

RobustLoc: Robust Camera Pose Regression in Challenging Driving Environments

Sijie Wang^{1*}, Qiyu Kang^{1*}, Rui She^{1*}, Wee Peng Tay¹,
Andreas Hartmannsgruber², Diego Navarro Navarro²

¹ Continental-NTU Corporate Lab, Nanyang Technological University

² Continental Automotive Singapore

{wang1679@e.; qiyu.kang@; rui.she@; wptay@}ntu.edu.sg,

{andreas.hartmannsgruber; diego.navarro.navarro}@continental.com

Abstract

Camera relocalization has various applications in autonomous driving. Previous camera pose regression models consider only ideal scenarios where there is little environmental perturbation. To deal with challenging driving environments that may have changing seasons, weather, illumination, and the presence of unstable objects, we propose RobustLoc, which derives its robustness against perturbations from neural differential equations. Our model uses a convolutional neural network to extract feature maps from multi-view images, a robust neural differential equation diffusion block module to diffuse information interactively, and a branched pose decoder with multi-layer training to estimate the vehicle poses. Experiments demonstrate that RobustLoc surpasses current state-of-the-art camera pose regression models and achieves robust performance in various environments. Our code is released at: <https://github.com/sijieaaa/RobustLoc>

Introduction

Accurate camera relocalization plays an important role in autonomous driving. Given query images, the camera relocalization task aims at estimating the camera poses for which the images are taken. In recent years, many camera relocalization approaches have been proposed. Generally speaking, they fall into two categories.

1. One solution is to treat relocalization as a matching task. This solution assumes the availability of a database or a map that stores prior information (e.g., 3D point clouds, images, or descriptors) of sample points. Given a query image, the matching model finds the best match between the query and the database based on a similarity score. The estimated camera pose is then inferred from the matched prior information in the database.
2. Another solution does not assume the availability of a database and uses only neural networks to regress the camera poses of the query images. This approach constructs an implicit relation between images and poses, which is called *camera pose regression* (CPR).

The prerequisite of a database on the one hand can boost the accuracy of the camera relocalization by storing useful prior

*These authors contributed equally.

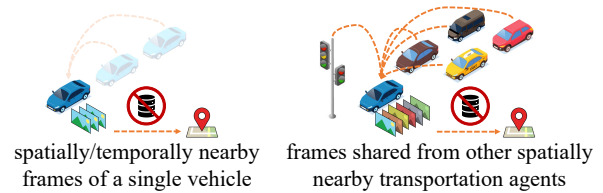


Figure 1: Multi-view camera pose regression with neighboring information, without the need for any database.

information. On the other hand, the computation and storage requirements are proportionate to the number of sample points in the database. To decouple relocalization from the need for a database, there has been a recent surge of research interest in the second category CPR.

The pioneering work PoseNet (Kendall, Grimes, and Cipolla 2015) uses a convolutional neural network (CNN) to extract features from a single image as vector embeddings, and the embeddings are directly regressed to the 6-DoF poses. To further improve the regression performance in driving scenarios, multi-view-based models extend the input from a single image to multi-view images. MapNet (Brahmbhatt et al. 2018) leverages pre-computed visual odometry to post-process the output pose trajectory. GNNMapNet (Xue et al. 2020) integrates a graph neural network (GNN) into CNN to make image nodes interact with neighbors.

The above-mentioned multi-view-based models show promising performance in benign driving environments. To operate well in challenging environments, the model must be robust to environmental perturbations (e.g., changing seasons, weather, illumination, and unstable objects), and effectively leverage neighboring information from spatially or temporally nearby frames of a single vehicle or multi-view images shared from other spatially nearby agents (e.g., using V2X communication) as shown in Fig. 1. Images sharing such neighboring information are said to be *covisible*.

Recently, neural Ordinary Differential Equations (ODEs) (Chen et al. 2018) and Partial Differential Equations (PDEs) (Chamberlain et al. 2021b,a) have demonstrated their robustness against input perturbations (Yan et al. 2018; Kang et al. 2021). Moreover, GNNs can effectively aggregate neighborhood information. We thus propose RobustLoc that not only

explores the relations between graph neighbors but also utilizes neural differential equations to improve robustness. We test our new multi-view-based model on three challenging autonomous driving datasets and verify that it outperforms existing state-of-the-art (SOTA) CPR methods.

Our main contributions are summarized as follows:

1. We represent the features extracted from a CNN in a graph and apply graph neural diffusion layers at each stage. I.e., we design feature diffusion blocks at both the feature map extraction and vector embedding stages to achieve robust feature representations. Each diffusion block consists of not only cross-diffusion from node to node in a graph but also self-diffusion within each node. We also propose multi-level training with the branched decoder to better regress the target poses.
2. We conduct experiments in both ideal and challenging noisy autonomous driving datasets to demonstrate the robustness of our proposed method. The experiments verify that our method achieves better performance than the current SOTA CPR methods.
3. We conduct extensive ablation studies to provide insights into the effectiveness of our design.

Related Work

Camera Pose Regression

Given the query images, CPR models directly regress the camera poses of these images without the need for a database. Thus, it does not depend on the scale of the database, which is definitely a born gift compared with those database methods. PoseNet (Kendall, Grimes, and Cipolla 2015) and Geo-PoseNet (Kendall and Cipolla 2017) propose the simultaneous learning for location and orientation by integrating balance parameters. MapNet (Brahmbhatt et al. 2018) uses visual odometry to serve as the post-processing technique to optimize the regressed poses. LsG (Xue et al. 2019) and LSTM-PoseNet (Walch et al. 2017) integrates the sequential information by fusing PoseNet and LSTM. AD-PoseNet and AD-MapNet (Huang et al. 2019) leverages the semantic masks to drop out the dynamic area in the image. AtLoc (Wang et al. 2020) introduces the global attention to guide the network to learn better representation. GNNMapNet (Xue et al. 2020) expands the feature exploration from a single image to multi-view images using GNN. IRPNet (Shavit and Ferens 2021) proposes to use two branches to regress translation and orientation respectively. Coordinet (Moreau et al. 2022) uses the coordconv (Liu et al. 2018) and weighted average pooling (Hu, Wang, and Lin 2017) to capture spatial relations.

Neural Differential Equations and Robustness

The dynamics of a system are usually described by ordinary or partial differential equations. The paper (Chen et al. 2018) first proposes trainable neural ODEs by parameterizing the continuous dynamics of hidden units. The hidden state of the ODE network is modeled as:

$$\frac{d\mathbf{y}(t)}{dt} = f_{\theta}(\mathbf{y}(t)) \quad (1)$$

where $\mathbf{y}(t)$ denotes the latent state of the trainable network f_{θ} that is parameterized by weights θ . Recent studies (Yan et al. 2018; Kang et al. 2021) have demonstrated that neural ODEs are intrinsically more robust against input perturbations compared to vanilla CNNs.

In addition, neural PDEs (Chamberlain et al. 2021b,a) have been proposed and applied to GNN, where the diffusion process is modeled on the graph. Furthermore, the stability of the heat semigroup and the heat kernel under perturbations of the Laplace operator (i.e., local perturbation of the manifold) is studied (Song et al. 2022).

Proposed Model

In this section, we provide a detailed description of our proposed CPR approach. We assume that the input is a set of images $\{\mathbf{I}_i\}_{i \in [N]}$ that may be covisible (see Fig. 1).¹ Our objective is to perform CPR on the input images.

RobustLoc Overview

We first summarize our multi-view CPR pipeline, which can be decomposed into three different stages, as follows (see Fig. 2 and Fig. 3):

1. Given N neighboring images, a CNN extracts the feature maps of all these images. Our proposed feature map diffusion block then performs cross-self diffusion on the feature maps.
2. After feature map diffusion, a global average pooling module aggregates the feature maps as vector embeddings, which contain global representations of these images. Similarly, those vector embeddings are then diffused by cascaded diffusion blocks.
3. Based on the vector embeddings, the branched decoder module regresses the output camera poses. During training, decoding is performed on multiple levels to provide better feature constraints.

Neural Diffusion for Feature Maps

The input images $\{\mathbf{I}_i\}_{i \in [N]}$ are passed through a CNN to obtain the feature maps $\{\mathbf{m}_i \in \mathbb{R}^{H \times W \times C}\}_{i \in [N]}$. Here, C is the channel dimension, while H and W are the dimensions of a feature map. For each feature map \mathbf{m}_i , we denote its j -th element as $\mathbf{m}_{i,j} \in \mathbb{R}^C, j \in [HW]$. We next describe the feature map diffusion block, where we perform cross-diffusion from node to node in a graph, and self-diffusion within each node. The two diffusion processes update the feature map by leveraging the neighboring information or only using each node’s individual information, respectively.

Cross-Diffusion Dynamics. To support the cross-diffusion over feature maps, we formulate the first graph in our pipeline as:

$$\mathcal{G}^{\text{feat}} = (\mathcal{V}^{\text{feat}}, \mathcal{E}^{\text{feat}}), \quad (2)$$

¹Notations: In this paper, we use $[N]$ to denote the set of integers $\{1, 2, \dots, N\}$. We use boldfaced lowercase letters like \mathbf{m} to denote vectors and boldface capital letters like \mathbf{W} to denote matrices.

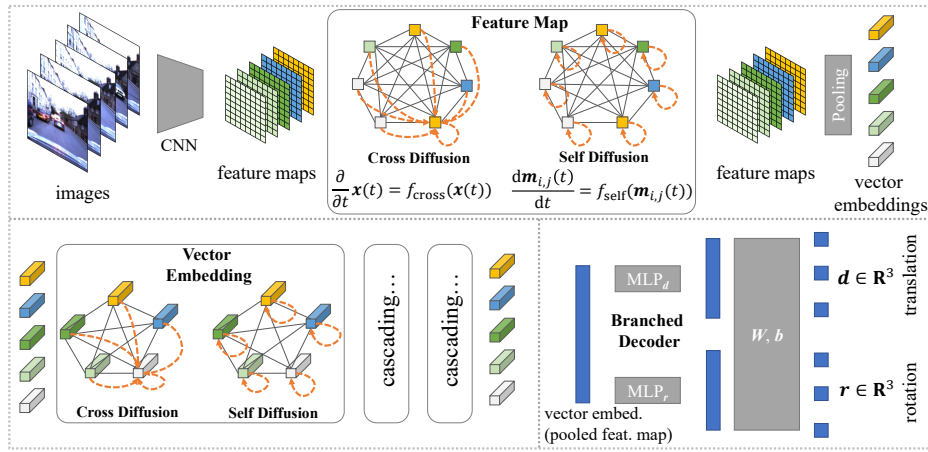


Figure 2: The main architecture of RobustLoc. Feature diffusion is performed at both the feature map stage and the vector embedding stage. The branched decoder regresses the 6-DoF poses based on the vector embeddings or the pooled feature maps. The details for multi-layer decoding are shown in Fig. 3.

where the node set $\mathcal{V}^{\text{feat}} = \{\mathbf{m}_{i,j}\}_{(i,j) \in [N] \times [HW]}$ contains element-wise features $\mathbf{m}_{i,j}$ and the edge set $\mathcal{E}^{\text{feat}}$ is defined as the complete graph edges associated with attention weights as discussed below. And the complete graph architecture is demonstrated to be an effective design shown in Table 5.

To achieve robust feature interaction, we next define the cross-diffusion process as:

$$\frac{\partial}{\partial t} \mathbf{x}(t) = f_{\text{cross}}(\mathbf{x}(t)), \quad (3)$$

where $f_{\text{cross}}(\mathbf{x}(t))$ is a neural network and can be approximately viewed as a neural PDE with the partial differential operations over a manifold space replaced by the attention modules that we will introduce later. We denote the input to the feature map diffusion module as the initial state at $t = t_0$ as $\mathbf{x}(t_0) = \{\mathbf{m}_{i,j}\}_{(i,j) \in [N] \times [HW]}$, where $\mathbf{x}(t) = \{\mathbf{m}_{i,j}(t)\}_{(i,j) \in [N] \times [HW]}$ denotes the hidden state of the diffusion. The diffusion process is known to have robustness against local perturbations of the manifold (Chen et al. 1998) where the local perturbations in our CPR task include challenging weather conditions, dynamic street objects, and unexpected image noise. Therefore, we expect our module (3) is simultaneously capable of leveraging the neighboring image information and holding robustness against local perturbations.

We next introduce the computation of attention weights in $f_{\text{cross}}(\mathbf{x}(t))$ for node features at time t . We first generate the embedding of each node using multi-head fully connected (FC) layers with learnable parameter matrix \mathbf{W}_k and bias \mathbf{b}_k at each head $k = [K]$, where K is the number of heads. The output at each head k can be written as:

$$\mathbf{m}_{i,j;k}^{\text{FC}}(t) = \mathbf{W}_k \mathbf{m}_{i,j}(t) + \mathbf{b}_k. \quad (4)$$

The attention weights are then generated by computing the dot product among all the neighboring nodes using the fea-

tures $\{\mathbf{m}_{i,j;k}^{\text{FC}}(t)\}_{(i,j) \in [N] \times [HW]}$. We have

$$\{a_{(i,j),(i',j');k}(t)\}_{(i',j') \in \mathcal{N}_{i,j}} = \text{Softmax}_{(i',j') \in \mathcal{N}(i,j)}(\mathbf{m}_{i,j;k}^{\text{FC}}(t) \cdot \mathbf{m}_{i',j',k}^{\text{FC}}(t)), \quad (5)$$

where $\mathcal{N}_{i,j}$ denotes the set of neighbors of node $\mathbf{m}_{i,j}$. Let

$$\mathbf{m}_{i,j;k}^{\text{weighted}}(t) = \sum_{(i',j') \in \mathcal{N}_{i,j}} a_{(i,j),(i',j');k}(t) \mathbf{m}_{i',j',k}^{\text{FC}}(t). \quad (6)$$

Finally, the updated node features are obtained by concatenating the weighted node features from all heads as

$$f_{\text{cross}}(\mathbf{x}(t)) = \left\{ \left\| \left\| \mathbf{m}_{i,j;k}^{\text{weighted}}(t) \right\| \right\|_{k \in [K]} \right\}_{(i,j) \in [N] \times [HW]}. \quad (7)$$

Based on the above pipeline, the output of the cross-diffusion at time $t = t_1$ can be obtained as:

$$\mathbf{x}(t_1) = F_{\text{cross}}(\mathbf{x}(t_0)), \quad (8)$$

where $F_{\text{cross}}(\cdot)$ denotes the solution of (3) integrated from $t = t_0$ to $t = t_1$.

Self-Diffusion Dynamics. In the next step, we update each node feature independently. The node-wise feature update can be regarded as a rewiring of the complete graph to an edgeless graph, and the node-wise feature update is described as:

$$\frac{d\mathbf{m}_{i,j}(t)}{dt} = f_{\text{self}}(\mathbf{m}_{i,j}(t)) = \text{MLP}(\mathbf{m}_{i,j}(t)). \quad (9)$$

And the output of self-diffusion can be obtained as:

$$\mathbf{m}_{i,j}(t_2) = F_{\text{self}}(\mathbf{m}_{i,j}(t_1)). \quad (10)$$

where $F_{\text{self}}(\cdot)$ denotes the solution of (9) integrated from $t = t_1$ to $t = t_2$. As neural ODEs are robust against input perturbations (Yan et al. 2018; Kang et al. 2021), we expect the updating of each node feature according to the self-diffusion (9) to be robust against perturbations like challenging weather conditions, dynamic street objects, and image noise.

Vector Embeddings and Diffusion

After the feature map neural diffusion, we feed the updated feature maps into a global average pooling module to generate the vector embeddings $\{\mathbf{h}_i \in \mathbb{R}^C\}_{i \in [N]}$, where

$$\mathbf{h}_i = \text{Pooling}(\mathbf{m}_i). \quad (11)$$

Each vector embedding contains rich global representations for the input image together with the information diffused from the neighboring images. To enable diffusion for the global information, we propose to design the vector embedding graph as:

$$\mathcal{G}^{\text{vect}} = (\mathcal{V}^{\text{vect}}, \mathcal{E}^{\text{vect}}), \quad (12)$$

where the node set $\mathcal{V}^{\text{vect}} = \{\mathbf{h}_i\}_{i \in [N]}$ contains image vector embeddings \mathbf{h}_i and the edge set $\mathcal{E}^{\text{vect}}$ is also defined to be the complete graph. Based on this graph $\mathcal{G}^{\text{vect}}$, we construct the cascaded diffusion blocks, to perform global information diffusion. Within the cascaded blocks, each basic diffusion block consists of two diffusion layers: a cross-diffusion layer and a self-diffusion layer, similar to the two diffusion schemes introduced at the feature map diffusion phase.

Pose Decoding

In this subsection, we explain the pose decoding operations.

Branched Pose Decoder. Each camera pose $\mathbf{p} = \{\mathbf{d}, \mathbf{r}\} \in \mathbb{R}^6$, consists of a 3-dimensional translation $\mathbf{d} \in \mathbb{R}^3$ and a 3-dimensional rotation $\mathbf{r} \in \mathbb{R}^3$. Thus CPR can be viewed as a multi-task learning problem. However, since the translation and rotation elements of \mathbf{p} do not scale compatibly, the regression converges in different basins. To deal with it, previous methods consider regression for translation and rotation respectively and demonstrate it is an effective way to improve performance (Shavit and Ferens 2021). In our paper, we also follow this insight to design the decoder.

Firstly, the feature embeddings $\{\mathbf{h}_d, \mathbf{h}_r\}$ for translation and rotation are extracted from the feature embedding \mathbf{h} using different non-linear MLP layers as:

$$\mathbf{h}_d = \text{MLP}_d(\mathbf{h}), \quad (13)$$

$$\mathbf{h}_r = \text{MLP}_r(\mathbf{h}), \quad (14)$$

Thus, the features of translation and rotation are decoupled. Next in the second stage, the pose output can be regressed as:

$$\mathbf{p} = \mathbf{W}(\mathbf{h}_d \parallel \mathbf{h}_r) + \mathbf{b} \quad (15)$$

where \mathbf{W}, \mathbf{b} are learnable parameters. During training, we compute the regression loss of decoded poses from multiple levels, which we will introduce below. During inference, we use the decoded pose from the last layer as the final output pose.

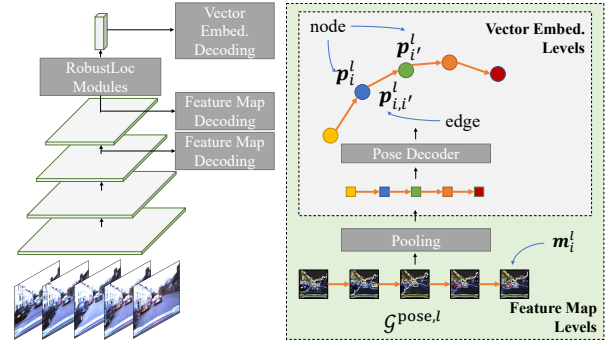


Figure 3: Multi-level pose decoding. Decoding can be directly applied to vector embeddings. Feature maps are first pooled and then decoded.

Multi-level Pose Decoding Graph. To better regularize the whole regression pipeline, we propose to leverage the feature maps at multiple levels. As shown in Fig. 3, at the vector embedding stage, we use the vector embeddings to regress the poses, while at the feature map stage, we use the feature maps. Denoting the feature maps at layer l as $\{\mathbf{m}_i^l \in \mathbb{R}^{H \times W \times C}\}_{i \in [N]}$, the pose decoding graph at layer l can be formulated as:

$$\mathcal{G}^{\text{pose},l} = (\mathcal{V}^{\text{pose},l}, \mathcal{E}^{\text{pose},l}), \quad (16)$$

where edge set $\mathcal{E}^{\text{pose},l}$ is defined to be connected with two spatially adjacent nodes which can be viewed as the odometry connection, while the node set $\mathcal{V}^{\text{pose},l}$ is defined depending on layers since the information used to regress poses is different:

$$\mathcal{V}^{\text{pose},l} = \begin{cases} \{\mathbf{h}_i\}_{i \in [N]} & \text{if } l = L, \\ \{\mathbf{m}_i^l\}_{i \in [N]} & \text{otherwise,} \end{cases} \quad (17)$$

where L represents the last layer in our network.

At the last layer where there are vector embeddings, we can directly apply the pose decoder to generate absolute pose messages. By contrast, at feature map layers, we first apply a global average pooling module on the feature maps to formulate feature vectors, and pose messages can be obtained using the pose decoder:

$$\mathbf{p}_i^l = \begin{cases} f_{\text{decoder}}^l(\mathbf{h}_i^l) & \text{if } l = L, \\ f_{\text{decoder}}^l(\text{Pooling}(\mathbf{m}_i^l)) & \text{otherwise.} \end{cases} \quad (18)$$

where $f_{\text{decoder}}^l(\cdot)$ is the pose decoder at layer l . Using the simplified relative pose computation technique in (Wang et al. 2020), the relative pose messages $\mathbf{p}_{i,i'}^l$ at layer l can be generated as:

$$\mathbf{p}_{i,i'}^l = \mathbf{p}_{i'}^l - \mathbf{p}_i^l. \quad (19)$$

By leveraging multi-layer information, we expect not only the last layer but also the preceding middle-level layers can directly learn the implicit relation between images and poses, which helps to improve the robustness against perturbations.

	Model	Loop (cross-day)		Loop (within-day)		Full	
		Mean	Median	Mean	Median	Mean	Median
+ Extra Data	GNNMapNet + <i>post.</i>	7.96 / <u>2.56</u>	-	-	-	17.35 / <u>3.47</u>	-
	ADPoseNet	-	-	-	6.40 / 3.09	-	33.82 / 6.77
	ADMapNet	-	-	-	6.45 / 2.98	-	19.18 / 4.60
	MapNet+	8.17 / 2.62	-	-	-	30.3 / 7.8	-
	MapNet+ + <i>post.</i>	6.73 / 2.23	-	-	-	29.5 / 7.8	-
CPR Only	GeoPoseNet	27.05 / 18.54	6.34 / 2.06	-	-	125.6 / 27.1	107.6 / 22.5
	MapNet	9.30 / 3.71	5.35 / <u>1.61</u>	-	-	41.4 / 12.5	17.94 / 6.68
	LsG	9.08 / 3.43	-	-	-	31.65 / 4.51	-
	AtLoc	8.74 / 4.63	5.37 / 2.12	-	-	29.6 / 12.4	11.1 / 5.28
	AtLoc+	<u>7.53</u> / 3.61	<u>4.06</u> / 1.98	-	-	21.0 / 6.15	6.40 / 1.50
	CoordiNet	-	-	4.06 / <u>1.44</u>	<u>2.42</u> / 0.88	14.96 / 5.74	3.55 / <u>1.14</u>
	RobustLoc (ours)	4.68 / 2.67	3.70 / 1.50	2.49 / 1.40	1.97 / 0.84	9.37 / 2.47	<u>5.93</u> / 1.06

Table 1: Median and mean translation/rotation estimation error (m^o) on the Oxford RobotCar dataset. The best and the second-best results in each metric are highlighted with bold and underline respectively. “-” denotes no data provided.

Loss Function

Following the approach in (Wang et al. 2020), we use a weighted balance loss for translation and rotation predictions. For the input image I_i , we denote the translation and rotation targets as $\mathbf{d}_i^* \in \mathbb{R}^3$ and $\mathbf{r}_i^* \in \mathbb{R}^3$ respectively. Then the absolute pose loss term \mathcal{L}_i^l and the relative pose loss term $\mathcal{L}_{i,i'}^l$ at decoding layer l are computed as:

$$\mathcal{L}_i^l = \|\mathbf{d}_i^l - \mathbf{d}_i^*\| \exp(-\alpha) + \alpha + \|\mathbf{r}_i^l - \mathbf{r}_i^*\| \exp(-\beta) + \beta, \quad (20)$$

$$\mathcal{L}_{i,i'}^l = \|\mathbf{d}_{i,i'}^l - \mathbf{d}_{i,i'}^*\| \exp(-\gamma) + \gamma + \|\mathbf{r}_{i,i'}^l - \mathbf{r}_{i,i'}^*\| \exp(-\lambda) + \lambda, \quad (21)$$

where $\mathbf{d}_i^l, \mathbf{r}_i^l, \mathbf{d}_{i,i'}^l, \mathbf{r}_{i,i'}^l$ are outputs at layer l , while $\alpha, \beta, \gamma, \lambda$ are all learnable parameters. Finally, the overall loss function can be obtained as:

$$\mathcal{L} = \sum_{l \in \{3,4,L\}} \sum_{i \in [N], i' \in \mathcal{N}_i^l} \mathcal{L}_i^l + \mathcal{L}_{i,i'}^l, \quad (22)$$

where \mathcal{N}_i^l is the neighborhood of node i in $\mathcal{G}^{\text{pose},l}$. We use the logarithmic form of the quaternion to represent rotation \mathbf{r} as:

$$\mathbf{r} = \log \mathbf{q} = \begin{cases} \frac{(\mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4)}{\|(\mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4)\|} \cos^{-1} \mathbf{q}_1 & \text{if } \|(\mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4)\| \neq 0, \\ 0 & \text{otherwise,} \end{cases} \quad (23)$$

where $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4) \in \mathbb{R}^4$ represents a quaternion.

Experiments

In this section, we first evaluate our proposed model on three large autonomous driving datasets. We next present an ablation study to demonstrate the effectiveness of our model design.

Datasets and Implementation Details

Oxford RobotCar. The Oxford RobotCar dataset (Maddern et al. 2017) is a large autonomous driving dataset collected



Figure 4: Visualization of the Perturbed RobotCar dataset. Medium is with fog, snow, rain, and spatter on the lens. Hard is with added Gaussian noise.

by a car driving along a route in Oxford, UK. It consists of two different routes: 1) Loop with a trajectory area of $8.8 \times 10^4 \text{m}^2$ and length of 10^3m , and 2) Full with a trajectory area of $1.2 \times 10^6 \text{m}^2$ and length of $9 \times 10^3 \text{m}$.

4Seasons. There are only a few existing methods designed for robust CPR in driving environments, and the experiment on the Oxford dataset is insufficient for comparison. Thus we also conduct experiments on another driving dataset to cover more driving scenarios. The 4Seasons dataset (Wenzel et al. 2020) is a comprehensive dataset for autonomous driving SLAM. It was collected in Munich, Germany, covering varying perceptual conditions. Specifically, it contains different environments including the business area, the residential area, and the town area. In addition, it consists of a wide variety of weather conditions and illuminations. In our experiments, we use 1) Business Campus (business area), 2) Neighborhood (residential area), and 3) Old Town (town area).

Perturbed RobotCar. To further evaluate the performance under challenging environments, we inject noise into the RobotCar Loop dataset and call this the Perturbed RobotCar dataset as shown in Fig. 4. We create three scenarios: 1) Medium (with fog, snow, rain, and spatter on the lens), 2) Hard (with added Gaussian noise), and 3) Hard (+ *noisy training*) (i.e., training with noisy augmentation).

Implementation We use ResNet34 as the backbone, which is pre-trained on the ImageNet dataset. We set the maximum number of input images as 11. We resize the shorter side of each input image to 128 and set the batch size to 64. The

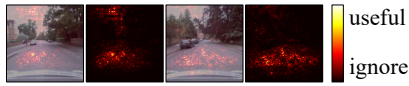


Figure 5: Robust features from RobustLoc.

Adam optimizer with a learning rate 2×10^{-4} and weight decay 5×10^{-4} is used to train the network. Data augmentation techniques include random cropping and color jittering. We set the integration times $t_0 = 0$, $t_1 = 1$, and $t_2 = 2$. The number of attention heads is 8. We train our network for 300 epochs. All of the experiments are conducted on an NVIDIA A5000.

Main Results

On the Oxford RobotCar dataset, as shown in Table 1, we obtain the best performance in 10 out of 12 metrics. Using the mean error, which is easily influenced by outlier predictions, RobustLoc outperforms the baselines by a significant margin. In the most challenging route Full, to the best of our knowledge, RobustLoc is the first to achieve less than 10m mean translation error for CPR.

The 4Seasons dataset consists of more varied driving scenes. As shown in Table 2, RobustLoc achieves the best performance in 11 out of 12 metrics. Again, using the mean error metric, RobustLoc outperforms the baselines by a significant margin.

On the Perturbed RobotCar dataset, where the images contain more challenging weather conditions and noisy perturbations, RobustLoc achieves the best in all metrics. The superiority of RobustLoc over other baselines is more obvious in Table 3.

Analysis

Ablation Study. We justify our design for RobustLoc by ablating each module. From Table 4, we observe that every module in our design contributes to the final improved estimation. We see that making use of neighboring information from covisible frames and learning robust feature maps contribute to more accurate CPR.

Salience Visualization. Salience maps shown in Fig. 5 suggest that in driving environments, RobustLoc pays more attention to relatively robust features such as the skyline and the road, similar to PixLoc (Sarlin et al. 2021). In addition, dynamic objects such as vehicles are implicitly suppressed in RobustLoc’s regression pipeline.

Diffusion and Augmentation. Using multi-level features is an effective method in dense prediction tasks such as depth estimation (Yan et al. 2021). To test if this holds in CPR, we use the feature maps from the lower stage 3 (see Fig. 3), which however does not lead to performance improvement shown in Table 4. We also utilize the multi-level concatenation strategy used in GNNMapNet. This does not lead to significant changes. These experiments demonstrate that CPR benefits more from high-level features with more semantic information than from low-level local texture features. Finally,

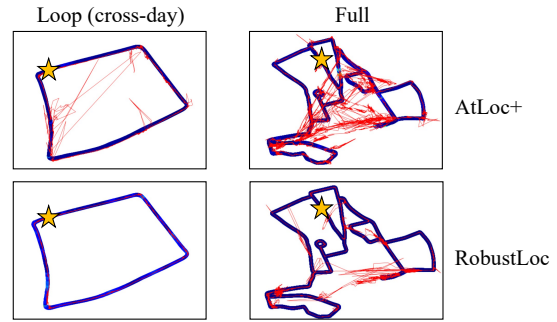


Figure 6: Trajectory visualization on the Oxford RobotCar dataset. The ground truth trajectories are shown in bold blue lines, and the estimated trajectories are shown in thin red lines. The stars mark the start of the trajectories.

we test the performance when training with less data augmentation, which leads to worse performance. This suggests that more extensive data augmentation can enhance the model robustness in challenging scenarios, which is consistent with the experimental results on the Perturbed Robotcar dataset in Table 3.

Graph Design. We next explore the use of different graph designs for feature map diffusion and vector embedding diffusion. The grid graph stacks an image with two other spatially adjacent images as a cube, and the attention weights are formulated within the 6-neighbor area (for feature maps) or the 2-neighbor area (for vector embeddings). The self-cross graph computes attention weights first within each image and then across different images. From Table 5, we see that the complete graph has the best performance. This is because, in the complete graph, each node can interact with all other nodes, allowing the aggregation of useful information with appropriate attention weights.

Rotation Representation. We compare different representations of rotation in Table 5, where the log form of the quaternion is the optimal choice. The other three representations, including the vanilla quaternion, the Lie group, and the vanilla rotation matrix, show similar performance.

Trajectory Visualization. We visualize the output pose trajectories as shown in Fig. 6, where a significant gap can be seen from the comparison. RobustLoc outputs more smooth and globally accurate poses compared with the previous method, which shows the effectiveness of our design.

Inference Speed. We finally test the performance using a different number of input frames. The inference speed does not drop significantly when increasing the input frames. And even the slowest one (using 11 frames) can run 50 iterations per second and achieve real-time regression. On the other hand, more frames can bring performance improvement when the input size is small, while further increasing frame size does not bring significant change.

Model	Business Campus		Neighborhood		Old Town	
	Mean	Median	Mean	Median	Mean	Median
GeoPoseNet	11.04 / 5.78	5.93 / 2.03	2.87 / 1.30	1.92 / 0.88	64.81 / 6.67	15.03 / 1.57
MapNet	10.35 / 3.78	5.66 / 1.83	2.81 / 1.05	1.89 / 0.92	46.56 / 7.14	16.52 / 2.12
GNNMapNet	<u>7.69</u> / 4.34	<u>5.52</u> / 2.16	3.02 / 2.92	2.14 / 1.45	<u>41.54</u> / 7.30	19.23 / 3.26
AtLoc	11.53 / 4.84	5.81 / <u>1.50</u>	2.80 / 1.16	1.83 / 0.93	84.17 / 7.81	17.10 / 1.73
AtLoc+	13.70 / 6.41	5.58 / 1.94	2.33 / 1.39	1.61 / 0.88	68.40 / 5.51	14.52 / 1.69
IRPNet	10.95 / 5.38	5.91 / 1.82	3.17 / 2.85	1.98 / 0.90	55.86 / 6.97	17.33 / 3.11
CoordiNet	11.52 / <u>3.44</u>	6.44 / 1.38	<u>1.72</u> / <u>0.86</u>	<u>1.37</u> / <u>0.69</u>	43.68 / <u>3.58</u>	<u>11.83</u> / <u>1.36</u>
RobustLoc (ours)	4.28 / 2.04	2.55 / <u>1.50</u>	1.36 / 0.83	1.00 / 0.65	21.65 / 2.41	5.52 / 1.05

Table 2: Median and mean translation/rotation estimation error (m°) on the 4Seasons dataset. The best and the second-best results in each metric are highlighted with bold and underline respectively.

Model	Medium		Hard		Hard (+ noisy training)	
	Mean	Median	Mean	Median	Mean	Median
GeoPoseNet	20.47 / 8.76	8.70 / 2.30	41.71 / 17.63	14.02 / 3.13	24.03 / 11.14	7.14 / 1.70
MapNet	17.93 / 7.01	6.89 / 2.00	49.36 / 20.01	18.37 / <u>2.58</u>	21.22 / 8.38	6.38 / 1.97
GNNMapNet	<u>16.17</u> / 7.24	8.02 / 2.35	73.97 / 35.57	61.47 / 19.73	<u>14.55</u> / <u>7.62</u>	6.69 / <u>1.57</u>
AtLoc	19.92 / 7.25	7.26 / <u>1.74</u>	52.56 / 23.46	15.01 / 3.17	23.48 / 11.43	7.42 / 2.38
AtLoc+	17.68 / 7.48	<u>6.19</u> / 1.80	<u>37.92</u> / 18.65	<u>12.17</u> / 2.93	22.61 / 11.23	<u>6.21</u> / 1.83
IRPNet	16.35 / 7.56	8.71 / 2.28	45.72 / 21.84	17.99 / 3.50	24.73 / 11.20	6.73 / 1.82
CoordiNet	17.67 / <u>6.66</u>	7.63 / 1.79	44.11 / <u>16.42</u>	17.21 / 2.70	24.06 / 12.27	6.25 / 1.61
RobustLoc (ours)	8.12 / 3.83	5.34 / 1.53	27.75 / 9.70	11.59 / 2.64	10.06 / 4.95	5.18 / 1.43

Table 3: Median and mean translation/rotation estimation error (m°) on the Perturbed RobotCar dataset. The best and the second-best results in each metric are highlighted with bold and underline respectively. RobustLoc achieves the best in all metrics.

Method	Mean Error (m°) on Loop (c.)
base model	8.38 / 4.29
+ feature map graph	7.01 / 3.86
+ vector embedding graph	6.24 / 3.21
+ diffusion	5.53 / 2.95
+ branched decoder	5.14 / 2.79
+ multi-level decoding	4.68 / 2.67
diffusion at stage 3	5.27 / 2.90
diffusion at stage 3,4	4.86 / 3.18
diffusion at stage 4	4.68 / 2.67
multi-layer concatenation	5.80 / 3.26
more augmentation	4.68 / 2.67
less augmentation	5.32 / 3.17

Table 4: Ablation study, diffusion design, and augmentation design comparison on the Oxford RobotCar dataset.

Conclusion

We have proposed and verified the performance of a robust CPR model RobustLoc. The model’s robustness derives from the use of information from covisible images and neural graph diffusion to aggregate neighboring information, which is present in challenging driving environments.

Acknowledgments

This work is supported under the RIE2020 Industry Alignment Fund-Industry Collaboration Projects (IAF-ICP) Fund-

Method	Mean Error (m°) on Full
grid graph	15.67 / 2.95
self-cross graph	15.31 / 3.28
complete graph	9.37 / 2.47
Mean Error ($^\circ$) on Business Campus	
quaternion	2.23
Lie group	2.20
rotation matrix	2.25
log (quaternion)	2.04

Table 5: Graph design comparison on the Oxford RobotCar dataset and rotation representation comparison on the 4Seasons dataset.

#frames	3	5	7	9	11
Speed (iters/s)	56	55	53	52	50
Mean Error (m)	5.28	5.09	4.96	4.68	4.72

Table 6: The performance using different numbers of frames on the Oxford RobotCar Loop (cross-day).

ing Initiative, as well as cash and in-kind contribution from the industry partner(s), and by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme.

References

- Brahmbhatt, S.; Gu, J.; Kim, K.; Hays, J.; and Kautz, J. 2018. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2616–2625.
- Chamberlain, B. P.; Rowbottom, J.; Eynard, D.; Di Giovanni, F.; Xiaowen, D.; and Bronstein, M. M. 2021a. Beltrami Flow and Neural Diffusion on Graphs. In *Proc. Advances Neural Inf. Process. Syst.*
- Chamberlain, B. P.; Rowbottom, J.; Goronova, M.; Webb, S.; Rossi, E.; and Bronstein, M. M. 2021b. GRAND: Graph Neural Diffusion. In *Proc. Int. Conf. Mach. Learn.*
- Chen, R. T.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. 2018. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*.
- Chen, Z.-Q.; Qian, Z.; Hu, Y.; and Zheng, W. 1998. Stability and approximations of symmetric diffusion semigroups and kernels. *J. functional anal.*, 152(1): 255–280.
- Hu, Y.; Wang, B.; and Lin, S. 2017. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4085–4094.
- Huang, Z.; Xu, Y.; Shi, J.; Zhou, X.; Bao, H.; and Zhang, G. 2019. Prior guided dropout for robust visual localization in dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2791–2800.
- Kang, Q.; Song, Y.; Ding, Q.; and Tay, W. P. 2021. Stable neural ODE with Lyapunov-stable equilibrium points for defending against adversarial attacks. In *Proc. Advances Neural Inf. Process. Syst.*
- Kendall, A.; and Cipolla, R. 2017. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5974–5983.
- Kendall, A.; Grimes, M.; and Cipolla, R. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, 2938–2946.
- Liu, R.; Lehman, J.; Molino, P.; Petroski Such, F.; Frank, E.; Sergeev, A.; and Yosinski, J. 2018. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31.
- Maddern, W.; Pascoe, G.; Linegar, C.; and Newman, P. 2017. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research*, 36(1): 3–15.
- Moreau, A.; Piasco, N.; Tsishkou, D.; Stanciulescu, B.; and de La Fortelle, A. 2022. CoordiNet: uncertainty-aware pose regressor for reliable vehicle localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2229–2238.
- Sarlin, P.-E.; Unagar, A.; Larsson, M.; Germain, H.; Toft, C.; Larsson, V.; Pollefeys, M.; Lepetit, V.; Hammarstrand, L.; Kahl, F.; et al. 2021. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3247–3257.
- Shavit, Y.; and Ferens, R. 2021. Do We Really Need Scene-specific Pose Encoders? In *2020 25th International Conference on Pattern Recognition (ICPR)*, 3186–3192. IEEE.
- Song, Y.; Kang, Q.; Wang, S.; Zhao, K.; and Tay, W. P. 2022. On the Robustness of Graph Neural Diffusion to Topology Perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*. New Orleans, USA.
- Walch, F.; Hazirbas, C.; Leal-Taixe, L.; Sattler, T.; Hilsenbeck, S.; and Cremers, D. 2017. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, 627–637.
- Wang, B.; Chen, C.; Lu, C. X.; Zhao, P.; Trigoni, N.; and Markham, A. 2020. Atloc: Attention guided camera localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10393–10401.
- Wenzel, P.; Wang, R.; Yang, N.; Cheng, Q.; Khan, Q.; von Stumberg, L.; Zeller, N.; and Cremers, D. 2020. 4Seasons: A cross-season dataset for multi-weather SLAM in autonomous driving. In *DAGM German Conference on Pattern Recognition*, 404–417. Springer.
- Xue, F.; Wang, X.; Yan, Z.; Wang, Q.; Wang, J.; and Zha, H. 2019. Local supports global: Deep camera relocalization with sequence enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2841–2850.
- Xue, F.; Wu, X.; Cai, S.; and Wang, J. 2020. Learning multi-view camera relocalization with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11372–11381. IEEE.
- Yan, H.; Du, J.; Tan, V. Y.; and Feng, J. 2018. On robustness of neural ordinary differential equations. In *Proc. Advances Neural Inf. Process. Syst.*, 1–13.
- Yan, J.; Zhao, H.; Bu, P.; and Jin, Y. 2021. Channel-Wise Attention-Based Network for Self-Supervised Monocular Depth Estimation. In *2021 International Conference on 3D Vision (3DV)*, 464–473. IEEE.