SuperPlane: 3D Plane Detection and Description from a Single Image

Weicai Ye*[1] Hai Li†[1] Tianxiang Zhang[2] Xiaowei Zhou‡[1] Hujun Bao§[1] Guofeng Zhang¶[1]
[1]State Key Lab of CAD&CG, Zhejiang University [2]Beijing Institute of Spacecraft System Engineering

Abstract

We present a novel end-to-end plane detection and description network named SuperPlane to detect and match planes in two RGB images. SuperPlane takes a single image as input and extracts 3D planes and generates corresponding descriptors simultaneously. A mask-attention module and an instance-triplet loss are proposed to improve the distinctiveness of the plane descriptor. For image matching, we also propose an area-aware Kullback-Leibler (KL) divergence retrieval method. Extensive experiments show that the proposed method outperforms state-of-the-art methods and retains good generalization capacity. The applications in image-based localization and augmented reality also demonstrate the effectiveness of SuperPlane.

Index Terms:

Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Computing methodologies—Artificial intelligence—Computer vision—Computer vision problems

1 Introduction

Finding the correspondence between different views is a key problem in 3D vision tasks such as augmented reality (AR) applications [2–4, 38] and image-based localization (IBL) task [7, 22, 31, 32, 34]. In AR applications, some virtual objects are often placed on the extracted planes [15, 18]. The traditional plane extraction usually follows this paradigm: triangulate the matched feature points to 3D coordinate points from multiple views, and then estimate the planes' parameters by clustering and expanding 3D points. However, it is non-trivial to obtain enough matching feature points in challenging conditions, such as textureless scenes (Fig. 1). Some methods directly perform depth estimation and then triangulate the plane, so that virtual objects can be placed on the plane. But they cannot differentiate semantically different areas. For example, the wall and the door may have the same depth and there will be only one plane detected, which is insufficient to realize the AR effect of hanging a hat on the door. Human-made scenes generally contain rich planar structures, and human perception of the world may be based on individual planar features, rather than low-level feature points or global image features. The mid-level feature such as plane structure can simulate the way humans perceive the world to some extent. In view of this, we highlight that plane detection and description deserves more attention.

While in image-based localization (IBL) task, most existing methods utilize global [1, 20] or semantic [10, 23, 33] features which are highly influenced by salient regions and sensitive to dynamic objects [8], such as moving people, resulting to ambiguous matches. Although SFRS [8] introduces image-to-region supervisions for training image features in a self-supervised manner, it ignores multi-region-to-multi-region supervision.

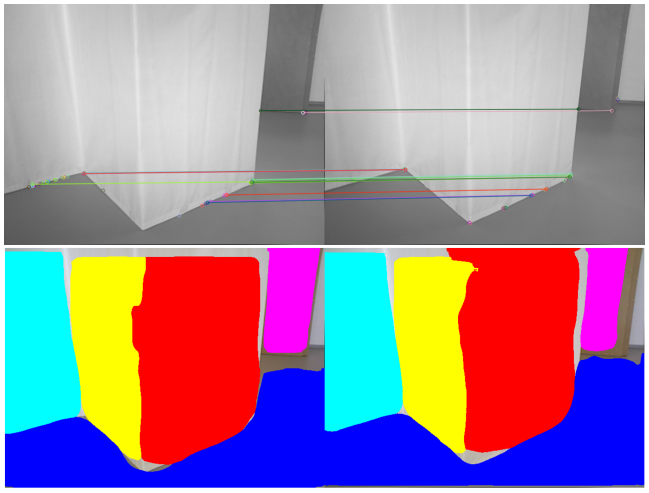


Figure 1: **Weak-texture challenge in man-made scenes.** The left shows the query image and the right shows the reference image [29]. Feature point based methods fail to extract enough matching points to group the plane (the first row) while SuperPlane (the second row) can directly detect the plane and generate the plane description. The matched planes are indicated with the same color.

Recently, there has been several works [16, 17, 37, 39] to detect planes from a single image. Note that all of them only emphasize plane detection but ignore plane description. Even though PlaneMatch [28] introduces a RGB-D patch descriptor designed for detecting coplanar surfaces for RGB-D reconstruction, but overlooks the fact that even if it is a small patch, the patch may still consist of multiple small planes, and using this patch directly to establish constraints may cause some errors. It also fails to detect planes and generates corresponding descriptors from a single image. These existing methods do not perform joint plane detection and description due to quite a few challenging problems, such as how to determine the number of planes and how to describe the detected planes. How to construct data set to supervise plane detection and description is also a critical issue.

For the problems aforementioned, we do the following analysis: plane detection should be related to the object instance in the real world. With the different images obtained, the number of planes detected should also change. For plane descriptor, it should retrain discriminative ability to handle viewpoint changes, even the illumination. We can follow the plane detection network such as PlaneRCNN [16] to detect the planes and construct the triplet samples as supervision of the corresponding plane descriptors. The triplet samples consist of the detected planes, not the complete images.

Based on the above considerations, we propose a novel framework named SuperPlane (Fig. 2) to detect 3D planes and generate corresponding descriptions from a single image. Our model can detect 3D planes and generate the globally consistent descriptors handling the illumination or large viewpoint changes. To the best of our knowledge, we are the first attempt to detect 3D planes with description

*e-mail: yeweicai@zju.edu.cn
†e-mail: garyli@zju.edu.cn
‡e-mail: xwzhou@zju.edu.cn
§e-mail: baohujun@zju.edu.cn
¶e-mail: zhangguofeng@zju.edu.cn, the corresponding author

2642-5254/21/$31.00 ©2021 IEEE
DOI 10.1109/VR50410.2021.00042

2021 IEEE Virtual Reality and 3D User Interfaces (VR) | 978-1-6654-1838-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/VR50410.2021.00042

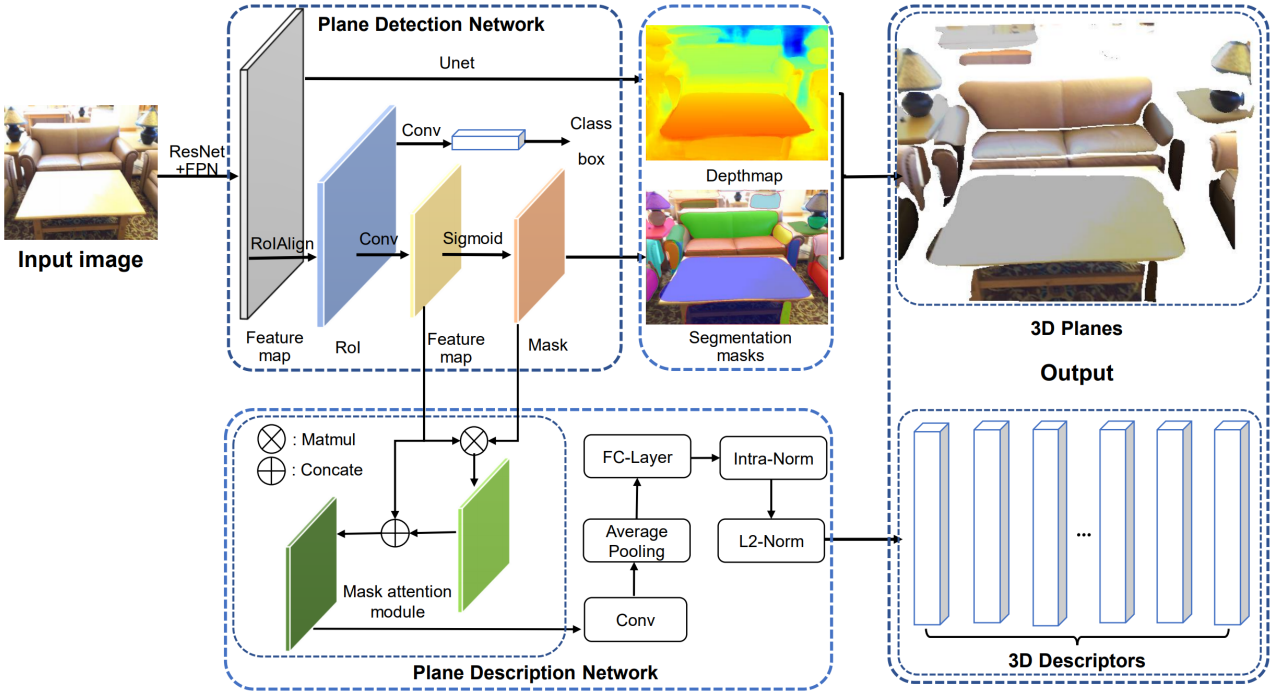


Figure 2: **SuperPlane Framework.** Our framework takes a single RGB image as input, and outputs the 3D planes and corresponding descriptors. Our network consists of three sub-networks: a plane detection network (upper), a plane description network (bottom-right) and a joint warping network (Fig. 3 left) respectively. The plane description network (bottom-left) is composed of a mask-attention module, 3 conv-layer, an average pooling layer, a fully-Connected layer, intra-normalization, and L_2 Normalization.

from a single image. In addition, due to the lack of available datasets for training our framework, we introduce a Plane Description Benchmark (PDB). Furthermore, we propose a novel pipeline (Fig. 3) for IBL task using our 3D plane detection and description network (SuperPlane). Different from the existing retrieval-based methods, they directly use the global features of query and gallery images to calculate the similarity between each other. We additionally use the multi-plane descriptors to get the similarity of the two images. It is non-trivial to merge the many-to-many plane similarity into the similarity of two images. Kullback–Leibler Divergence (KL) is generally utilized to estimate the average difference between the distributions P and Q. We cast every plane descriptor of the image as the distribution, and we can exploit the KL Divergence to estimate the difference of the two images. Since our system detects planes of different sizes, each plane has a different effect on the similarity of the image. We expand the traditional KL divergence to Area-Aware Kullback–Leibler Divergence similarity method to retrieve the similar images, which shows better performance. Overall, our contributions are summarised as four-fold.

- We propose a novel unified framework for simultaneously 3D planes detection and description for a single image.
- We introduce a new training and test benchmark for plane description from a single image and propose an instance-triplet loss to train our model.
- We apply our SuperPlane to image-based localization task and further introduce an Area-Aware Kullback–Leibler divergence retrieval method to retrieve similar images.
- The proposed system outperforms previous state-of-the-art methods on image-based localization benchmarks and demonstrates significant generalization capability.

2 Related Work

In this section, we briefly review the related methods.

2.1 Feature Extraction and Matching

Robust feature extraction and correct matching are particularly important in SLAM or AR applications. There exist several low-level feature detectors and descriptors such as SIFT [21], ORB [26], SuperPoint [6], and so on. It is still a challenge to extract robust features in weakly texture scenes, such as the white wall (Fig. 1 first row). While high-level global or semantic features pay attention to the whole feature, omitting the multi-region features. Human-made environments generally contain rich planar regions, which are often weakly or repeated textures. Furthermore, human perception and understanding of the real world may be more from multiple planes structure, and in AR application, the virtual objects are placed in the planar surfaces [13]. Therefore, we argue that mid-level features, such as plane detection and description, are necessary.

2.2 Plane Detection and Description

Recently several works of plane detection from a single image have emerged. A pioneer work name PlaneNet [17] is proposed for end-to-end planar reconstruction from a single RGB image. PlaneRecover [37] casts the 3D plane recovery problem as a depth prediction problem. However, the above two methods can only handle a fixed amount of planes. While PlaneRCNN [16] based on Mask R-CNN [11], performs a detection network to extract an arbitrary number of planar regions. It further improves the accuracy of plane parameters and depth map with the geometry constraints. PlanarReconstruction [39] exploits pixel-wise embedding and proposes a fast mean-shift clustering algorithm to group pixel embedding into plane instances. It fails to generate the full planar descriptors. Note that all of them only focus on plane detection but ignoring plane description. The most similar work PlaneMatch [28] introduces an RGB-D patch

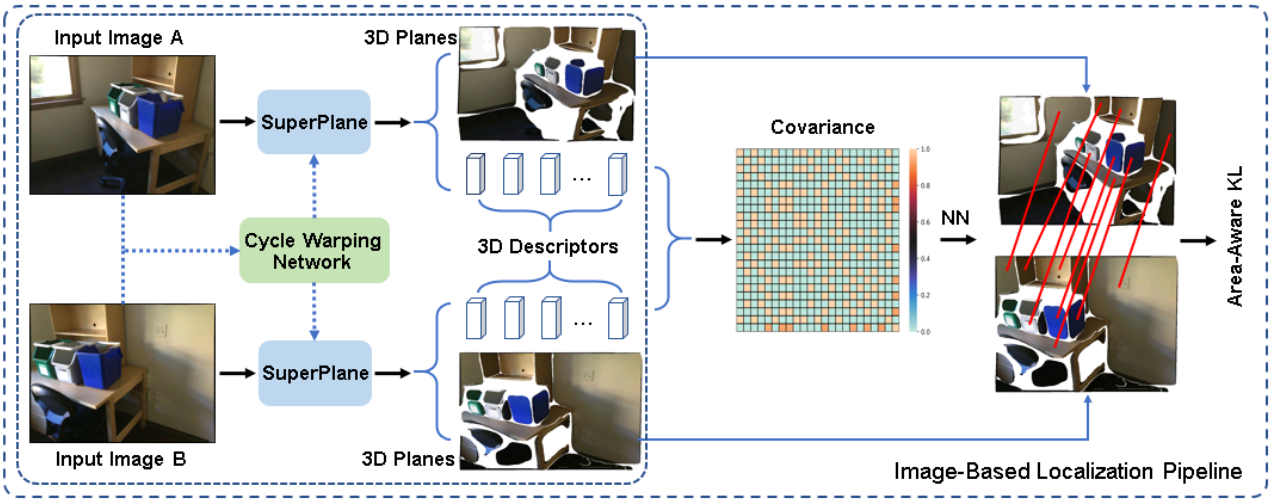


Figure 3: **Image-Based Localization Pipeline with SuperPlane.** We use the SuperPlane network with shared weights to inference the planes and descriptors of the query image and gallery image, respectively. Assume that the query image consists of m planes, and the gallery image consists of n planes. We then calculate the distance between different image planes. In other words, it will form an $m \times n$ matrix. Then we use Nearest Neighbor Search to get the index of the minimum similarity of each row, so that m pairs of matches can be formed. We regard each set of matching planes as two discrete distributions P and Q, so that KL divergence can be used to measure the difference between two images. The planes detected in each group of images are different, and we then propose an Area Aware KL divergence to measure the difference between the two images. Finally, we add the global feature differences of the two images with multiple local planar feature differences to get the final differences between the two images.

descriptor designed for detecting coplanar surfaces and uses the plane constrain to improve RGB-D reconstruction. It fails to detect planes and generates corresponding descriptors from a single integral image. On the contrary, we introduce a novel framework to jointly detect 3D planes and generate corresponding descriptors from a single image.

2.3 Image-Based Localization

Image-based localization task is also considered Place Recognition. Image-based localization is proposed to identify reference images captured at the same places from a geo-tagged database with the given query image. Existing works can be divided into image retrieval-based [1,8,20], per-position classification-based [12,35,36], 2D-3D registration-based [19,27] methods. Our method generates multiple planar descriptors, which are utilized to retrieve similar images. We, therefore, discuss related solutions that cast image-based localization as an image retrieval task. NetVLAD [1] transformed CNN features to local descriptors with learnable semantic centers for localization by proposing a learnable VLAD layer. SARE [20] further looked into effective metric learning to achieve better performance. While SFRS [8] introduces the image-to-region supervisions to mine difficult positive samples for more effective local feature learning. Different from SFRS [8], we utilize multi-region-to-multi-region supervisions to strengthen the discriminability of the feature vectors.

3 Methodology

Our goal is to detect plane instances and generate corresponding plane descriptors from a single RGB image. We propose a novel framework with a multi-branch network to tackle this problem. Our framework, named SuperPlane, consists of three main components (see Fig. 2): a plane detection network, a plane description network with a mask-attention module, and a cycle warping optimization network (see Fig. 3 left). The plane detection network is similar to PlaneRCNN [16]. The plane description network will be elaborated in Sect. 3.2, while the cycle warping optimization network will be demonstrated in Sect. 3.3 . We further apply our baseline to image-based localization tasks and propose an Area-Aware Kullback-Leibler Divergence Retrieval method to retrieve the more similar images, which are highlighted in Sect. 3.6.1. In the following, we will elaborate on the details of each section.

3.1 Plane Detection Network

Similar to [16], we use a plane detection network to detect the plane instances in the image. As shown in Fig. 2, we throw an image into the backbone model to obtain the feature map. The feature map is used to infer the corresponding depth map and segmentation masks of each planar region. For depth map, we use a *U-Net* structure network to recover the depth value with skip connections between Conv and Deconv layers. For the segmentation masks, we use the RoIAlign layer to extract local region information, which is essential for each instance mask inference.

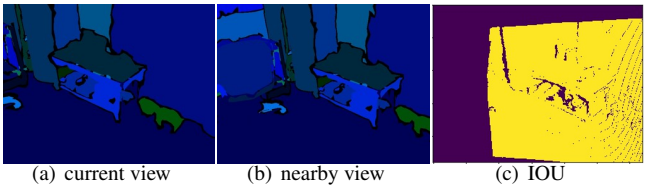
3.2 Plane Description Network

Our framework extends the plane detection branch of PlaneRCNN. It adds a plane description branch after a Region Proposal Network (RPN) module [25], which is detailed in Fig. 2 (bottom). To acquire the compact descriptor, the plane description branch exploits a NetVLAD [1]-like module to represent a plane feature. After obtaining the ROI [25] from RPN, we add three convolution layers, followed by a global average pooling layer. The $L2$-normalized (intra-normalization) converted the matrix generated from the Fully connected layer into a vector, and finally $L2$-normalized in its entirety. And the multiple plane descriptors will be required. Due to the plane detection network that can generate the plane masks, we further propose a mask-attention module to improve the descriptor, which is demonstrated in Sect. 5.3.1. In the mask-attention module, we multiply the feature map in front of the sigmoid layer with the feature map after sigmoid by pixel by pixel to obtain the feature map after the mask and then concatenate the feature map in front of the sigmoid layer with the feature map after the mask. The reason is that the previous paper [28] has shown that mask attention can enhance the distinguishing ability of feature points to a certain extent.

3.3 Cycle Warping Optimization Network

To train the plane descriptor, we develop a Siamese SuperPlane with shared weights (Fig. 3 left), and we construct a triplet sample to pull the positive plane descriptor to the anchor plane descriptor, whereas push the negative plane descriptor away from the anchor plane descriptor. The supervision loss is called Instance-Triplet Loss in Sect. 3.4.1. A cycle warping module is proposed to improve the quality of plane detection and depth estimation (Fig. 3 left).

Inspired by PlaneRCNN [16], we enforce the consistency of reconstructed 3D planes between current view with nearby view during training in cycle warping optimization network. Our Siamese-SuperPlane framework takes two overlapped view images as input and outputs multiple 3D Planes and corresponding descriptors. Note that the pose of the two overlapped view images is known. Each SuperPlane branch takes each frame as input and outputs a pixel-wise depth map. Let M_c and M_n denote the 3D coordinate maps of the current and nearby frames respectly. For every 3D point $P_c \in M_c$ in the current view, we exploit the pose information to project to the nearby frame. And the bilinear interpolation is utilized to read the 3D coordinate P_n from M_n. Based on the camera pose, we transform P_n to the coordinate frame of the current view and compute the 3D distance between the transformed coordinate P_t^c and P_n. The 3D point $P_n \in M_n$ in the nearby view also exploits the projection, un-projection, and coordinate frame transformation to keep the consistency.



(a) current view (b) nearby view (c) IOU

Figure 4: **Plane Description Benchmark.** We warp the current frame through a known pose to the adjacent frame, and then calculate the overlap Intersection over Union (IOU) from the current frame to the adjacent frame. The yellow part is the IOU after warping.

3.4 Loss Function

In this section, we mainly talk about the training loss to supervise our SuperPlane framework. We exploit the plane detection loss in PlaneRCNN [16], and propose an Instance-Triplet loss to train our plane description branch. The plane detection training loss is briefly summarized as follows.

Similar to Fast-RCNN [9], we use L_{cls} and L_{reg} losses for objects detecting and coarse location regression. L_{loc} and L_{mask} losses are used for fine location regression and binary mask predicting. For specific information, please refer to [11].

For depth estimation, we also use the smooth L_1 Loss:

$$L_{depth} = \sum smooth_{L_1}(d_{gt}, d_{pt}) \tag{1}$$

where d_{gt} means the ground-truth depth, while d_{pt} means the predicted depth.

Due to our baseline generates several plane descriptors, we extend standard Triplet loss to $Instance-Triplet$ loss to train our plane description branch. The standard triplet loss function can be described using a Euclidean distance function:

$$L_{triplet} = \sum_m (max(||f(A_m) - f(P_m))||^2 - ||f(A_m) - f(N_m)||^2 + \alpha, 0)) \tag{2}$$

where A_m is an anchor input, P_m is a positive input of the same class as A_m, N_m is a negative input of a different class from A. α is a margin between positive and negative pairs, and f is an embedding function. The standard triplet loss construct the image-level triplet, which is a coarse way to learn the global image representation.

3.4.1 Instance Triplet Loss For Fine-Grained Retrieval

Traditional methods learn a descriptor for each image and construct an image triplet to make the image-level descriptor more discriminative. However, this global matching strategy also lacks some detailed information within an image. For our SuperPlane, we try to learn more detailed information for similar image search. To achieve this goal, we learn a descriptor for each plane instance. Moreover, we propose a plane instance-level triplet loss for fine-grained discriminative feature learning. We do not construct image-level triplets among a training batch but construct the plane instance triplet within an image. For each plane instance within an image, this strategy enhances the discriminability of the plane descriptor. Thus, it can make the discrete probability distribution more discriminative. So, we can evaluate the KL Divergence of two plane descriptor distributions between images with more detailed information, which will lead to better performance. Our *Instance Triplet Loss* randomly selects different plane matching pairs of a set of images for supervision, the negative plane is randomly chosen, which can obtain more supervision information:

$$L_{Instance-Triplet} = \sum_m \frac{1}{k} \sum_{i,j}^{k} (max(||f(A_i) - f(P_i))||^2 - ||f(A_i) - f(N_j)||^2 + \alpha, 0)) \tag{3}$$

where i means the index of the matched planes and j is the random index except for the anchor and positive plane. We do not use the fixed triplet sample, which may contain arbitrary triplet samples, such as hard samples. To reconstruct enough semi-hard samples to train the framework, we propose a selection mechanism, which will be argued in Sect. 3.5.

Finally, the total loss function of our system is defined as follows:

$$L_{total} = \lambda_1 L_{RPN} + \lambda_2 L_{loc} + \lambda_3 L_{mask} + \lambda_4 L_{depth} + \lambda_5 L_{Instance-Triplet} \tag{4}$$

3.5 Plane Description Benchmark

To train our SuperPlane, we use the processed ScanNet [5] data from *PlaneRCNN* [16] and select training triplets by following steps.

- First, we keep the plane indices generated from PlaneRCNN [16]. Directly using every 20 adjacent frames in PlaneRCNN [16] to extract matching pairs may result in some easy samples, so we warp the current frame through a known pose to the adjacent frame, and then calculate the overlap Intersection over Union (IOU) from the current frame to the adjacent frame, as shown in Fig. 4.
- Second, with the computed IOU, we can divide the data set into three levels: simple(0.7-1.0), semi-hard(0.4-0.7), and hard(0.1-0.4). We mainly do experiments on semi-hard data sets. For all scenes in Scannet [5], we follow the Scannet train/val/test split metric. For a single scene, we also split the dataset to train/val/test subset with ratio 90%, 5%, 5%.
- Therefore, for every image pairs, it has multi-corresponding planes. Each pair contains a corresponding match (plane indices – plane indices), the relative pose, and the camera pose of each image. Every plane consists of plane parameters (such as Normal N and Offset d), mask information, depth, and global plane indices.

Table 1: **Comparison with state-of-the-arts on image-based localization benchmarks.** Note that the network is only trained on the proposed Plane Description Benchmark (PDB-train) and directly evaluate on Tokyo 24/7, Pitts250k-test and Pitts30k-test datasets. Our method outperforms existing state-of-the-art methods. We mark best results **bold**.

	Tokyo 24/7 [31]			Pitts250k-test [31]			Pitts30k-test [31]		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [1]	73.33	82.86	86.03	85.95	93.20	95.13	80.5	91.8	95.2
CRN [14]	75.20	83.80	87.30	85.50	93.50	95.50	-	-	-
SARE [20]	79.68	86.67	90.48	88.97	95.50	96.79	-	-	-
SFRS [8]	85.40	91.11	**93.33**	**90.68**	96.39	97.57	89.38	94.70	95.88
Ours	**85.71**	**92.38**	93.02	90.62	**96.47**	**97.67**	**89.48**	**94.73**	**96.01**

Table 2: **Comparison with state-of-the-arts on depth estimation benchmarks.** We evaluate the depth estimation on the ScanNet [5] dataset with state-of-the arts methods. Our method is slightly better than PlaneRCNN [16]. We mark best results **bold**.

	Lower the better (LTB)					Higher the better (HTB)		
	Rel	Rel(sqr)	$\log_{10}$	RMSE	RMSE_{log}	1.25	1.25^2	1.25^3
PlaneRCNN [16]	0.139	0.063	0.064	0.329	0.183	0.784	0.955	**0.994**
Ours	**0.137**	**0.063**	**0.061**	**0.319**	**0.177**	**0.810**	**0.963**	0.990

3.6 Image-Based Localization Pipeline

The retrieval-based image localization method mainly takes this paradigm: firstly, obtain the descriptor of the query image and the gallery image, and then calculate the similarity between the query and the gallery image to confirm whether the query image and the gallery image are obtained under the same GPS. Our pipeline is different from the existing scheme. What we obtain is not only the global feature vector of a single image, but also the multiple plane feature vectors of the image. We developed a strategy for integrating multi-plane matching similarity into the similarity of the entire image. As shown in Fig. 3, we first compute multi-plane descriptors and the global feature from each image, and the covariance between the query image and the gallery image is computed. We further perform the nearest neighbor search to capture the most similar plane for each plane in the query image. Then we will get the candidate matching for each row in the covariance matrix. We aggregate the multi-plane distances to a whole distance between two images with the proposed Area-Aware Kullback-Leibler Divergence Retrieval method. Finally, we add the global feature differences (from existing global feature methods [1, 8, 20]) of the two images with multiple local planar feature differences to get the final differences between the two images.

3.6.1 Area-Aware Kullback-Leibler Divergence Retrieval

In mathematical statistics, the standard Kullback-Leibler divergence is a measure of the difference between two probability distributions. For discrete probability distribution P and Q define on the same probability space, the Kullback–Leibler divergence from Q to P is defined to be:

$$D_{KL}(p||Q) = \sum_{x \in \mathscr{X}} P(x) log(\frac{P(x)}{Q(x)}). \quad (5)$$

While retrieval-based IBL task can be cast as the Kullback-Leibler divergence between the query image and the gallery images. Our proposed SuperPlane can generate multiple plane descriptors from a single image, so every plane descriptors can be naturally cast as the discrete distribution. Since the area of each plane is different, its importance is also different. We further propose an Area-Aware Kullback-Leibler Divergence Retrieval to adjust the measure metric. The metric can be designed as follow:

$$D_{Area-AwareKL}(p||Q) = \sum_{x \in \mathscr{X}} Area(x) P(x) log(\frac{P(x)}{Q(x)}) \quad (6)$$

where $D_{KL}(p||Q)$ means the distance of the plane descritor distributions between two images. Small distance mean they are similar and vice versa.

4 Implementation Details

Our framework is implemented by Pytorch [24], an imperative style, high-performance deep learning library. We adopt the same framework used in PlaneRCNN [16] and exploit a VLAD [1] layer for encoding and aggregating plane feature descriptors. Different from PlaneRCNN [16], we use the proposed PDB semi-hard dataset to train the plane description branch. We first fixed the pretrained plane detection branch in PlaneRCNN, and only train the plane description. When the description branch is close to convergence, we do not fix the plane detection branch and continue training until the network converges. Our model is trained by 600 iterations with a fixed plane detection branch, and further trained to 1200 iterations. The Adam algorithm is utilized to optimize the loss function, with a constant learning rate 1e-4, momentum 0.99, and weight decay 0.0001.

5 Experiments

We conduct the experiments in five aspects: ablation studies of our proposed framework on the proposed Plane Description Benchmark (PDB), comparison with state-of-the-art depth estimation methods on the ScanNet benchmark, compared with the retrieval-based methods on several image-based localization Benchmark, generalization capability and limitations and the application for AR with our SuperPlane.

5.1 Datasets

ScanNet [5] is a dataset of richly-annotated RGB-D scans of real-world environments containing 2.5M RGB-D image in 1,513 scans acquired in 707 distinct spaces. We followed the split metric of the PlaneRCNN [16] to evaluate the performance of the depth estimation.

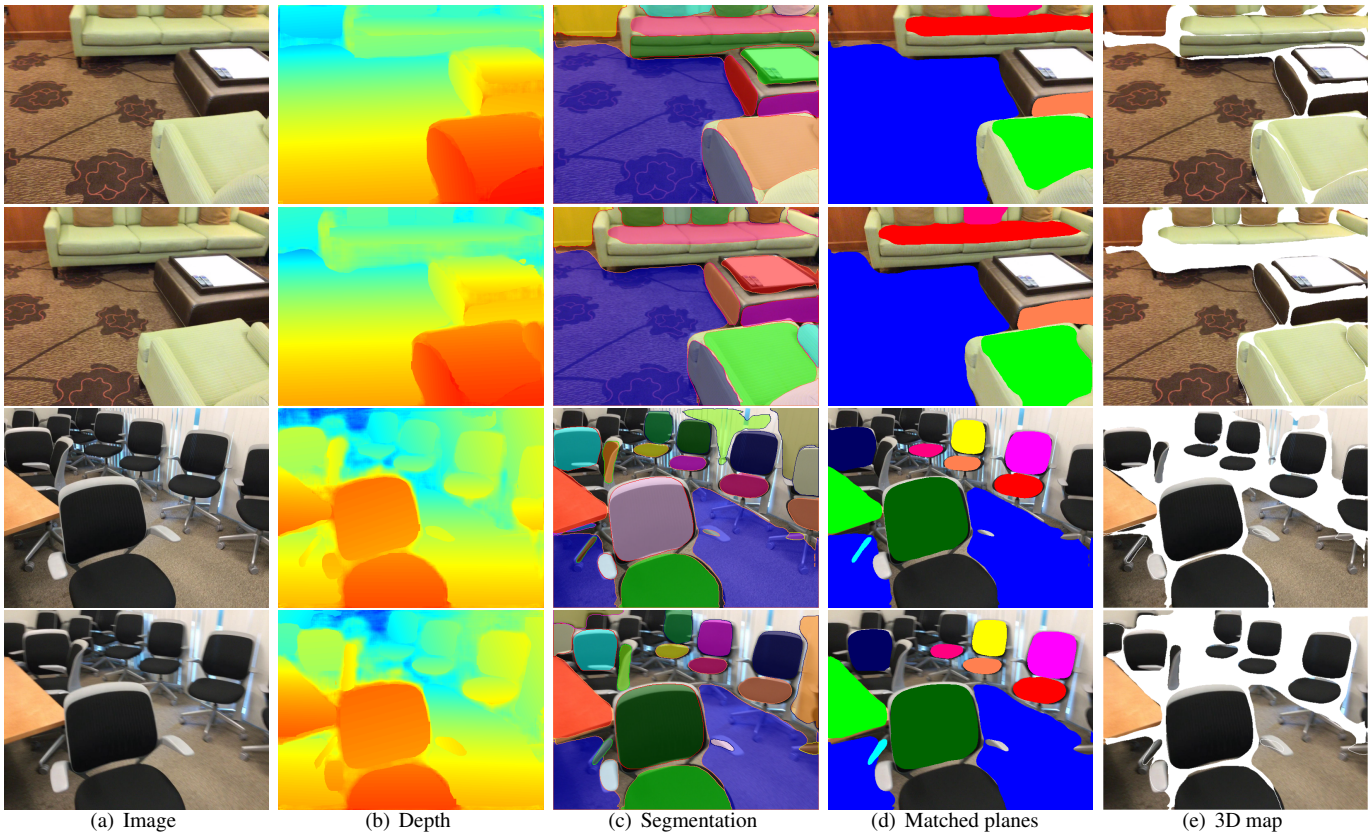
Pittsburgh [31] is the unified IBL dataset that consists of a large scale of panoramic images captured at different times and is associated with noisy GPS locations. The Pitts30k-val consists of 7, 608 queries and 10, 000 gallery images, and the Pitts250k-test contains 8, 280 probes and 83, 952 database images.

Tokyo 24/7 [30] is also widely used on IBL task. It is quite challenging since the queries were taken in varying conditions.

In addition, to verify the power of our method, we further apply our trained SuperPlane to IBL task and evaluate on Pitts30K-val, Pitts250K-val, and Tokyo 24/7 dataset. Note that we do not train on the above datasets. We follow the state-of-the-art retrieval-based IBL method for a fair comparison.

5.2 Evaluation

We evaluate the plane matching of our method using precision and recall metric on the proposed PDB dataset. Precision measures



(a) Image (b) Depth (c) Segmentation (d) Matched planes (e) 3D map

Figure 5: **Qualitative results of our method.** Every two rows are a pair of images with the change of viewpoint. From left to right: RGB image, depth map, segmentation, matched planes, and 3D map. The results demonstrate that our framework can yield stable plane detection and maintain matching consistency in repeated texture scenes.

the plane matching result relevancy, while recall measures how many truly plane matching relevant results are returned. For depth estimation, we follow the same evaluation metric used in [16] to evaluate the accuracy between the predicted depth map with the ground-truth depth. On retrieval-based IBL tasks, we follow the same evaluation metric proposed by [8], where the top-k recall is measured. If at least one of the top-k retrieved reference images is located within d=25 meters from the query image, it is determined that the query image has been successfully retrieved from the top-k.

Table 3: **Ablation studies on the proposed PDB dataset.** The baseline directly combines the Plane Detection Network of PlaneRCNN [16] and VLAD [1] layer. The second model reserves the baseline, with additional mask attention module. Then we add the cycle wraping module to improve the precision and recall.

method	precision	recall
baseline	0.7386	0.9123
+mask attention	0.7525	0.9332
+mask attention & cycle wraping	**0.7864**	**0.9715**

5.3 Ablation studies

5.3.1 Framework study

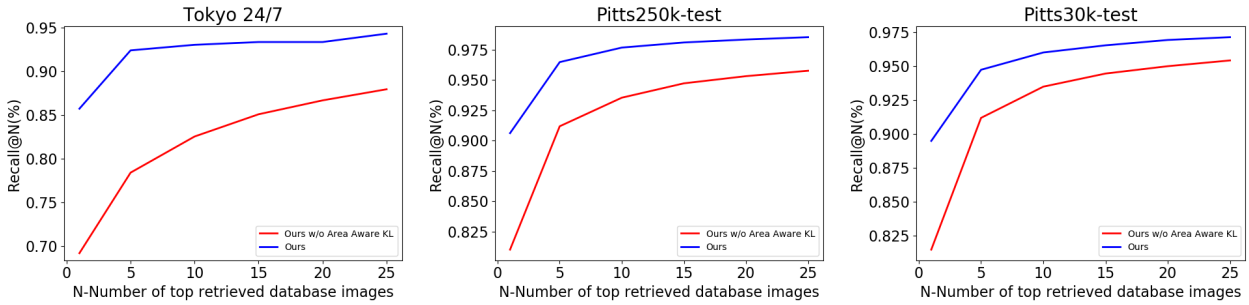
We utilize the proposed Plane Description Benchmark (PDB) for optimizing our Siamese-like SuperPlane Network following the experimental setting of state-of-the-art Plane Detection methods [16]. To our best knowledge, we are the first to propose detecting 3D planes and descriptions from a single image. Because we can not find the same work, we only do experiments and report some results on our PDB dataset. In the proposed Plane Description Benchmark (PDB), we perform precision and recall metrics to analyze the effectiveness of the proposed method. Table 3 demonstrates that the mask-attention module strengthens the plane descriptor's discriminate power. And the cycle warping optimization module further improves the precision and recall values. The qualitative results shown in Fig. 5 demonstrate that our framework can yield stable plane detection and maintain matching consistency in repeated texture scenes. We also provide supplementary videos to show the temporal consistency of our method in plane detection and matching.

5.3.2 Kullback-Leibler Divergence Study

In our image-based localization pipeline, we exploit two KL divergence methods to retrieval the similar images. "Our w/o area aware KL" is the baseline using the standard KL divergence. Fig. 6 demonstrates that the proposed Area Aware KL divergence outperforms standard KL divergence on Tokyo 24/7, Pitts250K-test and Pitts30k-test datasets.

5.4 Comparison with state-of-the-art depth estimation methods

We evaluate our depth estimation on the ScanNet dataset [5] and compare with state-of-the-art depth estimation methods. PlaneRCNN [16] is the most relevant work. Table 2 demonstrates that our method is generally better than PlaneRCNN [16]. The left five columns show different depth error metrics including rooted-mean-square-error (RMSE) and Related Difference (Rel) on the average



(a) Tokyo 24/7 (b) Pitts250k-test (c) Pitts30k-test

Figure 6: Ablation studies for our proposed Area-Aware Kullback-Leibler Divergence with the standard Kullback-Leibler Divergence in our pipeline on IBL tasks. Area-Aware KL method has a superior performance compared with the standard KL method.

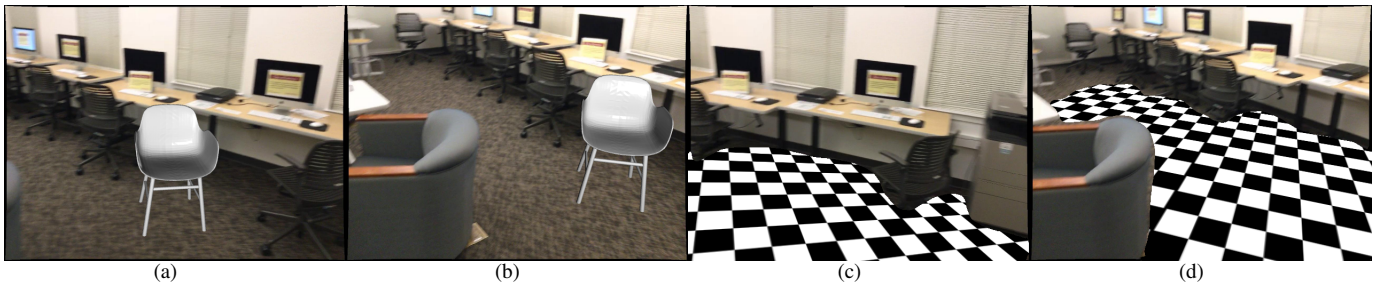
(a) Query (b) Our top-1 (c) SFRS's top-1 (d) Query matched planes (e) Our matched planes

Figure 7: **Qualitative results for viewpoint and illumination changes of our method and state-of-the-art SFRS [8]**. Compared with SFRS, the retrieved results of our method can maintain good discrimination ability under large viewpoint or illumination changes. The first two rows are for viewpoint changes while the last two rows are for illumination changes. The first column is the query image, the second column is our top1 retrieval results, the third column is SFRS's top1 results, the last two columns are the matched planes between query image and our top1 results, respectively. The same color indicates the matched planes.

per-pixel depth errors [17]. The lower the better. The right three columns present the ratio of pixels for which the relative difference between the ground truth depths and the predicted depths is below the threshold. The higher the better.

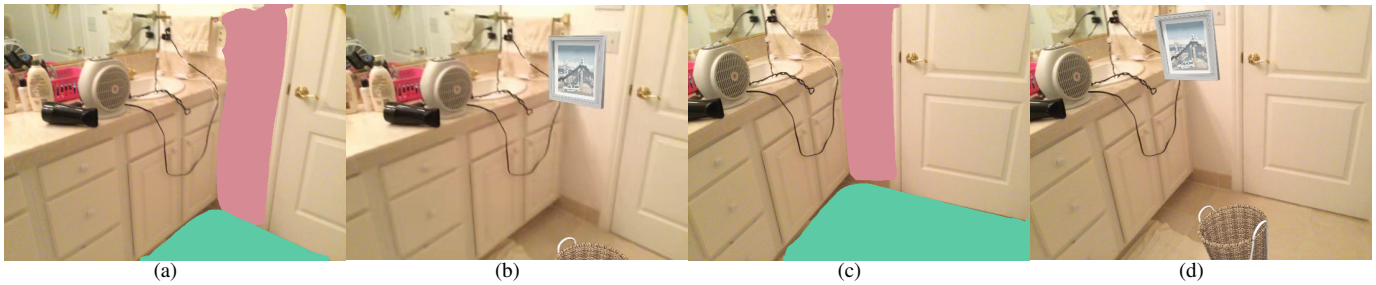
5.5 Comparison with state-of-the-art retrieval-based methods

We compare the proposed IBL Pipeline with state-of-the-art image localization methods NetVLAD [1], CRN [14], SARE [20] and SFRS [8] on localization datasets Pitts30k-test, Pitts250k-test, and Tokyo 24/7 in this experiment. We combine the standard Kullback-Leibler Divergence with our generated plane descriptors. We further exploit Area-Aware Kullback-Leibler Divergence Retrieval method, which has a superior performance compared with the standard Kullback-Leibler Divergence method. Experimental results demonstrate that our method outperforms the state-of-the-art methods, as shown in Table 1. These methods extract a global feature from the entire image which may be sensitive to dynamic objects, resulting



(a) (b) (c) (d)

Figure 8: **Single-plane detection for AR applications.** We place the chair on the detected floor plane, as shown in (a) and (b). As the viewpoint changes, our plane can be stably detected, and the resulting plane descriptor can maintain robust tracking. We apply texture mapping to the detected floor plane, and the plane texture can be smoothly generated as the viewpoint changes, as shown in (c) and (d).



(a) (b) (c) (d)

Figure 9: **Multi-planes detection for AR applications.** Our method can detect multiple planes, which can support multiple virtual objects placing. The plane detection and tracking are stable even the viewpoint is largely changed.

in false matches. Instead, our method is not only based on global feature but local plane features in the background and is supposed to be more robust to dynamic foreground.

For better understanding the superior performance of our method on the IBL tasks, we show the retrieval image compared with SFRS [8] with the given image. The recall top-1 images shown in Fig. 7 demonstrate that with the proposed method, our retrieval system can handle large change with illumination or viewpoint. The reason is two-fold. On the one hand, during the training process, the images are selected according to the IOUs, which can cover large viewpoint changes. In addition, the training dataset also contains some illumination changes. On the other hand, our model implicitly encodes various local cues including plane, contour and semantic information, so it can handle complex scenes more robustly.

5.6 Generalization Ability and Limitations

Since our framework is only trained on the proposed Planar Description Benchmark (PDB) and evaluated on several new datasets, extensive experiments show that our method retains the significant generalization capability on standard image retrieval tasks.

Our method assumes that the intrinsic parameters of the captured images are known. If the gap between the ground-truth and the given intrinsic parameters is large, it may cause inaccurate plane detection and description. Also, if the number of planes is not enough, the plane-based image matching accuracy may degrade. We will explore self-supervised training methods and combine with optical flow estimation to refine the matching accuracy in the future.

5.7 Application for Augmented Reality

We adopt some AR applications to show the Plane Detection and Plane Description's capability with our framework. Plane detection is a basic task in AR applications, which is usually used to place virtual objects. It is non-trivial for feature-based methods to capture enough matching feature points to construct the planes in weakly-texture scenes. However, our method can easily detect the multiple planes and can support the user to conveniently place the target objects. In AR applications, long-term user interaction will inevitably accumulate errors, and the system needs to automatically eliminate errors. Commonly used solutions may be loop closure detection or re-localization, where image retrieval is usually needed. As demonstrated, our plane-based image retrieval method can handle weak texture, repeated texture, perspective changes, illumination changes and other challenge scenarios. As shown in Fig. 8 and Fig. 9, our method yields stable plane matching results.

6 Conclusions and Future Work

This paper introduces a novel framework named SuperPlane to detect 3D planes and generate corresponding descriptors from a single image and build a new Plane Description Benchmark to facilitate future research in this direction. The proposed Area-Aware Kullback-Leibler divergence retrieval method gives rise to state-of-the-art IBL results on Tokyo 24/7, Pitts250k and Pitts30k datasets. Through the applications in image-based localization and augmented reality, SuperPlane demonstrates the strong power of plane matching in the challenge scenarios.

In the future, we will explore the self-supervised method to train our network to mitigate the need of the given intrinsic parameters, and combine with the optical flow estimation or other methods to further refine the plane matching accuracy.

Acknowledgments

The authors thank Hongjia Zhai, Ziyang Zhang and He Wang for their kind help in proofreading, and all the reviewers for their constructive comments to improve this paper. This work was partially supported by NSF of China (Nos. 61822310 and 61932003).

REFERENCES

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5307, 2016.

[2] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. Recent advances in augmented reality. *IEEE computer graphics and applications*, 21(6):34–47, 2001.

[3] R. T. Azuma. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385, 1997.

[4] J. Carmigniani, B. Furht, M. Anisetti, P. Ceravolo, E. Damiani, and M. Ivkovic. Augmented reality technologies, systems and applications. *Multimedia tools and applications*, 51(1):341–377, 2011.

[5] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

[7] D. Gálvez-López and J. D. Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

[8] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li. Self-supervising fine-grained region similarities for large-scale image localization. *arXiv preprint arXiv:2006.03926*, 2020.

[9] R. Girshick. Fast RCNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[10] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-match: an algorithm and an implementation of semantic matching. In *European semantic web symposium*, pp. 61–75. Springer, 2004.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

[12] P. Hongsuck Seo, T. Weyand, J. Sim, and B. Han. CPlaNet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 536–551, 2018.

[13] N. Huang, J. Chen, and Y. Miao. Optimization for RGB-D slam based on plane geometrical constraint. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 326–331, 2019.

[14] H. J. Kim, E. Dunn, and J.-M. Frahm. Learned contextual feature reweighting for image geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3251–3260. IEEE, 2017.

[15] G. Klein and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 225–234, 2007.

[16] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz. PlaneRCNN: 3D Plane Detection and Reconstruction From a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4450–4459, 2019.

[17] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa. PlaneNet: Piece-wise planar reconstruction from a single RGB image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2579–2588, 2018.

[18] H. Liu, G. Zhang, and H. Bao. Robust Keyframe-Based Monocular SLAM for Augmented Reality. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pp. 340–341, 2016.

[19] L. Liu, H. Li, and Y. Dai. Efficient global 2D-3D matching for camera localization in a large-scale 3D map. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2372–2381, 2017.

[20] L. Liu, H. Li, and Y. Dai. Stochastic attraction-repulsion embedding for large scale image localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2570–2579, 2019.

[21] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 vol.2, 1999.

[22] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2015.

[23] J. McCormac, A. Handa, A. Davison, and S. Leutenegger. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *Proceedings of the IEEE International Conference on Robotics and automation (ICRA)*, pp. 4628–4635. IEEE, 2017.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 8026–8037, 2019.

[25] S. Ren, K. He, R. Girshick, and J. Sun. Faster RCNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the International Conference on Computer Vision*, pp. 2564–2571. Ieee, 2011.

[27] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *Proceedings of the International Conference on Computer Vision*, pp. 667–674. IEEE, 2011.

[28] Y. Shi, K. Xu, M. Niessner, S. Rusinkiewicz, and T. Funkhouser. PlaneMatch: Patch coplanarity prediction for robust rgb-d reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 750–766, 2018.

[29] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D slam systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573–580. IEEE, 2012.

[30] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, 2015.

[31] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 883–890, 2013.

[32] A. Torralba, K. P. Murphy, W. T. Freeman, M. A. Rubin, et al. Context-based vision system for place and object recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 3, pp. 273–280, 2003.

[33] N. Ufer and B. Ommer. Deep semantic feature matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6914–6923, 2017.

[34] I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of the IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, vol. 2, pp. 1023–1029. Ieee, 2000.

[35] N. Vo, N. Jacobs, and J. Hays. Revisiting IM2GPS in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2621–2630, 2017.

[36] T. Weyand, I. Kostrikov, and J. Philbin. Planet-photo geolocation with convolutional neural networks. In *Proceedings of the European Conference on Computer Vision*, pp. 37–55. Springer, 2016.

[37] F. Yang and Z. Zhou. Recovering 3D planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100, 2018.

[38] H. Yu, W. Ye, Y. Feng, H. Bao, and G. Zhang. Learning bipartite graph matching for robust visual localization. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 146–155, 2020.

[39] Z. Yu, J. Zheng, D. Lian, Z. Zhou, and S. Gao. Single-image piece-wise planar 3D reconstruction via associative embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1029–1037, 2019.