2019 IEEE/CVF International Conference on Computer Vision (ICCV)

Prior Guided Dropout for Robust Visual Localization in Dynamic Environments

Zhaoyang Huang[1,2] Yan Xu[2] Jianping Shi[2] Xiaowei Zhou[1] Hujun Bao[1*] Guofeng Zhang[1*]

[1]State Key Lab of CAD&CG, Zhejiang University[†]
[2]SenseTime Research

Abstract

*Camera localization from monocular images has been a long-standing problem, but its robustness in dynamic environments is still not adequately addressed. Compared with classic geometric approaches, modern CNN-based methods (*e.g. *PoseNet) have manifested the reliability against illumination or viewpoint variations, but they still have the following limitations. First, foreground moving objects are not explicitly handled, which results in poor performance and instability in dynamic environments. Second, the output for each image is a point estimate without uncertainty quantification. In this paper, we propose a framework which can be generally applied to existing CNN-based pose regressors to improve their robustness in dynamic environments. The key idea is a prior guided dropout module coupled with a self-attention module which can guide CNNs to ignore foreground objects during both training and inference. Additionally, the dropout module enables the pose regressor to output multiple hypotheses from which the uncertainty of pose estimates can be quantified and leveraged in the following uncertainty-aware pose graph optimization to improve the robustness further. We achieve an average accuracy of 9.98m/3.63° on RobotCar dataset, which outperforms the state-of-the-art method by 62.97%/47.08%. The source code of our implementation is available at https://github.com/zju3dv/RVL-Dynamic.*

1. Introduction

Localization is a fundamental problem in many applications including robotics, AR/VR, autonomous driving, *etc*. A typical localization scenario is that a robot equipped by sensors, *e.g*., camera, lidar, *etc*., is locating itself in a large-scale urban scene. Visual localization only requires a camera and has drawn increasing attention because of its low cost and broad applicability compared to other localization techniques. Traditional geometric localization methods [32, 14, 13] mainly use handcrafted features and descriptors and are rather sensitive to illumination variation, viewpoint change, and dynamic elements, which are common in unconstrained environments.

Figure 1. Example images from the RobotCar dataset. Foreground objects are different in training and test images, which introduces bias when learning a camera pose regressor and leads to unstable localization.

Recently, convolutional neural networks (CNNs) have shown outstanding performance in object and place recognition, which also motivates researchers to exploit the potential of CNNs in visual localization. Many endeavors have been made to address the limitations of traditional methods, which will be discussed in Sec. 2. Among these explorations, PoseNet [28] is a pioneering work that leverages CNNs originally designed for object recognition to solve camera pose regression, which has validated the feasibility of visual localization through end-to-end neural networks. We refer to this kind of visual localization methods as neural pose regressors since they directly regress 6-DOF camera poses from images via a neural network.

Currently, most visual localization methods assume that the environment is static, which is obviously not true in practical scenarios. While neural pose regressors didn't explicitly make this assumption, foreground moving objects would inevitably degrade their accuracy and reliability as shown in Fig. 1. To the best of our knowledge, handling the dynamic objects for neural pose regressor has not been

*Corresponding authors: {bao, zhangguofeng}@cad.zju.edu.cn

[†]The authors from State Key Lab of CAD&CG, Zhejiang University are also affiliated with ZJU-SenseTime Joint Lab of 3D Vision. This work was partially supported by NSF of China (Nos. 61822310 and 61672457), and the Fundamental Research Funds for the Central Universities (Nos. 2018FZA5011 and 2019XZZX004-09).

2380-7504/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCV.2019.00288

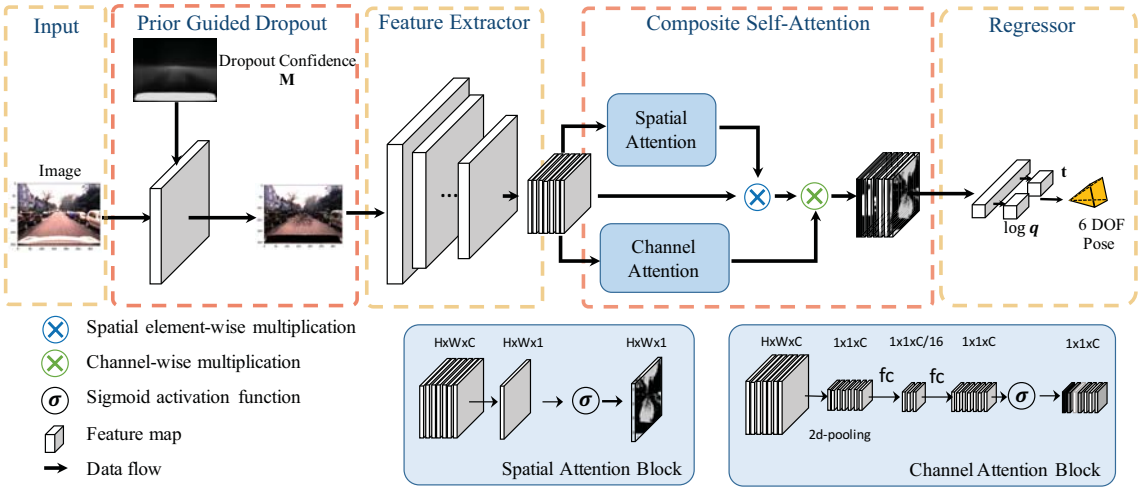

Figure 2. The proposed framework. Modern neural pose regressors [4, 28] generally comprises a feature extractor and a regressor (yellow boxes). We propose a prior guided dropout module and a composite self-attention module (red boxes), which can be universally embedded into the architecture of existing neural pose regressors, to alleviate the adverse impact from the unexpected movable objects in training and test phases. Before feature extraction, the input pixels are randomly discarded by the dropout module based on a prior probability obtained from object segmentation in training. After the feature extraction, the self-attention module reweights the extracted feature maps, *i.e.*, a 2-D attention weight map, and a 1-D attention weight vector, to filter out the misleading features. Finally, the pose regressor predicts the 6-DOF camera poses from the re-weighted feature maps.

thoroughly discussed in literature. An intuitive approach is to detect and subtract foreground objects from images before feeding images into CNNs, but our empirical results suggested that this method gave a poor performance as the subtraction resulted in salient image patterns (*e.g.*, sharp edges) that would affect the learning process.

In this paper, we propose a general framework to improve the robustness of neural pose regressors in dynamic environments with two novel modules as shown in Fig. 2. The prior guided dropout module randomly drops pixels based on a specified prior distribution obtained from an object segmentation method and the self-attention module reweights the extracted feature maps. With these two modules, the neural pose regressor can be guided to capture essential features for localization and the impact from moving objects can be alleviated. Besides achieving high accuracy, a robust localization should provide confidence levels for predicted results. In regression problems, bootstrapping [12, 18] is often utilized for asymptotic distribution estimation. Similarly, the distribution of poses predicted by a neural network is also hard to track analytically, so we propose to approximate the pose distribution with multiple hypotheses generated by prior guided dropout, and further improve the robustness by leveraging the pose distribution (mean and variance) in uncertainty-aware pose graph optimization.

Our main contributions are summarized as follows:

- We propose a prior guided dropout module and a composite self-attention module that can be naturally applied to existing neural pose regressors and guide the networks to ignore distracting information from foreground objects and focus on essential landmarks in the background for robust localization.
- We propose to quantify the uncertainty of pose estimation from multiple hypotheses given by the proposed dropout method and feed the uncertainty measures into uncertainty-aware pose-graph optimization to further improve the robustness of pose estimation.
- We report the state-of-the-art results on the challenging RobotCar dataset and outperform the existing methods by a large margin.

2. Related Work

Given an image or a video, a visual localization system attempts to compute the location of the camera relative to some representation of the environment, which is often called a map. According to different goals, visual localization methods can be categorized into topological localization and metric localization.

Topological localization, which is also called place recognition, aims to find whether the place or location where the query image is taken has been visited before [30]. Traditionally, topological localization represents the map as a bunch of images. Then, the problem is formulated as image retrieval where the query image is matched to database images based on image descriptors such as BoW (Bag-of-Words) [16], VLAD [23, 9], and Fisher vector [24]. In this framework, some efforts have been made to improve accuracy and efficiency. For example, FAB-MAP 2.0 [7, 8] uses an inverted index with BoW model and probabilistic inference. Schindler *et al*. [37] accelerates retrieving speed via hierarchical vocabulary tree [34].

Metric localization targets to figure out the metric position and orientation of cameras. A notable solution for this problem is visual simultaneous localization and mapping (VSLAM) [32, 14, 13], which is able to simultaneously build the map and localize the camera by consecutively estimating relative pose transformation between image pairs. The map in VSLAM contains a number of 3D landmarks such as points, edges and planes, and the 6-DOF camera pose can be computed by feature matching between 2D points in images and 3D points in the map.

While remarkable results have been achieved, visual localization in dynamic environments is still very challenging due to appearance variations and moving objects. Some works proposed more robust local features such as SIFT [29] and edge features [11, 22], which are invariant to lighting, orientation and scale in certain circumstances, but these handcrafted descriptors still have limited representability and robustness. In recent years, CNN has shown its ability to learn more powerful features or representations [35, 40]. In topological localization, Chen *et al*. [5] was among the first that replaced handcrafted descriptors [29, 36, 2] with CNN features. In metric localization, PoseNet [28] proposed to memorize the environment as parameters in a neural network and regress absolute 6-DOF camera poses from single images. Since then, a bunch of works has appeared to improve the localization accuracy for CNN-based pose regressors. For example, Kendall *et al*. [26] introduced a dropout procedure into PoseNet to measure the model's uncertainty. Clark *et al*. [6] proposed to localize the cameras in a video sequence by bidirectional LSTM (long short term memory) [38]. Kendall *et al*. [27] managed to learn an optimal loss weight to regress position and orientation simultaneously. Tayyab *et al*. [33] exploited a 3D space data augmentation to reduce insufficiency of labeled data in wild scenes. Recently, Abhinav *et al*. [41] proposed to learn visual odometry simultaneously and localization and Brahmbhatt *et al*. [4] utilized relative camera poses as an extra supervision signal along with constraints from unlabeled data.

The existence of movable objects is another challenge in dynamic environments. Some efforts were devoted to eliminating the influence of dynamic objects. For example, Johns and Yang [25] and Hafez *et al*. [17] used the BoW model and filtered out useless features determined by feature distinctiveness and feature reliability. Wang *et al*. [43] and Dong *et al*. [10] proposed to detect moving objects to obviate their disturbance. Yin and Shi [44] reasoned about static and dynamic parts of a scene separately to compute relative pose between image pairs. Vijayanarasimhan *et al*. [42] segmented the moving objects in a dynamic scene for motion estimation. To the best of our knowledge, there is no prior work taking dynamic objects into account for a neural pose regressor. Similar to [43, 10, 25], we aim to reduce the influence of movable objects. The difference is that it is easy to downweight object pixels in optimization based methods but not straightforward in neural networks, so we propose a prior guided dropout method to realize it.

3. Proposed Method

Recent works have employed CNNs to learn localization and mapping implicitly, but they do not address the dynamic objects issue in the scene and inevitably model them as part of the map, which dramatically degrades the localization accuracy and robustness. In this paper, we propose a framework incorporating three essential components, *i.e*., prior guided dropout, composite self-attention and pose refinement with uncertainty-aware PGO (pose graph optimization), to distill the reliable landmarks and filter out the interferences from movable objects. Fig. 2 gives an overview of our framework. Sec. 3.1 elaborates the prior distribution modeling and prior guided dropout, which provides a priority guiding the network to concentrate on the valuable landmarks. Then, a learnable attention module (Sec. 3.2) is embeded after the feature extractor endowing the model with the ability to select faithful features and sense spatial differences. Finally, with a set of 6-DOF pose predictions generated by the regressor, we further apply an uncertainty-aware PGO to refine the poses for the whole sequence, which will be discussed in Sec. 3.3.

3.1. Prior Guided Dropout

Previous works of neural pose regressor [28, 4], tend to make inaccurate predictions when dynamic objects (*e.g*., pedestrians) exist in the training or test frames, suggesting that moving objects are actually noisy features that contaminate the data. As illustrated in Fig. 1, since the movable vehicles occupy a significant portion in the view and always remain in an image sequence, the network would regard them as landmarks easily if no special supervision is provided. We argue that neural pose regressors should be guided to pay more attention to invariant features of the scene and downweight the contaminated features.

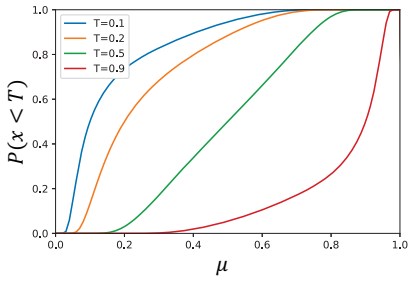

Figure 3. The CDF of dropout probability with different threshold T. Larger μ decreases $P(x < T; \mu)$ under the same T, which means the pixels with higher probability belonging to movable objects will be discarded more likely. All the pixels will be discarded if $T = 0$ and all the pixels will be preserved if $T = 1$.

Dropout [39, 20] is a common strategy to mitigate overfitting where the dropout probability specified to a feature can be viewed as an importance weighting. On the other hand, deep neural network is a mapping where dropout can be treated as Bayesian approximation for posterior estimation [15]. To guide neural pose regressors to concentrate on valuable features and evaluate the posterior pose distribution, we add a prior guided dropout module described in Alg. 1 at the beginning of the neural pose regressor. More specifically, the dropout module generates a random number x for each pixel in the input image and set the pixel value as zero if x is smaller than a predefined threshold T. The random number is generated from a Gaussian distribution whose mean and variance are spatially-varying which depend on the frequency of being occupied by foreground objects in training images. This prior guided dropout module both improves the robustness of the model and provides an uncertainty measurement for each predicted pose.

Informative prior can benefit both regression and Bayesian inference. In our dropout module, the prior is represented by the Gaussian parameters which are in charge of the dropout probability, so we also propose a statistical strategy to compute the parameters. Firstly, we apply an off-the-shelf segmentation approach, *i.e.*, Mask R-CNN [19], to generate a binary mask $\mathbf{m}_k$ for each training image $\mathbf{I}_k$ where $\mathbf{m}_k(i,j) = 1$ if pixel (i,j) belongs to a foreground object, otherwise $\mathbf{m}_k(i,j) = 0$. Then, the parameters for pixel (i,j) are calculated from the mask $\mathbf{m}_k$ by

$$\begin{cases} \boldsymbol{\mu}(i,j) = \frac{1}{n}\Sigma_{k=1}^{n}\mathbf{m}_k(i,j) \\ \boldsymbol{\sigma}^2(i,j) = \boldsymbol{\mu}(i,j)(1-\boldsymbol{\mu}(i,j)) \end{cases} \tag{1}$$

where n denotes the numer of training images.

A problem is that the prior may be biased, *e.g.*, false detection or leak detection of detectors, so we set a threshold T to smoothly moves the overall probability for bias compensation. Theoretically, the threshold T controls the shape

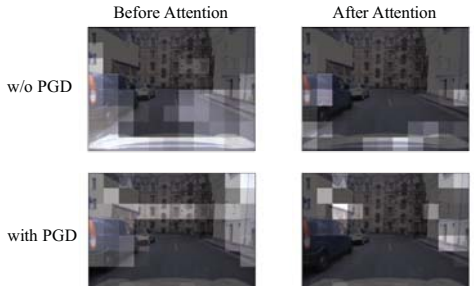

Figure 4. Activation maps before and after the self-attention module. With PGD (prior guided dropout), the feature map before the attention module highlights areas in the image occupied by a bus, while the feature map after that concentrates on the door and the walls, indicating that our proposed self-attention module successfully grasped the essential features. The feature map without PGD still highlights the vehicles, which illustrates the importance of PGD.

of CDF (cumulative distribution function) of dropout probability as illustrated in Fig. 3 deduced from the Q-function:

$$P(x < T; \mu) = F(T; \mu, \sigma) = 1 - Q(\frac{T-\mu}{\sigma}) \tag{2}$$

where F denotes the CDF of Gaussian distribution.

Algorithm 1 Prior Guided Dropout

INPUT:
$I(i,j)$: Image intensity of pixel (i,j)
$\boldsymbol{\mu}(i,j), \boldsymbol{\sigma}^2(i,j)$: Prior Gaussian distribution parameters
T: Predefined threshold
OUTPUT:
$O(i,j)$: Intensity assigned to (i,j) in an output image

```
for all (i,j) do
    x ←sample from N(μ(i,j), σ²(i,j))
    if x < T then
        O(i,j) = 0
    else
        O(i,j) = I(i,j)
    end if
end for
```

3.2. Composite Self-Attention Module

Recently, the network architecture of neural pose regressors mainly contains two stages, *i.e.*, a feature extractor and a global average pooling followed by a fully connected regressor. These works take advantage of transfer learning from image classification tasks, but apart from the benefit, neural pose regressor should be more robust to noises and perceive spatial information, which is obfuscated by global pooling. Considering these two factors, we propose

to embed a composite self-attention module containing a spatial attention block and a channel attention block before the global average pooling layer.

In the spatial attention block, as illustrated in Fig. 1, the network spontaneously learns a 1-channel spatial weight map from the input feature maps and then re-weight the input features spatially to generate the final output. This mechanism endows the model with the ability to automatically choose the valuable locations to focus on. Moreover, inspired by the work of Hu *et al*. [21], we employ the SE block in the channel attention module which enables the model to filter out worthless feature channels. Fig. 4 presents a visually enhanced comparison to show the feature maps before/after self-attention model with/without the prior guided dropout suggesting that prior guided dropout help self-attention map achieve a stronger response in reasonable regions, whereas the raw one focus on the movable bus. Consequently, the composite self-attention module further enhances the model's capability in distilling the reliable features from a hodgepodge. Exhaustive experiments demonstrate that this scheme boosts the convergence speed and prediction accuracy especially in the environments with piles of movable objects.

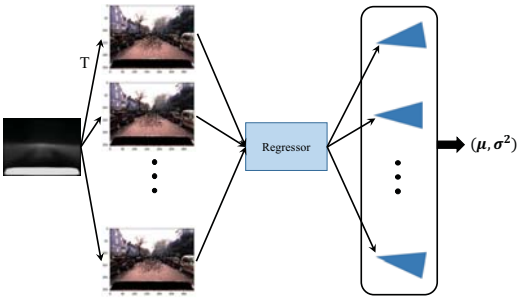

Figure 5. Pose distribution estimation. We generate multiple hypotheses for each input image and estimate mean and variance.

3.3. Uncertainty-Aware PGO

Although we trained a model more robust to noises, instability caused by moving objects yet can not be avoided in dynamic environments. Therefore, a robust visual localization system should provide not only camera poses, but also measurements of the uncertainty, so we propose to evaluate asymptotic pose distribution via prior guided dropout (Fig. 5).

The localization results provided by a neural pose regressor keep consistency in long term but fails occasionally, while the relative poses given by visual odometry (VO) are reliable between image pairs but will be up against drift with image sequence growing, so we propose to fuse the poses with PGO by minimizing the following energy function:

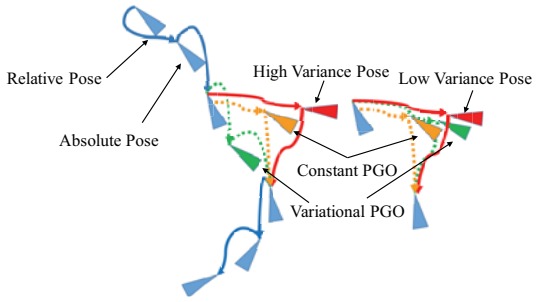

Figure 6. We present the pose sequence comprised of relative poses from visual odometry and absolute poses from neural pose regressor. Constant PGO assigns equivalent weight to each pose while the weight of absolute pose in variational PGO is given by reciprocal of estimated variance during optimization. Therefore, Variational PGO can benefit accuracy when variance is estimated accurately.

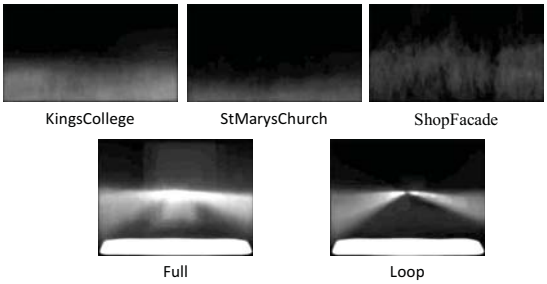

Figure 7. We present the μ parameter images of the five scenes. Movable objects always appear in the bottom of images in King's College and St Mary's Church, while in the middle of images in ShopFacade more frequently. In RobotCar dataset, movable objects show up on both sides of the road or in front of the car more frequently. King's College, full, and loop provide stronger prior.

$$\begin{aligned} E(\mathbf{p}^*, \boldsymbol{\theta}^*) \quad &= \Sigma_{i=1}^{N} w_i^p L_p(\mathbf{p}_i^*, \mathbf{p}_i) + w_i^\theta L_\theta(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i) \\ &+ \Sigma_{i=1}^{N-1} w_i^t L_t(\mathbf{v}_i^*, \mathbf{t}_i) + w_i^\delta L_\delta(\mathbf{r}_i^*, \boldsymbol{\delta}_i) \end{aligned} \tag{3}$$

$$\begin{cases} L_p(\mathbf{p}_i^*, \mathbf{p}_i) = \|\mathbf{p}_i^* - \mathbf{p}_i\|^2 \\ L_\theta(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i) = \|\log(R(\boldsymbol{\theta}_i^*)^T R(\boldsymbol{\theta}_i))\|^2 \\ L_t(\mathbf{v}_i^*, \mathbf{t}_i) = \|\mathbf{v}_i^* - \mathbf{t}_i\|^2 \\ L_\delta(\mathbf{r}_i^*, \boldsymbol{\delta}_i) = \|\mathbf{r}_i^* - \boldsymbol{\delta}_i\|^2 \\ \mathbf{v}_i^* = R(\boldsymbol{\theta}_i)^T(\mathbf{p}_{i+1}^* - \mathbf{p}_i^*) \\ \mathbf{r}_i^* = \log(R(\boldsymbol{\theta}_{i+1}^*)^T R(\boldsymbol{\theta}_i^*)) \end{cases} \tag{4}$$

where i denotes the index of image, $\mathbf{p}_i^*$ and $\boldsymbol{\theta}_i^*$ are absolute position and orientation to be optimized, $\mathbf{v}_i$ and $\mathbf{r}_i$ are relative translation and rotation calculated from $\mathbf{p}_i^*$, $\mathbf{p}_{i+1}^*$, $\boldsymbol{\theta}_i^*$ and $\boldsymbol{\theta}_{i+1}^*$. $\mathbf{p}_i$, $\boldsymbol{\theta}_i$ are the absolute position and orientation predicted by the neural pose regressor. $\mathbf{t}_i$, $\boldsymbol{\delta}_i$ represent the relative translation and rotation between pose of i and $i+1$

Scene	PoseNet	A-PoseNet	D-PoseNet	AD-PoseNet	AD-PoseNet+CPGO	AD-PoseNet+VPGO
Full	46.61m,10.45°	62.46m,11.95°	38.56m,10.45°	33.82m,6.77°	**27.35**m,6.88°	27.37m,**6.18°**
Loop	7.90m,3.53°	12.55m,4.63°	7.57m,3.61°	6.40m,3.09°	7.04m,3.03°	**6.49**m,**2.80°**
Average	27.26m,6.99°	37.51m,8.29°	23.07m,7.09°	20.11m,4.93°	17.20m,4.96°	**16.93**m, **4.49°**
Scene	MapNet	A-MapNet	D-MapNet	AD-MapNet	AD-MapNet+CPGO	AD-MapNet+VPGO
Full	44.61m,10.38°	30.02m,6.97°	32.64m,10.07°	19.18m,4.60°	18.84m,13.73°	**14.85**m,**4.30°**
Loop	9.29m,3.34°	8.41m,3.41°	9.72m,3.77°	6.45m,2.98°	6.37m,3.12°	**5.10**m,**2.96°**
Average	26.95m,6.86°	19.22m,5.19°	21.18m,6.92°	12.82m,3.79°	12.61m,8.43°	**9.98**m,**3.63°**

Table 1. Ablation study on RobotCar dataset. PoseNet [27] and MapNet [4] are used as baseline models. A and D denotes composite self-attention and prior guided dropout. As shown in the table, composite self-attention module alone achieves minor improvement (A-MapNet) and may even result in overfitting (A-PoseNet, Fig. 8), and prior guided dropout alone brings certain improvement (D-PoseNet, D-MapNet), but the cooperation of these two modules boosts the performance (AD-PoseNet, AD-MapNet). Moreover, our VPGO (Variational PGO) algorithm further raises the accuracy. By comparing with CPGO (Constant PGO) algorithm, it validates the effectiveness of both estimated variance and VPGO algorithm. Note that the whole framework we proposed (AD-PoseNet+VPGO and AD-MapNet+VPGO) separately outperforms the baseline (PoseNet and MapNet) by 37.89%/35.77% and 62.97%/47.08%.

Scene	Spatial Extent	PoseNet	AD-PoseNet	Bayesian PoseNet	Dense PoseNet	Dist. to NN
King's College	140×40m	1.61m,2.95°	**1.30**m, **1.67°**	1.74m,4.06°	3.34m,5.92°	1.66m,4.86°
St Mary's Church	80×60m	2.14m,5.06°	2.28m,**4.80°**	**2.11**m,8.38°	2.45m,7.96°	4.48m,11.3°
ShopFacade	35×25m	1.55m,**4.64°**	**1.22**m,6.17°	1.25m,7.54°	1.41m,7.18°	2.10m,10.4°
Average		1.77m,4.22°	**1.60**m, **4.21°**	1.70m,6.66°	1.84m,6.67°	2.74m,8.85°

Table 2. Comparison with related works on Cambridge Landmarks. PoseNet [27], Bayesian PoseNet [26] and Dense PoseNet [28] are state of the art neural pose regressor. Dist. to NN is a topological localization method introduced in [28]. Out framework improves PoseNet in both position and orientation and achieves state-of-the-art performance. The performance in the King's College sequence has been improved a lot compared to the others because its prior provides more information (see Fig. 7).

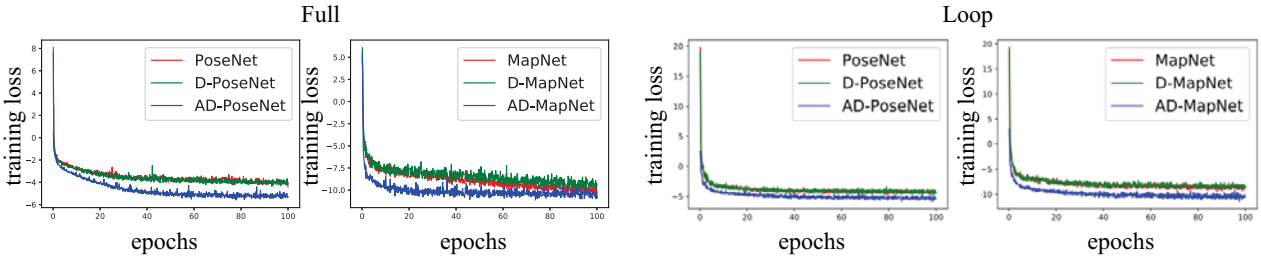

Figure 8. Training loss comparison. D-PoseNet equips PoseNet with prior guided dropout and AD-PoseNet further insert composite self-attention module into D-PoseNet. The composite self-attention module facilitates both PoseNet and MapNet to achieve lower training loss and converge faster.

provided by VO. L_p, L_θ, L_t and L_δ are loss functions, w_i^p, w_i^θ, w_i^t and w_i^δ are loss weights, R maps $\boldsymbol{\theta}$ to rotation matrix and log maps the lie group $SO(3)$ into its tangent space, *i.e.*, lie algebra $so(3)$.

Brahmbatt *et al.* [4] introduced a constant PGO method, which uses equivalent weights to smooth poses in temporal sequence. A critical defect of this algorithm is that a well-predicted pose will be erroneously dragged. By contrast, we propose a variational PGO which assigns reciprocal of estimated variances to loss weights of absolute poses, *i.e.*, w_i^p and w_i^θ. As illustrated in Fig. 6, absolute poses with high variance will be dragged easily by relative poses while low variance assist them in standing still in variational PGO, so good measurements of variance can further improve the accuracy of predicted poses.

4. Experiments

In this section, we first introduce the datasets we experiment on, and then exhibit a thorough ablation study to evaluate each proposed components. The comparison with the state-of-the-art is demonstrated finally. Some implementation details and more experimental results are provided in supplementary materials.

4.1. Datasets

We evaluate our method on two publicly available datasets: Cambridge Landmarks [28] and Oxford RobotCar [31], and to keep consistent with previous works, we compute median error on Cambridge Landmarks and mean error on Oxford RobotCar. The Cambridge Landmarks contains several short image sequences captured in differ-

Scene	DSAC	ORB-SLAM2	DBoW3	Stereo VO	PoseNet	MapNet	AD-MapNet
Full	N/A	N/A	222.49m,33.80°	80.32m,13.73°	46.61m,10.45°	44.61m,10.38°	**19.18m,4.60°**
Loop	N/A	N/A	7.88m,3.87°	22.42m,45.50°	7.90m,3.53°	9.29m,3.34°	**6.45m,2.98°**
Average	N/A	N/A	115.19m,18.84°	51.37m,29.62°	27.26m,6.99°	26.95m,6.86°	**12.82m,3.79°**

Table 3. Comparison with related works on Oxford RobotCar. DSAC [3] needs a dense 3D model so it can not figure out poses from RGB image. ORB-SLAM2 [32], a geometry based localization system, takes a long time to initialize and fails to track when the car turns. DBoW3 [1] is a topological localization method, which works better on loop than on full because full is more dynamic (which is also the reason our model achieves more improvement on full). The trajectory of Stereo VO is provided by Oxford RobotCar dataset, and we compute its accuracy after aligning it to ground truth. Our model (AD-MapNet) outperforms previous works and achieves superior accuracy on both full and loop.

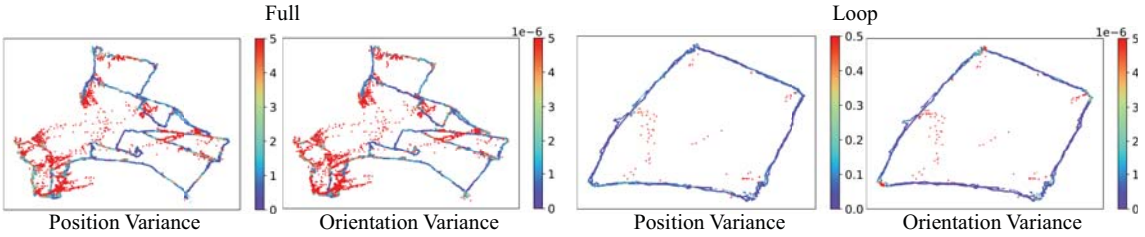

Figure 9. We sample 9 times with AD-MapNet to estimate the posterior pose distribution on RobotCar dataset including mean and variance of position and orientation. The scatter points whose position and color is given by mean pose and variance show that higher variance denotes lower localization accuracy, which means the variance provided by our method is a good uncertainty measurement. A detailed variance analysis is provided in supplementary.

Scene	T=0.05	T=0.1	T=0.2	T=0.4
Full (P)	29.74m,8.86°	25.20m,9.31°	19.18m,4.60°	33.03m,7.59°
Loop (P)	6.62m,3.05°	5.65m,2.53°	6.45m,2.98°	6.51m,3.06°
Full (U)	48.35m,12.58°	54.81m,13.95°	53.69m,13.25°	46.53m,11.69°
Loop (U)	10.57m,4.58°	10.58m,5.70°	9.53m,3.60°	15.71m,6.06°

Table 4. Localization accuracy of AD-MapNet trained with different T is presented in the table. P and U respectively denotes prior guided dropout and uniform dropout. The difference between the accuracy of P and U shows that prior guided dropout surpass uniform dropout, which validates the effectiveness of prior guided dropout module. $T = 0.2$ and $T = 0.1$ are the best choices on full and loop, and both higher and lower T degrades the performance. It proves that T is an effective bias compensation approach.

ent places, along with corresponding pose ground-truths computed by the structure from motion. King's College, St Mary's Church and ShopFacade are three challenging scenes containing significant clutters as reported by [28].

The Oxford RobotCar contains over 100 sequences captured from a consistent route in Oxford, UK, by a car equipped with sensors. The video sequences are replete with complex traffic conditions and abundant with movable objects such as vehicles and pedestrians. Following the previous work of Brahmbhatt *et al*. [4], we extract several sequences from the whole dataset constituting two different scenes to experiment on, *i.e*., *full* and *loop*.

4.2. Experimental Setup

Following the work of Brahmbhatt *et al*. [4], in all our experiments, we apply logarithm of a unit quaternion *i.e*., $\boldsymbol{\theta} = \log \mathbf{q}$, to measure the camera orientation, and use the loss function given by

$$\text{loss}(\mathbf{x}, \hat{\mathbf{x}}) = ||\mathbf{p} - \hat{\mathbf{p}}||_1 e^{-\beta} + \beta + ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_1 e^{-\gamma} + \gamma \quad (5)$$

where $\mathbf{x} = (\mathbf{p}, \boldsymbol{\theta})$ denotes the ground-truth of camera pose composed of the position $\mathbf{p}$ and orientation $\boldsymbol{\theta}$, $\hat{\mathbf{x}} = (\hat{\mathbf{p}}, \hat{\boldsymbol{\theta}})$ represents the predicted camera pose, while β and γ are two learnable parameters to adaptively balance the position loss and orientation loss. Specifically, we use Mask R-CNN [19] to segment the movable objects including six classes (*i.e*., bus, car, person, bicycle, truck and motorcycle) generating the binary masks, based on which the parameters of Gaussian distribution (Fig. 7) is calculated as mentioned in Sec. 3.1, and the dropout threshold T used in the prior guided dropout is set to 0.2.

4.3. Ablation Study

We carry out a thorough ablation study to demonstrate the effectiveness of each proposed module. To demonstrate the universality of the proposed modules, we test our methods by applying them to state-of-the-art pose regressors PoseNet [28] and MapNet [4].

We assess the effectiveness of our methods by sequentially adding the proposed modules to the raw neural pose regressors, *i.e*., PoseNet and MapNet, and evaluate the performance step by step. To be more specific, the D-* and A-* ones in Table. 1 denotes the model only applied with the prior guided dropout or self-attention module in training, and the models who named with prefix 'AD' (self-

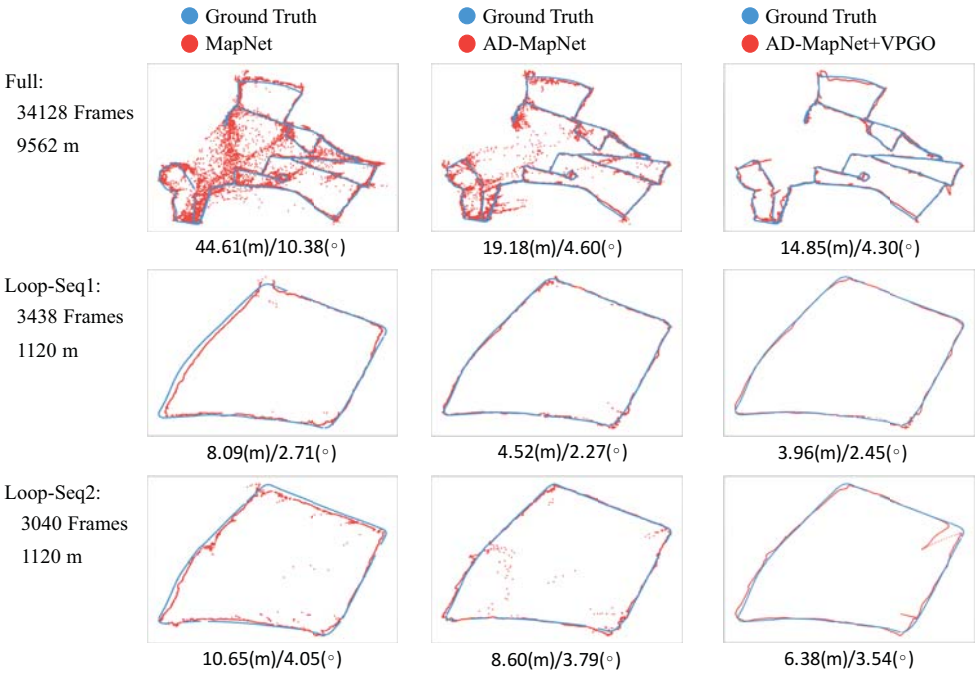

Figure 10. **Comparison of localization results given by MapNet, AD-MapNet, and AD-MapNet+VPGO on full and loop**. As there are two sequences (Seq1 is 2014-06-23-15-36-04, and Seq2 denotes 2014-06-26-08-53-56) in the test of loop, we perform PGO for them respectively. Compared to MapNet, AD-MapNet significantly reduces failure localization results, and AD-MapNet+VPGO converges most outliers because of the highly confidential variance estimation. Our framework improves MapNet by 66.7%/58.6% on 2014-12-09-13-21-02 (full), 51.1%/9.6% on 2014-06-23-15-36-04 and 40.0%/12.6% on 2014-06-26-08-53-56.

attention and prior guided dropout) are armed with both prior guided dropout and self-attention module. Furthermore, our enhanced PGO method (*+VPGO) described in Sec. 3.3 is also evaluated by comparing with the version (*+CPGO) adopted by Brahmbhatt *et al*. [4], which uses constant weights on each term in Eq. (3). The mean and variance utilized in VPGO are calculated through prior guided dropout in test. As illustrated in Table 1, the accuracy is improved as the integrity is increasing.

We investigate the influence of threshold T and prior guided dropout module on the prediction results exhibited in Table 4, which shows that the prior guided dropout exerts a positive impact to filter out the irrelevant objects and improves the localization accuracy. Moreover, we visualize the relationship between the prediction accuracy and the variance measured by our prior guided dropout module. As displayed in Fig. 9, the predictions with lower variances incline to lie nearer to the ground-truth trajectory, while those with high variances deviates from the trajectory. The strong negative correlation between the prediction accuracy and variance manifests that the prior guided dropout can be exploited as weight terms in PGO to further refine the predicted poses globally as mentioned in Sec. 3.3. As plotted in Fig. 10, our proposed VPGO rectify the predictions far from the trajectory by global optimization, which significantly improves the localization accuracy and robustness.

4.4. Comparison with Related Works

We compare with the state-of-the-art works on both Cambridge Landmarks and Oxford RobotCar datasets as demonstrated in Table. 2 and Table. 3.3. The experiments show the existing neural pose regressors are seriously impacted by dynamic objects, while our framework significantly improves the robustness of neural pose regressor in dynamic environments and outperforms the existing methods by a large margin on the challenging RobotCar dataset.

5. Conclusions

We proposed a novel visual localization framework which significantly improves the accuracy and robustness of modern neural pose regressor in dynamic environments, especially in complicated urban traffic. The integration of prior guided dropout and self-attention strategy is quite useful and can be easily incorporated into a modern neural network, which provides a chance to introduce prior into neural network and an uncertainty evaluation method. These modules might be also useful for other tasks where data is contaminated by gross errors whose distribution can be precomputed by some other methods from data.

References

[1] DBoW3. 2017. https://github.com/rmsalinas/DBow3.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Proceedings of the European conference on computer vision*, pages 404–417. Springer, 2006.

[3] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC-differentiable RANSAC for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6684–6692, 2017.

[4] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2616–2625, 2018.

[5] Zetao Chen, Obadiah Lam, Adam Jacobson, and Michael Milford. Convolutional neural network-based place recognition. *arXiv preprint arXiv:1411.1509*, 2014.

[6] Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, and Hongkai Wen. Vidloc: A deep spatio-temporal model for 6-DoF video-clip relocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2652–2660, 2017.

[7] Mark Cummins. Highly scalable appearance-only SLAM-FAB-MAP 2.0. *In Proceedings of Robotics: Sciences and Systems (RSS)*, 2009.

[8] Mark Cummins and Paul Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011.

[9] Jonathan Delhumeau, Philippe-Henri Gosselin, Hervé Jégou, and Patrick Pérez. Revisiting the VLAD image representation. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 653–656. ACM, 2013.

[10] Jun Feng Dong, W. Sardha Wijesoma, and Andrew P. Shacklock. Extended rao-blackwellised genetic algorithmic filter SLAM in dynamic environment with raw sensor measurement. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1473–1478. IEEE, 2007.

[11] Ethan Eade and Tom Drummond. Edge landmarks in monocular SLAM. *Image and Vision Computing*, 27(5):588–596, 2009.

[12] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[13] Jakob Engel, Jörg Stückler, and Daniel Cremers. Large-scale direct SLAM with stereo cameras. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1935–1942. IEEE, 2015.

[14] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast semi-direct monocular visual odometry. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 1050–1059, 2016.

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

[17] A. H. Abdul Hafez, Manpreet Singh, K. Madhava Krishna, and C. V. Jawahar. Visual localization in highly crowded urban environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2778–2783. IEEE, 2013.

[18] Josiah P Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 538–546. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, 2017.

[20] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7:7132–7141, 2017.

[22] Marco Imperoli and Alberto Pretto. Active detection and localization of textureless objects in cluttered environments. *arXiv preprint arXiv:1603.07022*, 2016.

[23] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3304–3311. IEEE, 2010.

[24] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34:1704–1716, 2012.

[25] Edward Johns and Guang-Zhong Yang. Feature co-occurrence maps: Appearance-based localisation throughout the day. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3212–3218. IEEE, 2013.

[26] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *Proceedings of the IEEE international conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE, 2016.

[27] Alex Kendall, Roberto Cipolla, et al. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6555–6564, 2017.

[28] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DoF camera relocalization. In *Proceedings of the IEEE International*

Conference on Computer Vision (ICCV), pages 2938–2946, 2015.

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

[30] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.

[31] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017.

[32] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[33] Tayyab Naseer and Wolfram Burgard. Deep regression for monocular camera-based 6-DoF global localization in outdoor environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1525–1530. IEEE, 2017.

[34] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2161–2168. IEEE Computer Society, 2006.

[35] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1717–1724, 2014.

[36] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE international conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011.

[37] Grant Schindler, Matthew A. Brown, and Richard Szeliski. City-scale location recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2007.

[38] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[39] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[40] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of convnet features for place recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4297–4304. IEEE, 2015.

[41] Abhinav Valada, Noha Radwan, and Wolfram Burgard. Deep auxiliary learning for visual localization and odometry. *arXiv preprint arXiv:1803.03642*, 2018.

[42] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. SfM-Net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.

[43] Chieh-Chih Wang, Charles Thorpe, Sebastian Thrun, Martial Hebert, and Hugh Durrant-Whyte. Simultaneous localization, mapping and moving object tracking. *The International Journal of Robotics Research*, 26(9):889–916, 2007.

[44] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1983–1992, 2018.