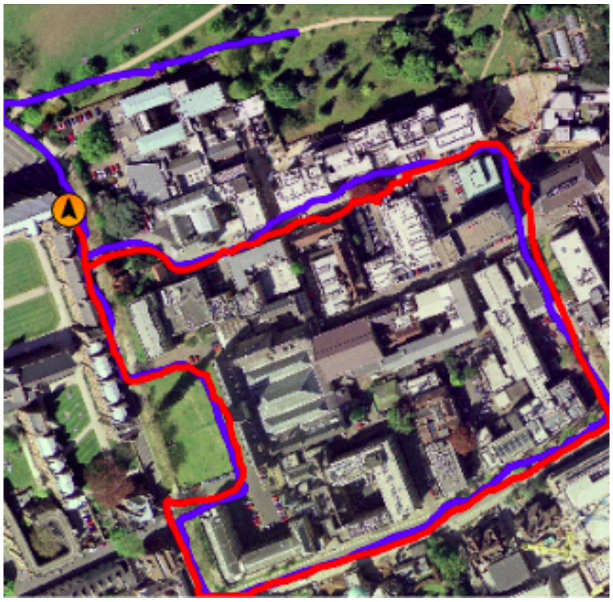


Figure 7: City Center Map. The red trajectory indicates loop closure in the vehicle path.

Moreover, Table 4 illustrates the significant positive impact of multi-scaling on the performance of PlaceNet on all test sets. It is clear that the performance increases as we add different scales of the input frames. This improvement is observed remarkably in New College and City Center datasets with an increase of 7% and 6%, respectively, for the models using semantic fusion when the three scales are employed contrary to the full-scale only.

Finally, we present an example [1] of our experiments on the City Center sequence in Fig. 5. This example highlights the performance of our model and its robustness in detecting loop closures along the trajectory shown in Fig. 7.

5. Conclusion

This paper proposes a novel plug-and-play model for loop closure detection based on a novel multi-scale convolutional auto-encoder network architecture for powerful feature extraction. The main components of this workflow are (i) the semantic fusion network, (ii) the multi-scale encoder-decoder architecture, (iii) the weighted scale-wise loss, and (iv) the similarity measurement with temporal checks. We showed the significant impact of our proposed multi-scale architecture in generating scale-invariant feature representations, besides the effect of semantic fusion that yields semantic aware features that efficiently deal with dynamic environments. We conducted several experiments on various challenging benchmarks and demonstrated that PlaceNet yields competitive results compared to the state-of-the-art approaches for loop closure detection, besides being computationally efficient. Furthermore, we demonstrated that our method is robust to perceptual aliasing, partial occlusion due to moving objects and condition change and powerful in detecting loops in indoor and outdoor scenes. Our future work includes enhancing the performance of PlaceNet in environments with severe weather conditions and non-urban environments. Furthermore, we can extend our model to handle loop closure detection in day-night sequences where the scene undergoes a significant change in illumination between day and night. In that regard, we may consider augmenting scene images with thermal-infrared images since monocular vision sensors operating in the visible spectrum alone suffer from the fundamental limitation of cyclic appearance change over 24 hours.

References

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, J. J. Leonard, Past, present, and future of simultaneous local-

[1] The recorded video of this experiment can be found in: `https://scholar.cu.edu.eg/abayoumi/PlaceNet`

ization and mapping: Toward the robust-perception age, IEEE Trans. Robot. 32 (6) (2016) 1309–1332.

[2] K. A. Tsintotas, L. Bampis, A. Gasteratos, The revisiting problem in simultaneous localization and mapping: A survey on visual loop closure detection, IEEE Transactions on Intelligent Transportation Systems (2022).

[3] M. Cummins, P. Newman, Fab-map: Probabilistic localization and mapping in the space of appearance, Int. J. Robot. Res. 27 (6) (2008) 647–665.

[4] M. Cummins, P. Newman, Highly scalable appearance-only slam - fab-map 2.0, in: Proc. Robot.: Sci. Syst., 2009.

[5] M. Milford, G. Wyeth, Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights, in: IEEE Int. Conf. Robot. Autom., 2012, pp. 1643–1649.

[6] S. Siam, H. Zhang, Fast-seqslam: A fast appearance based place recognition algorithm, in: IEEE Int. Conf. Robot. Autom., 2017, pp. 5702–5708.

[7] A. Torii, J. Sivic, M. Okutomi, T. Pajdla, Visual place recognition with repetitive structures, IEEE Trans. Pattern Anal. Mach. Intell. 37 (11) (2015) 2346–2359.

[8] J. Knopp, J. Sivic, T. Pajdla, Avoiding confusing features in place recognition, in: Eur. Conf. Comput. Vis., Springer, 2010, pp. 748–761.

[9] C. McManus, W. Churchill, W. Maddern, A. Stewart, P. Newman, Shady dealings: Robust, long-term visual localisation using illumination invariance, in: IEEE Int. Conf. Robot. Autom., 2014, pp. 901–906.

[10] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, M. Milford, On the performance of convnet features for place recognition, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2015, pp. 4297–4304.

[11] S. Garg, N. Suenderhauf, M. Milford, Don't look back: Robustifying place categorization for viewpoint-and condition-invariant place recognition, in: IEEE Int. Conf. Robot. Autom., 2018, pp. 3645–3652.

[12] P. Corke, R. Paul, W. Churchill, P. Newman, Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2013, pp. 2085–2092.

[13] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, T. Pajdla, 24/7 place recognition by view synthesis, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (2018) 257–271.

[14] S. Lowry, H. Andreasson, Lightweight, viewpoint-invariant visual place recognition in changing environments, IEEE Robot. Autom. Lett. 3 (2) (2018) 957–964.

[15] L. Wu, Y. Wu, Deep supervised hashing with similar hierarchy for place recognition, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2019, pp. 3781–3786.

[16] B. Talbot, S. Garg, M. Milford, Openseqslam2.0: An open source toolbox for visual place recognition under changing conditions, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2018, pp. 7758–7765.

[17] R. Arandjelović, A. Zisserman, Dislocation: Scalable descriptor distinctiveness for location recognition, in: Asian Conf. Comput. Vis., 2014, pp. 188–204.

[18] S. Cao, N. Snavely, Graph-based discriminative learning for location recognition, in: IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 700–707.

[19] P. Gronat, G. Obozinski, J. Sivic, T. Pajdla, Learning and calibrating per-location classifiers for visual place recognition, in: IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 907–914.

[20] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, J. Sivic, Netvlad: Cnn architecture for weakly supervised place recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1437–1451.

[21] R. Arandjelović, A. Zisserman, Three things everyone should know to improve object retrieval, in: IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 2911–2918.

[22] R. Arandjelovic, A. Zisserman, All about vlad, in: IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 1578–1585.

[23] O. Chum, A. Mikulik, M. Perdoch, J. Matas, Total recall ii: Query expansion revisited, in: IEEE Conf. Comput. Vis. Pattern Recog., 2011, pp. 889–896.

[24] J. Delhumeau, P.-H. Gosselin, H. Jégou, P. Pérez, Revisiting the vlad image representation, in: ACM Int. Conf. Multimedia, 2013, pp. 653–656.

[25] H. Jégou, O. Chum, Negative evidences and co-occurrences in image, in: Eur. Conf. Comput. Vis., 2012.

[26] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: IEEE Conf. Comput. Vis. Pattern Recog., 2010, pp. 3304–3311.

[27] H. Jégou, A. Zisserman, Triangulation embedding and democratic aggregation for image search, in: IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 3310–3317.

[28] G. Tolias, Y. Avrithis, H. Jégou, To aggregate or not to aggregate: Selective match kernels for image search, in: IEEE Int. Conf. Comput. Vis., 2013, pp. 1401–1408.

[29] G. Tolias, H. Jégou, Visual query expansion with or without geometry: refining local descriptors by feature aggregation, Pattern Recognition 47 (10) (2014) 3466–3476.

[30] A. Yandex, V. Lempitsky, Aggregating local deep features for image retrieval, in: IEEE Int. Conf. Comput. Vis., 2015, pp. 1269–1277.

[31] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronin, C. Schmid, Local convolutional features with unsupervised training for image retrieval, in: IEEE Int. Conf. Comput. Vis., 2015, pp. 91–99.

[32] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, M. Milford, Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free, Proc. Robot.: Sci. Syst. (2015) 1–10.

[33] H. Noh, A. Araujo, J. Sim, T. Weyand, B. Han, Large-scale image retrieval with attentive deep local features, in: IEEE Conf. Comput. Vis. Pattern Recog., IEEE, 2017.

[34] A. Babenko, A. Slesarev, A. Chigorin, V. Lempitsky, Neural codes for image retrieval, in: Eur. Conf. Comput. Vis., 2014, pp. 584–599.

[35] A. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, in: IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 806–813.

[36] T. Do, D. Le Tan, T. Pham, N. Cheung, Simultaneous feature aggregating and hashing for large-scale image search, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 4217–4226.

[37] A. Gordo, J. Almazán, J. Revaud, D. Larlus, Deep image retrieval: Learning global representations for image search, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Eur. Conf. Comput. Vis., 2016, pp. 241–257.

[38] Y. Tamaazousti, H. Le Borgne, C. Hudelot, Mucale-net: Multi categorical-level networks to generate more discriminating features, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 6711–6720.

[39] Y. Liu, H. Zhang, Visual loop closure detection with a compact image descriptor, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2012, pp. 1051–1056.

[40] N. Sünderhauf, P. Protzel, Brief-gist - closing the loop by simple means, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2011, pp. 1234–1241.

[41] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[42] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: Eur. Conf. Comput. Vis., Springer, 2006, pp. 404–417.

[43] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Orb: An efficient alternative to sift or surf, in: IEEE Int. Conf. Comput. Vis., 2011, pp. 2564–2571.

[44] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: Eur. Conf. Comput. Vis., Springer, 2006, pp. 430–443.

[45] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, P. Fua, Brief: Computing a local binary descriptor very fast, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1281–1298.

[46] D. Gálvez-López, J. Tardós, Bags of binary words for fast place recognition in image sequences, IEEE Trans. Robot. 28 (5) (2012) 1188–1197.

[47] R. Mur-Artal, J. D. Tardós, Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras, IEEE Trans. Robot. 33 (5) (2017) 1255–1262.

[48] L. Bampis, A. Amanatiadis, A. Gasteratos, Encoding the description of image sequences: A two-layered pipeline for loop closure detection, IEEE/RSJ Int. Conf. Intell. Robots Syst. (2016) 4530–4536.

[49] A. Angeli, D. Filliat, S. Doncieux, J. Meyer, Fast and incremental method for loop-closure detection using bags of visual words, IEEE Trans. Robot. 24 (5) (2008) 1027–1037.

[50] E. Garcia-Fidalgo, A. Ortiz, ibow-lcd: An appearance-based loop-closure detection approach using incremental bags of binary words, IEEE Robot. Autom. Lett. 3 (2018) 3051–3057.

[51] K. Tsintotas, L. Bampis, A. Gasteratos, Assigning visual words to places for loop closure detection, IEEE Int. Conf. Robot. Autom. (2018) 1–7.

[52] M. Labbé, F. Michaud, Appearance-based loop closure detection for online large-scale and long-term operation, IEEE Trans. Robot. 29 (3) (2013) 734–745.

[53] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Adv. Neural Inform. Process. Syst., 2012, pp. 1097–1105.

[54] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, in: Int. Conf. Learn. Represent., 2014.

[55] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: arXiv preprint arXiv:1409.1556, 2014.

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 1–9.

[57] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 770–778.

[58] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: AAAI, 2017.

[59] J. Yu, C. Zhu, J. Zhang, Q. Huang, D. Tao, Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition, IEEE Transactions on Neural Networks and Learning Systems 31 (2) (2020) 661–674. `doi:10.1109/TNNLS.2019.2908982`.

[60] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, M. Milford, Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free, Robotics: Science and Systems XI (2015) 1–10.

[61] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5188–5196.

[62] T. Kanji, Self-localization from images with small overlap, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 4497–4504. `doi:10.1109/IROS.2016.7759662`.

[63] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 224–236.

[64] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, T. Sattler, D2-net: A trainable cnn for joint description and detection

of local features, in: Proceedings of the ieee/cvf conference on computer vision and pattern recognition, 2019, pp. 8092–8101.

[65] P.-E. Sarlin, C. Cadena, R. Siegwart, M. Dymczyk, From coarse to fine: Robust hierarchical localization at large scale, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12716–12725.

[66] B. Cao, A. Araujo, J. Sim, Unifying deep local and global features for image search, in: European Conference on Computer Vision, Springer, 2020, pp. 726–743.

[67] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Adv. Neural Inform. Process. Syst., 2014, pp. 487–495.

[68] Y. Hou, H. Zhang, S. Zhou, Convolutional neural network-based image representation for visual loop closure detection, in: Int. Conf. Inf. Automat., 2015, pp. 2238–2245.

[69] D. Bai, C. Wang, B. Zhang, X. Yi, X. Yang, Cnn feature boosted seqslam for real-time loop closure detection, Chinese J. of Electronics 27 (3) (2018) 488–499.

[70] X. Zhang, Y. Su, X. Zhu, Loop closure detection for visual slam systems using convolutional neural network, in: Int. Conf. Automat. Comput., 2017, pp. 1–6.

[71] J. Lai, Z. Liu, J. Lin, Loop closure detection for visual slam systems using various cnn algorithms contrasts, in: Chinese Automat. Congress, IEEE, 2019, pp. 1663–1668.

[72] S. Wang, X. Lv, D. Ye, B. Li, Compressed holistic convolutional neural network-based descriptors for scene recognition, in: IEEE Int. Conf. Robot. Autom., IEEE, 2019, pp. 135–139.

[73] L. Zuo, C. Zhang, F. Liu, Y. Wu, Performance evaluation of deep neural networks in detecting loop closure of visual slam, in: Int. Conf. on Int. Human Machine Sys. and Cyber., Vol. 2, IEEE, 2019, pp. 171–175.

[74] X. Zhang, L. Wang, Y. Zhao, Y. Su, Graph-based place recognition in image sequences with cnn features, J. Int. Robot. Sys. 95 (2) (2019) 389–403.

[75] Q. Li, K. Li, X. You, S. Bu, Z. Liu, Place recognition based on deep feature and adaptive weighting of similarity matrix, Neurocomputing 199 (2016) 114–127.

[76] Z. Chen, F. Maffra, I. Sa, M. Chli, Only look once, mining distinctive landmarks from convnet for visual place recognition, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2017, pp. 9–16.

[77] X. Gao, T. Zhang, Unsupervised learning to detect loops using deep neural networks for visual slam system, Autonomous robots 41 (1) (2017) 1–18.

[78] N. Merrill, G. Huang, Lightweight unsupervised deep loop closure, in: Proc. Robot.: Sci. Syst., Pittsburgh, PA, 2018.

[79] N. Merrill, G. Huang, Calc2.0: Combining appearance, semantic and geometric information for robust and efficient visual loop closure, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2019, pp. 4554–4561.

[80] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, M. Milford, Deep learning features at scale for visual place recognition, in: IEEE Int. Conf. Robot. Autom., 2017, pp. 3223–3230.

[81] L. Herranz, S. Jiang, X. Li, Scene recognition with cnns: objects, scales and dataset bias, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 571–579.

[82] S. Garg, A. Jacobson, S. Kumar, M. Milford, Improving condition- and environment-invariant place recognition with semantic place categorization, IEEE/RSJ Int. Conf. Intell. Robots Syst. (2017) 6863–6870.

[83] Y. Hou, H. Zhang, S. Zhou, H. Zou, Use of roadway scene semantic information and geometry-preserving landmark pairs to improve visual place recognition in changing environments, IEEE Access 5 (2017) 7702–7713.

[84] M. Hu, S. Li, J. Wu, J. Guo, H. Li, X. Kang, Loop closure detection for visual slam fusing semantic information, in: Chinese Cont. Conf., 2019, pp. 4136–4141.

[85] S. Garg, N. Suenderhauf, M. Milford, Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics, Proc. Robot.: Sci. Syst. (2018).

[86] L. Camara, L. Přeučil, Spatio-semantic convnet-based visual place recognition, in: Eur. Conf. Mob. Robot., 2019, pp. 1–8.

[87] S. Wang, X. Lv, X. Liu, D. Ye, Compressed holistic convnet representations for detecting loop closures in dynamic environments, IEEE Access 8 (2020) 60552–60574.

[88] H. Liu, C. Zhao, W. Huang, W. Shi, An end-to-end siamese convolutional neural network for loop closure detection in visual slam system, in: Intl. Conf. on Acoust. Speech Sig. Proc. (ICASSP), IEEE, 2018, pp. 3121–3125.

[89] S. Appalaraju, V. Chaoji, Image similarity using deep cnn and curriculum learning, arXiv preprint arXiv:1709.08761 (2017).

[90] H. Osman, N. Darwish, A. Bayoumi, Loopnet: Where to focus? detecting loop closures in dynamic scenes, IEEE Robotics and Automation Letters 7 (2) (2022) 2031–2038. doi:10.1109/LRA.2022.3142901.

[91] F. Radenović, G. Tolias, O. Chum, Fine-tuning cnn image retrieval with no human annotation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2019) 1655–1668. doi:10.1109/TPAMI.2018.2846566.

[92] J. Revaud, J. Almazan, R. Rezende, C. D. Souza, Learning with average precision: Training image retrieval with a listwise loss, in: IEEE Int. Conf. Comput. Vis., 2019, pp. 5106–5115. doi:10.1109/ICCV.2019.00521.

[93] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, T. Pajdla, 24/7 place recognition by view synthesis, in: IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 1808–1817. doi:10.1109/CVPR.2015.7298790.

[94] S. An, H. Zhu, D. Wei, K. A. Tsintotas, A. Gasteratos, Fast and incremental loop closure detection with deep features and proximity graphs, Journal of Field Robotics 39 (4) (2022) 473–493.

[95] J. Company-Corcoles, E. Garcia-Fidalgo, A. Ortiz, Appearance-based loop closure detection combining lines and learned points for low-textured environments, Autonomous Robots 46 (3) (2022) 451–467.

[96] K. Zhang, X. Jiang, J. Ma, Appearance-based loop closure detection via locality-driven accurate motion field learning, IEEE Transactions on Intelligent Transportation Systems 23 (3) (2021) 2350–2365.

[97] K. Zhang, J. Ma, J. Jiang, Loop closure detection with reweighting netvlad and local motion and structure consensus, IEEE/CAA Journal of Automatica Sinica 9 (6) (2022) 1087–1090.

[98] Y. Xu, J. Huang, J. Wang, Y. Wang, H. Qin, K. Nan, Esa-vlad: a lightweight network based on second-order attention and netvlad for loop closure detection, IEEE Robotics and Automation Letters 6 (4) (2021) 6545–6552.

[99] M. Gehrig, E. Stumm, T. Hinzmann, R. Siegwart, Visual place recognition with probabilistic voting, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 3192–3199. doi:10.1109/ICRA.2017.7989362.

[100] K. A. Tsintotas, S. An, I. T. Papapetros, F. K. Konstantinidis, G. C. Sirakoulis, A. Gasteratos, Dimensionality reduction through visual data resampling for low-storage loop-closure detection, in: 2022 IEEE International Conference on Imaging Systems and Techniques (IST), IEEE, 2022, pp. 1–6.

[101] K. A. Tsintotas, V. Sevetlidis, I. T. Papapetros, V. Balaska, A. Psomoulis, A. Gasteratos, Bk tree indexing for active vision-based loop-closure detection in autonomous navigation, in: 2022 30th Mediterranean Conference on Control and Automation (MED), IEEE, 2022, pp. 532–537.

[102] K. A. Tsintotas, L. Bampis, A. Gasteratos, Modest-vocabulary loop-closure detection with incremental bag of tracked words, Robotics

and Autonomous Systems 141 (2021) 103782. doi:10.1016/j.robot.2021.103782.

[103] I. T. Papapetros, V. Balaska, A. Gasteratos, Visual loop-closure detection via prominent feature tracking, Journal of Intelligent & Robotic Systems 104 (3) (2022) 1–13.

[104] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in: Eur. Conf. Comput. Vis., 2018, pp. 418–434.

[105] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Intl. Conf. on Medical Image Comput., Springer, 2015, pp. 234–241.

[106] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 3213–3223.

[107] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ade20k dataset, Int. J. Comput. Vis. 127 (3) (2019) 302–321.

[108] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 2881–2890.

[109] F. Chollet, et al., Keras (2015).
URL https://github.com/fchollet/keras

[110] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 3354–3361.

[111] R. Arroyo, P. Alcantarilla, L. Bergasa, J. Yebes, S. Bronte, Fast and effective visual place recognition using binary codes and disparity information, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., 2014, pp. 3089–3094.

[112] D. Olid, J. Fácil, J. Civera, Single-view place recognition under seasonal changes, in: PPNIV Workshop at IROS, 2018.

[113] J. Sturm, N. Engelhard, F. Endres, W. Burgard, D. Cremers, A benchmark for the evaluation of rgb-d slam systems, in: IEEE/RSJ Int. Conf. Intell. Robots Syst., IEEE, 2012.