Leveraging Planar Motion Property for Robust Visual Matching and Localization

Yanmei Jiao, Bo Fu, Qunkang Zhang, Fuzhang Han, Yue Wang, and Rong Xiong

***Abstract*— One primary difficulty preventing the visual localization for service robots is the robustness against changes, including environmental changes and perspective changes. In recent years, learning-based feature matching methods have been widely studied and effectively verified in practical applications. Learning-based feature matching effectively solves the problem of environmental changes, including illumination changes and man-made changes. However, there is still room for improvement dealing with large perspective changes. In this paper, we leverage planar motion property to simplify the affine transform and propose an augmentation-based feature matching method that greatly enhances the robustness to perspective changes. The proposed feature matching approach maintains low matching costs as the augmentation is performed on the simplified affine matrix space. Combined with the planar motion property aided minimal solution for pose estimation, an end-to-end robust visual localization system is proposed which is shown to bring 67% improvement in localization performance under large perspective changes with a 20% increase in computation. In addition, a guide for map frame selection is presented to support robust localization with very sparse frames in storage. Experiments on the classified dataset with environmental changes and perspective changes validate the effectiveness of the proposed system.**

I. Introduction

The last decade has witnessed a fast-growing volume of service robots applied in various indoor scenarios e.g. bank lobby, restaurant, and home. One of the fundamental techniques for these robots is localization [1]–[3]. To further save the cost of robots, vision based localization becomes a popular solution [4], [5], which is widely studied and relatively mature for autonomous vehicles [6]–[8]. However, compared with autonomous vehicles, the localization for service robots has two additional difficulties. Firstly, when dealing with *environmental changes* (EC), we can not assume semantic features to be known as autonomous driving, so only low-level point-based features can be utilized. However, the point features are sensitive to long-term EC, such as illumination variations due to complex lighting conditions (including natural and artificial lighting) in the room, and changes and occlusion of objects due to frequent human activities in the indoor environment. Secondly, the *perspective changes* (PC) introduced by camera motion is more serious than in outdoors, as the distance between robot and scene is shorter, where small turns of the robot can lead to large changes in image appearance. In addition, indoor service robot's route is not constrained by lanes. The turn of service robot is thus free, and larger PC will exist. The difficulty in maintaining the robustness of large PC lies in the small overlap area between images and the existence of deformation. Therefore, not only the outlier rate in feature matches is high, but also the number of matches is small.

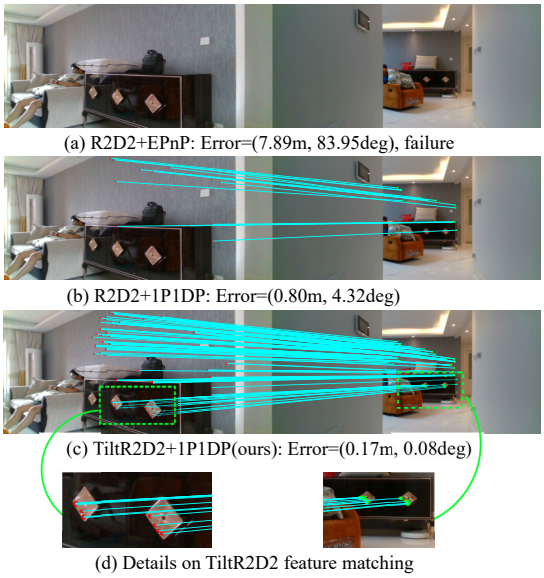


Fig. 1. Comparison of the localization performance in the same indoor scene. The cyan lines between two images are inliers found by the corresponding pose estimation algorithm among all correspondences matched by different features. The proposed augmentation based feature matching TiltR2D2 matches a large number of inliers under large perspective changes, which assists the subsequent pose estimation algorithm to achieve better localization accuracy.

To deal with the above changes, many learning-based features are proposed that significantly improves the visual matching robustness by training on a large number of images with various changes [9]. However, a previous study finds that even with learning-based feature matching, perspective changes still severely hurt the localization performance [10].

An intuitive idea to eliminate the effect of large viewpoint variation is to enrich feature matches by further augmenting query image with affine transform [11]. The augmentation brings more feature matches, consequently more inliers. Unfortunately, at the same time, the augmentation is expensive both in time and hardware due to multiple times of feature matching and may lead to a higher outlier rate.

In this paper, we combine the learning-based feature with classic image augmentation to overcome the large perspective changes and leverage the planar motion property to address

Yanmei Jiao, Bo Fu, Qunkang Zhang, Fuzhang Han, Yue Wang and Rong Xiong are with the State Key Laboratory of Industrial Control and Technology, Zhejiang University, Hangzhou, China, and with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China. (Corresponding author: Yue Wang, Email: wangyue@iipc.zju.edu.cn).

the two challenges. For efficiency, we introduce planar motion property to simplify affine transform, leading to reduced dimensions of augmentation and smaller matching costs. In addition, we leverage the robustness to perspective changes obtained by the proposed method to explore the map frame selection guide, which can support the localization with very sparse frames in storage, satisfying the service robots hardware requirement. For picking inliers, we use the minimal solution proposed in our previous work which models the reduced subspace to consider both features with and without depth, to enable robust pose estimation in features with a high percentage of outliers [12]. Based on the above two points, we propose a simple-yet-effective robust visual localization system for service robots. Comparison of localization performance shown in Fig. 1 demonstrates that the proposed localization system can obtain more inliers and achieve better localization accuracy under large perspective changes. In experiments, we manually divide data in the dimension of PC and EC, to study the effect of PC on localization performance of service robot firstly, and secondly compare with other methods in finer granularity. Results show that the proposed system is effective for both PC and EC, with a significant improvement for PC. In summary, the contributions of this work are listed as follows:

- An end-to-end planar motion aided visual localization system framework is proposed which integrates the augmentation-based feature matching at the front-end and the minimal solution at the back-end to deal with a high percentage of outliers.
- An efficient augmentation-based feature matching method is proposed after simplification of affine distortion matrix using planar motion property to deal with the large viewpoint variations.
- A simple map selection guide is explored which can support the localization with very sparse frames in storage, satisfying the service robots hardware requirement.
- Evaluation experiments about the effects of EC and PC in performance of indoor localization among different features are presented which validates the effectiveness of the proposed system, and the source code to divide EC and PC is available on github[1].

II. Related Works

A. Visual localization

Traditionally, there are two lines of visual localization approaches. The first is the image retrieval based localization, which can also be called place recognition based methods. The pose of query image is approximated by the pose of top-retrieved database images using the image-based representations for indexing [13]–[15]. The second is 3D model based approaches, which solve the camera pose using matched feature correspondences between query image and pre-built 3D map [16]–[18], leading to much higher accuracy. However, the performance will degenerate significantly facing serious changes in environment with a high outlier rate in feature correspondences. Recently, the hierarchical localization framework is widely used which converges the two lines of methods by using place recognition to reduce the local region and feature based localization for accurate pose estimation [6]. The idea of incorporating higher-level scene understanding into visual localization for autonomous vehicles has attracted more attention in recent years. The semantic features extracted on informative regions (such as sky, building, vegetation, road and pole) are integrated to enhance the feature matching robustness [19] or pose refinement [20].

B. Motion property for localization

When dealing with outliers in feature matching, the robust pose estimation algorithm is needed to reject outliers. The typical way to achieve robust estimation is random sample consensus (RANSAC) [21]. To enhance the robustness of RANSAC, many efforts have been made to reduce the minimum number of feature correspondences for pose estimation by leveraging motion property. One type of motion property is the measurement of pitch and roll angles of a robot obtained by gravity direction alignment from inertial measurements. With the known pitch and roll angles, the minimal solutions using two points are presented in [8], [22]. Another type of motion property is planar motion which is commonly used in indoor localization. In [23], the minimal solution using only a single affine correspondence is derived under planar motion. In our previous work, the minimal solution using one point correspondence with depth and another one without depth for planar moving robots is shown to achieve more robust performance that can deal with up to 80% outlier rate. However, the performance of RVL is still limited by the absolute number of inliers, even when outlier removal is perfect [24], [25].

C. Visual matching

For robust visual matching, learning-based features [26]–[28] are shown to outperform conventional hand-crafted methods, like SIFT [29] and ORB [30], in presence of environmental variations. To generate more matches, NC-Net [31] builds a dense volume encoding per-pixel similarity between two images, which enables a dense feature matches after learning. However, when dealing with large perspective changes, the performance of learning-based methods is significantly reduced due to the small overlap and deformation [10]. To obtain more inliers when large perspective changes exist, ASIFT [11] is the first to achieve full affine invariance by simulating all affine parameters for augmentation. However, it requires a considerable number of image augmentation, which is time-consuming and difficult to meet the requirement of robot applications in natural scene localization.

III. Motion Property aided Localization System

In this paper, we leverage the planar motion property of wheeled mobile robots to increase the robustness of both feature matching and pose estimation, thus improving the

[1] https://github.com/slinkle/Divided-Openloris

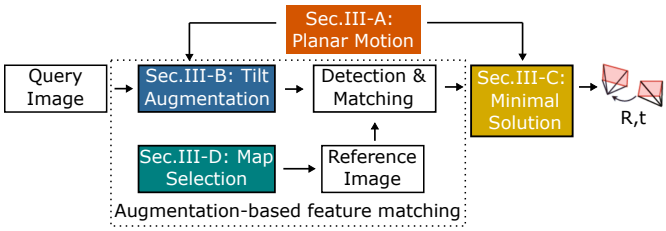


Fig. 2. The framework of the proposed planar motion aided robust visual localization system.

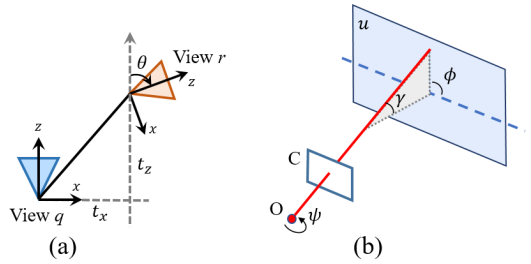


Fig. 3. (a) The illustration of planar motion between multiple query and reference views. (b) The geometric interpretation of affine decomposition.

end-to-end system performance of the visual localization. For the feature matching at the front-end, we simplify the affine distortion matrix with the motion property and propose a tilt augmentation strategy to suppress the variation caused by PC. Combined with the learning-based feature extraction, the proposed technique can increase a considerable number of matching inliers, as illustrated in Fig.1. For the pose estimation at the back-end, we use the minimal solution in our previous work [12] utilizing either map feature points with depth or ones without depth, thus taking full advantage of all correspondences. The framework of the proposed motion property aided visual localization system is shown in Fig. 2. Taking one query image as input, the augmentation-based feature matching is first applied to obtain matches, and the pose estimation algorithm is then performed in the RANSAC framework to solve the camera pose. In the following, we first introduce the planar motion property in indoor visual localization.

A. Planar motion property

Service robots generally move on planar surfaces, or at least locally planar surfaces. For planar motion shown in Fig. 3 (a), with the assumption that the camera plane is perpendicular to the ground, the pose from the coordinate system of reference view r to query view q can be expressed as

$$\mathbf{R} = \begin{bmatrix} cos\theta & 0 & sin\theta \\ 0 & 1 & 0 \\ -sin\theta & 0 & cos\theta \end{bmatrix}, \mathbf{t} = \begin{bmatrix} t_x \\ 0 \\ t_z \end{bmatrix} \tag{1}$$

where θ denotes the rotation angle around axis y and t_x, t_z denotes the translation along axis x and z. Note that the assumption is easily satisfied by rotating the appropriate pitch and roll angles from calibration.

Therefore the DoF of the transformation matrix under planar motion is reduced. Leveraging this property, the problem of indoor visual localization can be mathematically relaxed, and both the feature matching in the front-end and pose estimation in the back-end can be simplified. Specifically, in the front-end, to address the difficulty of feature matching caused by viewpoint difference due to camera motion, we reformulate the affine matrix by combining the planar motion characteristics, so as to achieve full affine invariance, as detailed in the next subsection.

B. Augmentation based feature matching

Affine model in planar motion: According to the camera imaging principle, the viewpoint variance between two images can be formulated by projective deformation, which can be further simplified to affine distortion if the subject is a smooth plane. Since the feature points in the indoor environment are mainly distributed on vertical walls, we next consider the affine model in detail. According to the affine theorem proposed in [11], any affine map $\mathbf{A}$ has a unique decomposition

$$\mathbf{A} = \lambda \begin{bmatrix} cos\psi & -sin\psi \\ sin\psi & cos\psi \end{bmatrix} \begin{bmatrix} \tau & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} cos\phi & -sin\phi \\ sin\phi & cos\phi \end{bmatrix} \tag{2}$$

where $\lambda > 0$, $\phi \in [0, \pi)$, and $\tau > 1$.

The geometric interpretation is illustrated in Fig. 3 (b), which shows the affine decomposition between the front view and a slanted view. The plane u is the object plane, C represents the camera image plane and O is the optical center. ϕ is called longitude angle which causes the image to rotate by the certain angle. γ is called latitude angle and $\tau = 1/cos\gamma$ which is called absolute tilt. The latitude angle results in apparent image deformation by resampling the image along the longitudinal direction, and so does the absolute tilt. λ denotes the distance between the camera and object plane along the optical axis, which is the scale factor. And ψ is induced when the camera rotates around its optical axis.

In indoor localization, the affine model can be simplified. Firstly, the rotation angle ψ can be eliminated as the camera is fixed on the robot. Then as illustrated in [32], there is a transformation relationship between the affine model and rotation matrix of the camera relative to the object. With the rotation matrix denoted as R and the entry in m row and n column as R_{mn}, the latitude and longitude angle can be expressed as:

$$\gamma = acos(R_{33}) \tag{3}$$

$$\phi = |atan(R_{31}/R_{32})| \tag{4}$$

In planar moving situation, the rotation matrix of the camera can be formulated by one rotation angle, which takes the form as in (1). By setting the coordinate system of the vertical flat object as the coordinate system of camera, and denoting the rotation angle around axis y as α, we have

$$\gamma = acos(cos\alpha) = \alpha \tag{5}$$

$$\phi = |atan(-sin\alpha/0)| = \pi/2 \tag{6}$$

Then the affine model in planar motion can be written as:

$$\mathbf{A} = \lambda \begin{bmatrix} \tau & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{7}$$

where $\tau = 1/cos\gamma = 1/cos\alpha$.

The above affine model illustrates the deformation between a front view and a slanted view. While in real world applications, the reference image may not be a front view facing the object plane. Therefore, we are going to derive the affine model between two slanted views. The affine model of the two cameras relative to the object plane is as follows:

$$\mathbf{A}_1 = \lambda_1 \begin{bmatrix} \tau_1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \mathbf{A}_2 = \lambda_2 \begin{bmatrix} \tau_2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{8}$$

Then the affine model between the two slanted views is

$$\mathbf{A}_{12} = \mathbf{A}_1 \mathbf{A}_2^{-1} = \frac{\lambda_1}{\lambda_2} \begin{bmatrix} \frac{\tau_1}{\tau_2} & 0 \\ 0 & 1 \end{bmatrix} \tag{9}$$

It's obvious that the transition tilt $\tau_{12} = \tau_1/\tau_2$ and transition rotation $\phi_{12} = 0$ from the expression of $\mathbf{A}_{12}$ in (9). Therefore, the affine model between two camera in planar motion can be expressed by one scale factor and one transition tilt.

Augmentation strategy: Lots of feature descriptors are fully scale-invariant such as SIFT [29], SuperPoint [26]. Therefore, we only focus on the tilt invariance strategy which can be then integrated with the scale-invariant descriptors to achieve full affine invariance. To achieve tilt invariance, the query image is augmented by applying all possible transition tilts and the obtained distorted query images are then matched with the reference image. Considering the translation tilt of the two tilted images in real world applications, we sample the latitude γ following the geometric series as: $1, \sqrt{2}, 2, 2\sqrt{2}, 4, 4\sqrt{2}$ to simulate different translation tilts for augmentation [11]. Finally, the feature correspondences obtained by matching all augmented query images are combined together after the inverse tilt operation, and the redundant matches are removed of which the Euclidean distance is less than 1 pixel. As the proposed augmentation-based feature matching algorithm is integrated with the planar motion property, it is more efficient and effective compared with the ASIFT [11].

C. Minimal solution

In this section, we briefly introduce the minimal solution namely 1P1DP proposed in our previous work [12], which utilizes one point correspondence with depth (DP) and one point correspondence without depth (P) to solve the camera pose by taking the planar property into consideration. Specifically, there are three unknown variables in robot pose with planar motion as shown in (1). According to projection geometry, two constraints about the pose can be derived using one 3D-2D correspondence. And one epipolar constraint can be derived using one 2D-2D correspondence. Then the pose estimation problem of the planar moving robot can be solved by combining the three constraints together. More details can be found in [12]. With the solution, all correspondences provided by the front-end can be utilized for pose estimation which enhances the robustness of the localization system as the sample set in RANSAC is maximal.

D. Map selection

As the proposed augmentation-based feature matching is fully affine invariant, we further propose a map selection guide to reduce the map storage. Specifically, the general indoor environment consists of some walls as the main structure, so we collect the reference images of the respective facing walls to construct the map. Then the transition tilt degenerates to the absolute tilt of query view $\tau_{12} = \tau_1$, as the reference view is a front view with $\tau_2 = 1$. This makes sense because the proposed method makes the map construction process simpler and easier to execute compared to existing methods of acquiring sequential image sequences for map building. And the capacity of the maps is substantially reduced for more devices equipped with low computing resources. Subsequent experimental results show that the strategy achieves smaller map storage without affecting localization performance and is suitable for robot platforms with limited computing resources.

IV. Experimental Results

For performance validation, we first conduct simulation experiments to evaluate i) the accuracy and robustness in feature matching of the proposed method and the performance in ii) homography estimation and iii) visual localization. And the computing efficiency is also discussed. In real world experiments, we use both public indoor localization dataset and self-collected dataset for validation. The experiments are performed on a desktop with CPU of Intel i7-7700 @ 3.60GHz $\times$ 8 and GPU of TITAN X $\times$ 4.

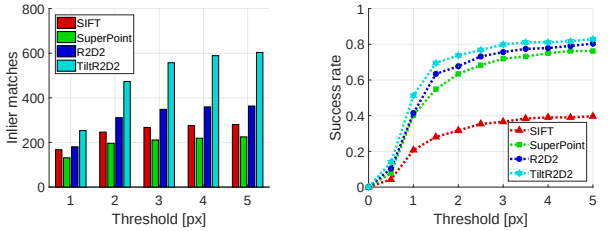
A. Simulation Experiments

In this section, we design an indoor simulation environment in Gazebo to satisfy the planar motion property.

Data setup: The simulation indoor environment is surrounded by four walls, and we randomly selected four images from the HPatches dataset [33] as textures to be attached to each of these four walls. A wheeled robot equipped with camera sensors is controlled in the environment for data collection. We first make the robot face the four walls to obtain four map reference images, and then control the robot to move randomly in the room to acquire query images. The resolution of the images is 752×480 and total of 164 query images are obtained for evaluation.

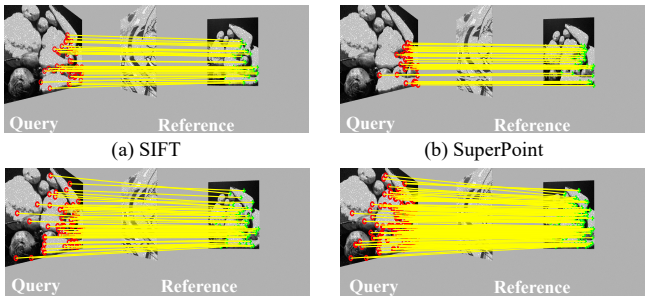
Feature matching: The well-known handcrafted descriptor SIFT [29] and learning-based descriptors SuperPoint [26] and R2D2 [27] are adopted for comparison. As the experimental results in our previous work [12] show that the matching performance of R2D2 is better than the other descriptors, we directly apply tilt augmentation on R2D2, and the obtained descriptor is denoted as TiltR2D2.

We match each query image with all four reference images and then select the one with the most feature matches as the result. For evaluation, the ground truth homography matrix between the query image and the reference image is computed using the known camera motion and plane parameters. Then the number of correct feature matches is counted for each image pair under the varying pixel error



(a) (b)

Fig. 4. (a) Comparison of feature matching accuracy. (b) Accuracy of homography estimation.



(a) SIFT (b) SuperPoint

(c) R2D2 (d) TiltR2D2

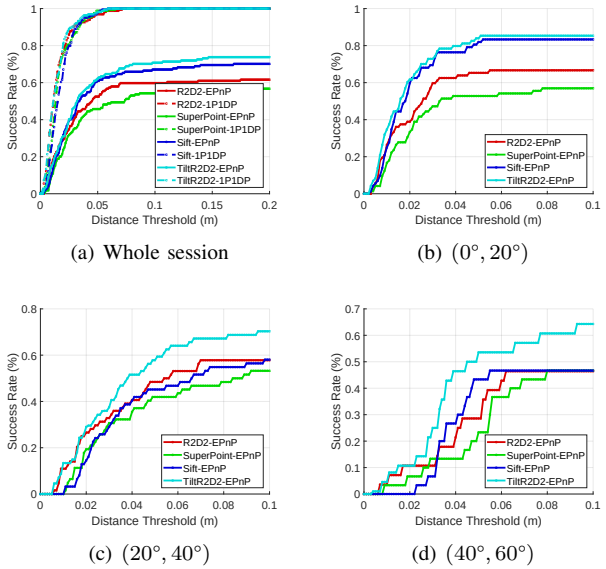
Fig. 5. Matching examples of simulation experiment. The shown matches are inliers with the error less than 1 pixel.

threshold, as in [11]. The result is shown in Fig. 4 (a). As can be seen from the result, thanks to the effective tilt augmentation strategy, the proposed TiltR2D2 achieves significantly more accurate feature matches at each threshold. Some feature matching cases are shown in Fig. 5 for better illustration.

Homography estimation: More accurate feature matching is not equivalent to more accurate geometric estimation, since the distribution of the matches is also essential. Therefore, we next conduct a homography estimation experiment. The corner correctness metric is utilized for evaluation which is computed by comparing the four corners on one image and the transformed corners on another using the estimated homography and the ground truth homography, respectively. The estimated homography is considered to be correct if the average error of the four corner points is less than a certain threshold. Then the success rate is counted as the homography estimation accuracy, following [26].

For all methods, RANSAC-based robust homography estimation is performed and the result is shown in Fig. 4 (b), from which we can find that the proposed TiltR2D2 achieves higher accuracy among all error thresholds, especially in the low threshold (1 pixel). The results demonstrate that the feature correspondences obtained by the proposed TiltR2D2 can be used to perform the better geometric estimation.

Localization performance: In the final, we run the complete visual localization system in the synthetic indoor environment mentioned before. As the plane parameters of each wall are known, the depth of each feature point can be computed by intersecting the projection ray with the plane,



(a) Whole session (b) $(0°, 20°)$

(c) $(20°, 40°)$ (d) $(40°, 60°)$

Fig. 6. (a): Localization performance on whole session. (b)-(d): Success rate comparison with increasing viewpoint difference.

such that the 2D-3D correspondences can be obtained. The success rate among different translation error thresholds of different algorithms is shown in Fig. 6 (a), and the rotation error threshold is fixed as 5 degree. For clear demonstration, we only show the estimation results of two solvers: the proposed 1P1DP and EPnP, which are tested with different feature matching methods. Results show that the proposed motion property aided localization system can achieve the best performance compared with other methods. And combined with the proposed solution, all feature matching methods can obtain great improvement, which validates the effectiveness of the proposed solution.

Furthermore, we group the query images according to the viewpoint difference with the reference image and test with EPnP for clear comparison. Results are shown in Fig. 6 (b-d), from which we can find that the advantage of the proposed tilt augmentation feature matching becomes more obvious when the viewpoint difference is bigger.

Runtime: We count the average computation time per simulation image in feature matching as follows: SIFT (0.08s), SuperPoint (0.04s), R2D2 (0.26s), TiltR2D2 (0.32s). Using parallel computing and batch processing technique of GPU, the computation time of the augmentation-based method increases little compared with the original one. The proposed method takes only 1.2 times longer by augmenting 5 images, which meets the real-time requirement.

B. Experiments on OpenLORIS

The Openloris dataset is collected by a wheeled robot equipped with a RealSense D435i. Challenging lifelong changes are presented in the dataset caused by illumination, viewpoint variations, and dynamic objects/human occlusions. And the dataset covers a variety of common scenarios in daily life, which is universal and representative for indoor localization testing.

TABLE I

PERFORMANCE COMPARISON ON SELECTED OPENLORIS.

		Small PC									Large PC								
	Changes	Case Num	SIFT		Superpoint		R2D2		TiltR2D2		Case Num	SIFT		Superpoint		R2D2		TiltR2D2	
			Inlier Num	Suc Rate	Inlier Num	Suc Rate	Inlier Num	Suc Rate	Inlier Num	Suc Rate		Inlier Num	Suc Rate	Inlier Num	Suc Rate	Inlier Num	Suc Rate	Inlier Num	Suc Rate
Home	Static	68	12.41	0.61	29.60	0.62	84.54	0.78	320.00	**0.79**	6	1.17	0.17	5.50	0.33	15.83	0.33	65.50	**0.50**
	Illumination	35	7.54	0.26	15.97	0.51	15.77	0.63	48.94	**0.70**	5	2.00	F	1.00	0.40	1.40	0.40	3.60	**0.60**
	Man-made	95	7.54	0.19	13.34	0.53	13.89	0.53	46.09	**0.55**	22	2.09	0.05	3.18	0.09	2.59	0.23	7.91	**0.41**
Corridor	Static	312	41.20	0.44	73.99	0.50	129.97	0.71	550.21	**0.73**	7	1.71	0.14	3.00	0.14	3.00	0.14	17.00	**0.29**
	Illumination	151	11.07	0.12	19.72	0.22	26.85	0.19	103.84	**0.25**	130	0.31	F	0.22	F	0.16	F	0.63	F
	Man-made	177	20.99	0.25	36.79	0.38	56.45	0.45	194.98	**0.45**	90	0.46	F	0.42	0.01	0.50	0.01	1.61	**0.02**
Office	Static	14	25.64	0.71	62.50	0.93	122.36	**1.00**	526.85	**1.00**	-	-	-	-	-	-	-	-	-
	Illumination	37	13.41	0.22	25.27	0.70	17.43	0.76	52.89	**0.89**	2	0	F	1.00	F	5.00	F	26.50	**0.50**
	Man-made	27	12.00	0.26	24.00	0.59	24.00	0.78	79.41	**0.81**	19	0.74	F	0.53	F	0.61	0.12	2.11	**0.21**
Cafe	Static	38	41.05	0.37	59.37	0.47	125.66	0.45	499.89	**0.53**	2	15.50	F	8.50	F	22.00	F	73.00	F
	Illumination	6	20.00	0.50	33.33	0.67	61.33	0.67	203.67	**0.67**	-	-	-	-	-	-	-	-	-
	Man-made	8	26.75	0.38	42.25	0.50	59.00	0.50	217.75	**0.63**	6	2.67	F	5.33	F	3.33	F	9.83	**0.33**

[1] F denotes failure localization.

TABLE II

SUCCESS RATE COMPARISON ON LOBBY DATASET.

Method	SIFT	SuperPoint	R2D2	TiltR2D2
m	0.25/0.50/1.0	0.25/0.50/1.0	0.25/0.50/1.0	0.25/0.50/1.0
deg	5/5/5	5/5/5	5/5/5	5/5/5
P3P	15.22/17.39/19.57	47.83/58.70/63.04	82.61/93.48/93.48	91.30/95.65/95.65
EPnP	4.35/4.35/4.35	43.48/50.00/52.17	76.09/82.61/82.61	80.43/91.30/91.30
AP3P	10.87/13.04/13.04	41.30/50.00/54.35	82.61/91.30/91.30	89.13/93.48/93.48
1P1DP	**60.87/67.39/67.39**	**60.87/78.26/78.26**	**84.78/93.48/93.48**	**95.65/97.83/97.83**

Data setup: As in [10], we first downsample each session so that the minimum distance between images is 1m or the minimum angular difference is 20 degree. Then the selected images can meet the requirements of covering a wide range of geographical locations and appearances while being representative enough. For each scene, there are multiple sessions collected in different time periods. To generate image pairs for evaluation, we select any two sessions for matching. Specifically, one of the sessions is selected as the reference map, and the nearest reference image is retrieved for each query image in the other to form an image pair. NetVLAD [15] is performed as image retrieval method.

Classification protocol: We extract feature correspondences using the baseline feature matching method SIFT and then perform two widely used pose estimation algorithms EPnP [34] and P3P [35] in RANSAC [21] framework for robust estimation. If both of the above methods localize successfully with translation error lower than 0.5m and rotation error smaller than 5 degree, the difficulty level of this image pair is easy. And this image pair is excluded as there is no point to exploring in an easy level. If localization fails, according to different PC, we categorize the image pairs into **large PC** category, if the view directional difference is larger than 25 degree, otherwise **small PC** category. We further classify the image pairs according to different changes to **static** if the query session and the reference session are the same one, and **EC** including illumination change and man-made change, as summarized in Table. I.

Evaluation measures: We define the count the average number of correct feature matches as *matching accuracy* for the image pairs under different categories. For evaluation, the ground truth fundamental matrix between the query image and reference image is computed using the ground truth camera pose and camera parameters. Then we find the correspondences that agree with the epipolar geometry between the image pair using Sampson distance as in [36]. We also report the success rate of localization to evaluate *localization robustness*. To get an accurate depth value for each matched feature point, we triangulate scene points using observation from multiple frames and refine the reconstruction with bundle adjustment. Then the translation error of the estimated relative pose and ground truth pose is computed as the Euclidean distance in meter. The rotation error is computed as $\arccos(0.5Tr(\mathbf{R}\mathbf{R}_{gt}^{T}) - 0.5)$ in degree, where $\mathbf{R}_{gt}$ denotes the ground truth relative rotation matrix and R denotes the estimated one. The localization is successful if the translation error is lower than 0.5m and rotation error is smaller than 5 degree as in [8].

Stratified evaluation: In this section, we perform 1P1DP as the pose estimation algorithm to test the performance of the whole system. Superpoint [26] and R2D2 [27] are selected as the representations of learning-based features which are shown to outperform traditional features and are tested in various real world datasets [12], [37], [38].

From the results in Table. I, we can find that the performance of the proposed system outperforms the others in all sessions with all categories, as the more inliers brought by the proposed method significantly promote the localization performance. This enhancement is more obvious in *office-illumination-largePC* and *cafe-man-made-largePC*, where the proposed method breaks through to successful localization when all other methods fail. Comparison of performance in EC and PC are summarized in Fig. 7. Results show that the improvement of the proposed method is significant under large PC, with 67% improvement with pure PC (Static+Large PC category) and even 120% improvement in condition with both EC and large PC. The proposed method also outperforms comparison methods under small PC, with an 8% improvement in condition with pure EC.

In addition, we can find that the improvement of learning-based feature matching is obvious compared to SIFT under small PC, especially in the presence of EC. However, with large PC, the performance of learning-based feature matching is basically no improvement over SIFT. In summary, using the learning-based approach to learn EC under small PC is significantly effective, but solving large viewpoint matching problems by learning is challenging. This is consistent with the results found in [10] and shows that the work in this

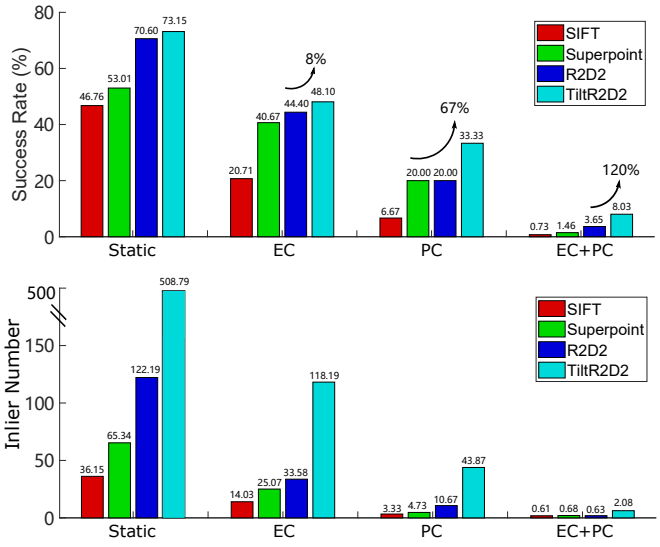
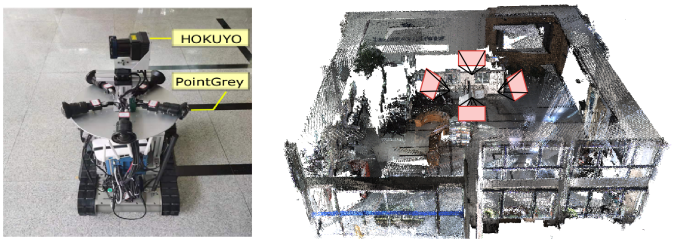


Fig. 7. Performance in dimensions of EC and PC.

TABLE III

SUCCESS RATE COMPARISON OF MAP SELECTION TEST.

Method m deg	SuperPoint-EPnP 0.25 / 0.50 / 1.0 5 / 5 / 5	TiltR2D2-EPnP 0.25 / 0.50 / 1.0 5 / 5 / 5	TiltR2D2-1P1DP 0.25 / 0.50 / 1.0 5 / 5 / 5
Selected 4 - map	36.67 / 43.33 / 46.67	90.00 / 93.33 / 93.33	**100.0 / 100.0 / 100.0**
Random 4 - map	30.00 / 33.33 / 33.33	56.67 / 63.33 / 63.33	63.33 / 70.00 / 70.00
Random 12 - map	50.00 / 63.33 / 63.33	93.33 / 93.33 / 93.33	96.67 / 100.0 / 100.0
Random 20 - map	63.33 / 70.00 / 70.00	96.67 / 96.67 / 96.67	100.0 / 100.0 / 100.0



(a) (b)

Fig. 8. (a) The platform used to collect Lobby dataset. (b) The reconstructed whole point cloud of the Lobby colored by projected pixels on images. The camera positions to get the selected map frames are shown.

paper does a good job of bridging this gap to promote the applications of service robots.

Full localization system performance: For complete system evaluation, we also show the localization performance of different solvers including EPnP [34], P3P [35], AP3P [39], and 1P1DP [12], and different feature matching methods on the whole OpenLORIS dataset. Results show that the proposed system achieves the best localization performance on all 12 sessions under all error thresholds, and the detailed success rate comparison can be found in Appendix-A[1]. Furthermore, with the proposed TiltR2D2 used as the front-end, all pose estimation algorithms are improved to different degrees. This enhancement is more apparent in the *cafe* session, where the distance from the robot to the wall is shorter and thus the change in viewpoint is more significant.

C. Experiments on Lobby

We also test the performance in a self-collected Lobby dataset. The dataset is collected by a mobile robot equipped with five pointgrey cameras and a 2D LiDAR as shown in Fig. 8 (a).

Data setup: We manually control the robot to move around in the lobby and acquire total 50 images. The resolution of each image is 1024×540. We obtain 3D scans by controlling the servo to rotate the 2D LiDAR for 360 degrees to improve the density of the acquired point cloud. The ground truth camera pose is obtained by aligning the dense synchronized 3D scans. And the 3D points of each image pixel are acquired by projecting the LiDAR scan into the image using the calibrated parameters.

Localization performance: We manually select 4 images facing the walls as the reference images as shown in Fig. 8(b) and the remaining 46 images as the query images for evaluation. Each query image is matched with these 4 reference images for the subsequent pose estimation. The localization performance is shown in Table II. Results show that the proposed system achieves the best performance, which is consistent with the results in Openloris. Furthermore, the results show that with the help of augmentation-based feature matching, the proposed system gets more robustness against viewpoint variances, such that four reference images facing the walls are sufficient for robust indoor localization.

Map selection test: To further demonstrate that the proposed localization system helps to achieve smaller map construction, we show the results with the different number of map images. To be specific, we first select 30 images as query images and randomly select 4, 8, 12, 16, and 20 images to construct the map among the remaining 20 images. We select the three most representative results on randomly selected map to show in Table III, and the detailed setting of the experiment and the full results can be found in Appendix-C[1]. Results with SuperPoint-EPnP show that the localization performance is enhanced with the increase of map size. However, the results with TiltR2D2-EPnP reflect that the performance on the selected map is similar to that of using 12 reference images and the promotion from a larger map is not significant. With TiltR2D2-1P1DP, the result of the selected map is the same as that using 20 images, which validates the effectiveness of the proposed system.

V. CONCLUSION

In this paper, we propose a complete indoor visual localization system that leverages the planar motion on both feature matching and pose estimation. Specifically, we find that the affine model can be simplified by taking planar motion into consideration and then propose an augmentation-based feature matching method to deal with large perspective variations. Combined with the motion property aided minimal solution in back end, a robust localization system is proposed and validated. Experimental results on the public dataset and self-collected dataset demonstrate the effectiveness of the proposed system in challenging long-term indoor localization. In the future, we are going to analyze strategies for eliminating affine distortion in outdoor localization.

TABLE IV

SUCCESS RATE COMPARISON ON WHOLE OPENLORIS DATASET.

Task	session	home2	home4	home5	corridor2	corridor3	corridor4
	m	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0
	degree	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 /5	5 / 5 / 5
R2D2	P3P	51.67 / 62.84 / 65.14	34.62 / 45.14 / 47.82	64.18 / 65.60 / 65.73	7.85 / 24.07 / 43.63	0.81 / 2.95 / 8.10	8.79 / 17.42 / 24.46
	EPnP	50.93 / 62.27 / 63.94	35.56 / 45.04 / 46.92	68.42 / 69.32 / 69.32	5.81 / 19.31 / 37.97	0.38 / 2.63 / 5.85	7.84 / 15.38 / 21.59
	AP3P	50.33 / 62.47 / 64.88	37.01 / 46.78 / 48.57	67.91 / 68.93 / 68.93	5.87 / 18.84 / 37.91	0.60 / 2.10 / 6.08	7.64 / 15.89 / 21.79
	1P1DP	**57.91 / 69.31 / 71.48**	**47.35 / 61.77 / 65.19**	**85.88 / 86.91 / 86.91**	**11.30 / 31.09 / 50.85**	**1.76 / 5.91 / 13.96**	**13.87 / 26.63 / 31.87**
TiltR2D2	P3P	52.90 / 66.74 / 69.55	37.11 / 49.18 / 52.14	64.06 / 65.60 / 65.73	10.04 / 27.61 / 44.26	1.19 / 4.29 / 9.48	15.14 / 21.23 / 24.83
	EPnP	51.80 / 64.51 / 67.21	36.36 / 46.92 / 49.55	69.51 / 70.31 / 70.31	10.58 / 27.64 / 43.63	0.81 / 3.33 / 7.86	15.30 / 19.38 / 21.97
	AP3P	52.77 / 66.64 / 69.38	36.40 / 49.04 / 52.14	69.46 / 71.26 / 71.26	10.01 / 27.41 / 43.80	1.19 / 4.10 / 9.29	14.88 / 20.86 / 24.35
	1P1DP	**60.34 / 72.11 / 74.28**	**49.98 / 62.66 / 68.15**	**87.03 / 87.80 / 87.80**	**15.59 / 35.17 / 52.98**	**1.76 / 7.19 / 16.39**	**18.32 / 26.79 / 31.97**

Task	session	corridor5	office3	office4	office5	office7	cafe1
	m	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0	0.25 / 0.5 / 1.0
	degree	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 /5	5 / 5 / 5
R2D2	P3P	29.97 / 50.25 / 68.43	56.11 / 56.11 / 56.11	70.92 / 71.15 / 71.15	81.94 / 83.51 / 83.51	87.03 / 88.96 / 88.96	30.50 / 62.06 / 80.15
	EPnP	30.08 / 50.59 / 69.92	55.56 / 55.56 / 55.56	71.15 / 71.61 / 71.61	81.18 / 82.82 / 82.88	87.03 / 88.87 / 88.87	31.79 / 62.65 / 80.68
	AP3P	29.88 / 50.39 / 69.19	56.11 / 56.11 / 56.11	70.92 / 71.15 / 71.26	81.91 / 83.53 / 83.53	83.70 / 85.49 / 85.49	30.21 / 62.47 / 79.98
	1P1DP	**34.70 / 56.69 / 74.69**	**58.33 / 58.61 / 58.61**	**74.14 / 76.44 / 76.44**	**89.36 / 90.25 / 90.25**	**90.97 / 92.11 / 92.11**	**39.99 / 72.54 / 82.32**
TiltR2D2	P3P	37.37 / 55.46 / 74.01	56.11 / 56.11 / 56.11	71.15 / 71.26 / 71.38	84.33 / 85.97 / 85.97	87.20 / 88.96 / 88.96	49.06 / 76.93 / 85.19
	EPnP	36.93 / 56.21 / 73.92	56.39 / 56.39 / 56.39	71.15 / 71.61 / 71.72	82.50 / 83.53 / 83.53	87.03 / 89.13 / 89.13	50.06 / 78.04 / 85.19
	AP3P	37.25 / 55.39 / 73.76	56.39 / 56.67 / 56.67	71.03 / 71.15 / 71.26	83.64 / 85.34 / 85.34	87.64 / 89.40 / 89.40	48.65 / 76.70 / 85.19
	1P1DP	**42.17 / 61.03 / 79.83**	**58.78 / 58.89 / 59.17**	**74.71 / 76.67 / 77.24**	**90.12 / 91.50 / 91.50**	**91.24 / 92.55 / 92.55**	**54.22 / 81.03 / 85.66**

[1] Considering the map coverage of the environment, the map sequences for each session are as follows: home1 and home3, corridor1, office1 and office2, cafe2.
[2] For testing on office6, all methods achieve 100 success rate over three thresholds.

TABLE V

SUCCESS RATE COMPARISON OF MAP SELECTION TEST.

Method	SuperPoint-EPnP	TiltR2D2-EPnP	TiltR2D2-1P1DP
m	0.25 / 0.50 / 1.0	0.25 / 0.50 / 1.0	0.25 / 0.50 / 1.0
deg	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5
Selected 4 - map	36.67 / 43.33 / 46.67	90.00 / 93.33 / 93.33	**100.0 / 100.0 / 100.0**
Random 4 - map	30.00 / 33.33 / 33.33	56.67 / 63.33 / 63.33	63.33 / 70.00 / 70.00
Random 8 - map	43.33 / 43.33 / 43.33	76.67 / 76.67 / 76.67	83.33 / 86.67 / 90.00
Random 12 - map	50.00 / 63.33 / 63.33	93.33 / 93.33 / 93.33	96.67 / 100.0 / 100.0
Random 16 - map	60.00 / 60.00 / 60.00	96.67 / 96.67 / 96.67	100.0 / 100.0 / 100.0
Random 20 - map	63.33 / 70.00 / 70.00	96.67 / 96.67 / 96.67	100.0 / 100.0 / 100.0

APPENDIX

A. Localization performance on whole OpenLORIS

For complete system evaluation, we show the localization performance of different solvers and feature matching methods on the whole OpenLORIS dataset. The success rate under different error thresholds is shown in Table IV.

B. Details of map selection test

To further demonstrate that the proposed localization system helps to achieve smaller map construction, we show the results with the different number of map images. To be specific, we first select 30 images as query images and randomly select 4, 8, 12, 16, and 20 images to construct the map among the remaining 20 images. Then the localization results are compared with that of the selected map which consists of four images facing walls. When testing on different maps, we choose the top 4 reference images with the highest number of feature matches for pose estimation. The results of all randomly selected map are shown in Table V.

REFERENCES

[1] Z. Chen, J. Guo, X. Xu, Y. Wang, H. Huang, Y. Wang, and R. Xiong, "Pregan: Pose randomization and estimation for weakly paired image style translation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2209–2216, 2021.

[2] S. Garg, B. Harwood, G. Anand, and M. Milford, "Delta descriptors: Change-based place representation for robust visual localization," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5120–5127, 2020.

[3] H. Huang, H. Ye, Y. Sun, and M. Liu, "Gmmloc: Structure consistent visual localization with gaussian mixture models," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5043–5050, 2020.

[4] L. Tang, Y. Wang, X. Ding, H. Yin, R. Xiong, and S. Huang, "Topological local-metric framework for mobile robots navigation: a long term perspective," *Autonomous Robots*, vol. 43, no. 1, pp. 197–211, 2019.

[5] X. Ding, Y. Wang, R. Xiong, D. Li, L. Tang, H. Yin, and L. Zhao, "Persistent stereo visual localization on cross-modal invariant map," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[6] P.-E. Sarlin, C. Cadena, R. Siegwart, and 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. 12 716–12 725.

[7] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8601–8610.

[8] Y. Jiao, Y. Wang, X. Ding, B. Fu, S. Huang, and R. Xiong, "2-entity ransac for robust visual localization: Framework, methods and verifications," *IEEE Transactions on Industrial Electronics*, 2020.

[9] C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, "A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence," *arXiv preprint arXiv:2006.12567*, 2020.

[10] A. Jafarzadeh, M. L. Antequera, P. Gargallo, Y. Kuang, C. Toft, F. Kahl, and T. Sattler, "Crowddriven: A new challenging dataset for outdoor visual localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9845–9855.

[11] J.-M. Morel and G. Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM journal on imaging sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[12] Y. Jiao, L. Liu, B. Fu, X. Ding, M. Wang, Y. Wang, and R. Xiong, "Robust localization for planar moving robot in changing environment: A perspective on density of correspondence and depth," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4006–4012.

[13] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[14] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a

versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[15] 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*, 2016, pp. 5297–5307.

[16] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart, "Get out of my lab: Large-scale, real-time visual-inertial localization." in *Robotics: Science and Systems*, vol. 1, 2015.

[17] T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys, "Hyperpoints and fine vocabularies for large-scale location recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2102–2110.

[18] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1455–1461, 2016.

[19] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6896–6906.

[20] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic match consistency for long-term visual localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 383–399.

[21] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[22] Z. Kukelova, M. Bujnak, and T. Pajdla, "Closed-form solutions to minimal absolute pose problems with known vertical direction," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 216–229.

[23] B. Guan, J. Zhao, Z. Li, F. Sun, and F. Fraundorfer, "Minimal solutions for relative pose with a single affine correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1929–1938.

[24] P. Speciale, D. Pani Paudel, M. R. Oswald, T. Kroeger, L. Van Gool, and M. Pollefeys, "Consensus maximization with linear matrix inequality constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4941–4949.

[25] Y. Jiao, Y. Wang, X. Ding, M. Wang, and R. Xiong, "Deterministic optimality for robust vehicle localization using visual measurements," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[26] 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*, 2018, pp. 224–236.

[27] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: repeatable and reliable detector and descriptor," *arXiv preprint arXiv:1906.06195*, 2019.

[28] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and 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.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[31] I. Rocco, M. Cimpoi, R. Arandjelovic, A. Torii, T. Pajdla, and J. Sivic, "Ncnet: Neighbourhood consensus networks for estimating image correspondences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[32] Q. Wang, W. Zhang, X. Liu, Z. Zhang, M. H. A. Baig, G. Wang, L. He, and T. Cui, "Line matching of wide baseline images in an affine projection space," *International Journal of Remote Sensing*, vol. 41, no. 2, pp. 632–654, 2020.

[33] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5173–5182.

[34] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[35] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[36] Q. Zhou, T. Sattler, and L. Leal-Taixe, "Patch2pix: Epipolar-guided pixel-level correspondences," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4669–4678.

[37] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[38] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 23–79, 2021.

[39] T. Ke and S. I. Roumeliotis, "An efficient algebraic solution to the perspective-three-point problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7225–7233.