KFNet: Learning Temporal Camera Relocalization using Kalman Filtering

Lei Zhou\textsuperscript{1} Zixin Luo\textsuperscript{1} Tianwei Shen\textsuperscript{1} Jiahui Zhang\textsuperscript{2}
Mingmin Zhen\textsuperscript{1} Yao Yao\textsuperscript{1} Tian Fang\textsuperscript{3} Long Quan\textsuperscript{1}

\textsuperscript{1}Hong Kong University of Science and Technology \hspace{1em} \textsuperscript{2}Tsinghua University \hspace{1em} \textsuperscript{3}Everest Innovation Technology

\{lzhouai, zluoaag, tshenaa, mzhen, yyaoag, quan\}@cse.ust.hk
\textsuperscript{1}jiahui-z15@mails.tsinghua.edu.cn \hspace{1em} \textsuperscript{3}fangtian@altizure.com

Abstract

Temporal camera relocalization estimates the pose with respect to each video frame in sequence, as opposed to one-shot relocalization which focuses on a still image. Even though the time dependency has been taken into account, current temporal relocalization methods still generally underperform the state-of-the-art one-shot approaches in terms of accuracy. In this work, we improve the temporal relocalization method by using a network architecture that incorporates Kalman filtering (KFNet) for online camera relocalization. In particular, KFNet extends the scene coordinate regression problem to the time domain in order to recursively establish 2D and 3D correspondences for the pose determination. The network architecture design and the loss formulation are based on Kalman filtering in the context of Bayesian learning. Extensive experiments on multiple relocalization benchmarks demonstrate the high accuracy of KFNet at the top of both one-shot and temporal relocalization approaches. Our codes are released at https://github.com/zl thinker/KFNet.

1. Introduction

Camera relocalization serves as the subroutine of applications including SLAM [15], augmented reality [9] and autonomous navigation [45]. It estimates the 6-DoF pose of a query RGB image in a known scene coordinate system. Current relocalization approaches mostly focus on one-shot relocalization for a still image. They can be mainly categorized into three classes [13, 50]: (1) the relative pose regression (RPR) methods which determine the relative pose w.r.t. the database images [3, 29], (2) the absolute pose regression (APR) methods regressing the absolute pose through PoseNet [25] and its variants [23, 24, 60] and (3) the structure-based methods that establish 2D-3D correspondences with Active Search [48, 49] or Scene Coordinate Regression (SCoRe) [52] and then solve the pose by PnP algorithms [18, 42]. Particularly, SCoRe is widely adopted recently to learn per-pixel scene coordinates from dense training data for a scene, due to its ability to form dense and accurate 2D-3D matches even in texture-less scenes [5, 6]. As extensively evaluated in [5, 6, 50], the structure-based methods generally show better pose accuracy than the RPR and APR methods, because they explicitly exploit the rules of the projective geometry and the scene structures [50].

Apart from one-shot relocalization, temporal relocalization with respect to video frames is also worthy of investigation. However, almost all the temporal relocalization methods are based on PoseNet [25], which, in general, even underperform the structure-based one-shot methods in accuracy. This is mainly because their accuracies are fundamentally limited by the retrieval nature of PoseNet. As analyzed in [50], PoseNet based methods are essentially analogous to approximate pose estimation via image retrieval, and cannot go beyond the retrieval baseline in accuracy.

In this work, we are motivated by the high accuracy of structure-based relocalization methods and resort to SCoRe to estimate per-pixel scene coordinates for pose computation. Besides, we propose to extend SCoRe to the time domain in a recursive manner to enhance the temporal consistency of 2D-3D matching, thus allowing for more accurate online pose estimations for sequential images. Specifically, a recurrent network named KFNet is proposed in the context of Bayesian learning [37] by embedding SCoRe into the Kalman filter within a deep learning framework. It is composed of three subsystems below, as illustrated in Fig. 1.

- **The measurement system** features a network termed SCoordNet to derive the maximum likelihood (ML) predictions of the scene coordinates for a single image.
- **The process system** uses OFlowNet that models the optical flow based transition process for image pixels across time steps and yields the prior predictions of scene coordinates. Additionally, the measurement and process systems provide uncertainty predictions [40, 23] to model the noise dynamics over time.
- **The filtering system** fuses both predictions and leads to the maximum a posteriori (MAP) estimations of the final scene coordinates.
Furthermore, we propose probabilistic losses for the three subsystems based on the Bayesian formulation of KFNet, to enable the training of either the subsystems or the full framework. We summarize the contributions as follows.

- We are the first to extend the scene coordinate regression problem \([52]\) to the time domain in a learnable way for temporally-consistent 2D-3D matching.
- We integrate the traditional Kalman filter \([22]\) into a recurrent CNN network (KFNet) that resolves pixel-level state inference over time-series images.
- KFNet bridges the existing performance gap between temporal and one-shot relocalization approaches, and achieves top accuracy on multiple relocalization benchmarks \([52, 57, 25, 43]\).
- Lastly, for better practicality, we propose a statistical assessment tool to enable KFNet to self-inspect the potential outlier predictions on the fly.

2. Related Works

Camera relocalization. We categorize camera relocalization algorithms into three classes: the relative pose regression (RPR) methods, the absolute pose regression (APR) methods and the structure-based methods.

The RPR methods use a coarse-to-fine strategy which first finds similar images in the database through image retrieval \([55, 2]\) and then computes the relative poses w.r.t. the retrieved images \([3, 29, 46]\). They have good generalization to unseen scenes, but the retrieval process needs to match the query image against all the database images, which can be costly for time-critical applications.

The APR methods include PoseNet \([25]\) and its variants \([23, 24, 60]\) which learn to regress the absolute camera poses from the input images through a CNN. They are simple and efficient, but generally fall behind the structure-based methods in terms of accuracy, as validated by \([5, 6, 50]\). Theoretically, \([50]\) explains that PoseNet-based methods are more closely related to image retrieval than to accurate pose estimation via 3D geometry.

The structure-based methods explicitly establish the correspondences between 2D image pixels and 3D scene points and then solve camera poses by PnP algorithms \([18, 42, 30]\). Traditionally, correspondences are searched by matching the patch features against Structure from Motion (SfM) tracks via Active Search \([48, 49]\) and its variants \([32, 10, 33, 47]\), which can be inefficient and fragile in texture-less scenarios. Recently, the correspondence problem is resolved by predicting the scene coordinates for pixels by training random forests \([52, 58, 36]\) or CNNs \([5, 6, 31, 7]\) with ground truth scene coordinates, which is referred to as Scene Coordinate Regression (SCoRe).

Besides one-shot relocalization, some works have extended PoseNet to the time domain to address temporal re-localization. VidLoc \([11]\) performs offline and batch re-localization for fixed-length video-clips by BLSTM \([51]\). Coskun et al. refine the pose dynamics by embedding LSTM units in the Kalman filters \([12]\). VLocNet \([56]\) and VLocNet++ \([43]\) propose to learn pose regression and the visual odometry jointly. LSG \([63]\) combines LSTM with visual odometry to further exploit the spatial-temporal consistency. Since all the methods are extensions of PoseNet, their accuracies are fundamentally limited by the retrieval nature of PoseNet, following the analysis of \([50]\).

Temporal processing. When processing time-series image data, ConvLSTM \([61]\) is a standard way of modeling the spatial correlations of local contexts through time \([59, 35, 28]\). However, some works have pointed out that the implicit convolutional modeling is less suited to discovering the pixel associations between neighboring frames, especially when pixel-level accuracy is desired \([21, 39]\). Therefore, in later works, the optical flow is highlighted as a more explicit way of delineating the pixel correspondences across sequential steps \([41]\). For example, \([41, 20, 28, 53, 39]\) commonly predict the optical flow fields to guide the feature map warping across time steps. Then, the warped features are fused by weighting \([65, 66]\) or pooling \([38, 41]\) to aggregate the temporal knowledge. In this work, we follow the practice of flow-guided warping, but the distinction from previous works is that we propose to fuse the predictions by leveraging Kalman filter principles \([37]\).

3. Bayesian Formulation

This section presents the Bayesian formulation of recursive scene coordinate regression in the time domain for temporal camera relocalization. Based on the formulation, the proposed KFNet is built and the probabilistic losses are defined in Sec. 4 ~ 6. Notations used below have been summarized in Table 1 for quick reference.

Given a stream of RGB images up to time \(t\), i.e., \(\mathcal{I}_t = \{\mathcal{I}_1, ..., \mathcal{I}_{t-1}, \mathcal{I}_t\}\), our aim is to predict the latent state for
each frame, i.e., the scene coordinate map, which is then used for pose computation. We denote the map as $\theta_t \in \mathbb{R}^{N \times 3}$, where $N$ is the pixel number. By imposing the Gaussian noise assumption on the states, the state $\theta_t$ conditioned on $I_t$ follows an unknown Gaussian distribution:

$$\theta_t^+ \triangleq (\theta_t | I_t) \sim \mathcal{N}(\hat{\theta}_t, \Sigma_t),$$  \hspace{1cm} (1)

where $\hat{\theta}_t$ and $\Sigma_t$ are the expectation and covariance to be determined. Under the routine of Bayesian theorem, the posterior probability of $\theta_t$ can be factorized as

$$P(\theta_t | I_t) \propto P(\theta_t | I_t, \bar{I}_t) P(I_t | \theta_t, \bar{I}_{t-1}),$$  \hspace{1cm} (2)

where $\bar{I}_t = I_{t-1} \cup \{I_t\}$.

The first factor $P(\theta_t | I_t, \bar{I}_{t-1})$ of the right hand side (RHS) of Eq. 2 indicates the prior belief about $\theta_t$ obtained from time $t - 1$ through a process system. Provided that no occlusions or dynamic objects occur, the consecutive coordinate maps can be approximately associated by a linear process equation describing their pixel correspondences, wherein

$$\theta_t = G_t \theta_{t-1} + w_t,$$  \hspace{1cm} (3)

with $G_t \in \mathbb{R}^{N \times N}$ being the sparse state transition matrix given by the optical flow fields from time $t - 1$ to $t$, and $w_t \sim \mathcal{N}(0, W_t)$. $W_t \in \mathbb{S}^{N \times 1}_{++}$ being the process noise. Given $I_{t-1}$, we already have the probability statement that $(\theta_{t-1} | I_{t-1}) \sim \mathcal{N}(\hat{\theta}_{t-1}, \Sigma_{t-1})$. Then the prior estimation of $\theta_t$ from time $t - 1$ can be expressed as

$$\hat{\theta}_t^- \triangleq (\theta_t | \bar{I}_{t-1}) \sim \mathcal{N}(\hat{\theta}_t^-, R_t),$$  \hspace{1cm} (4)

where $\hat{\theta}_t^- = G_t \hat{\theta}_{t-1}, R_t = G_t \Sigma_{t-1} G_t^T + W_t$.

The second factor $P(I_t | \theta_t, \bar{I}_{t-1})$ of the RHS of Eq. 2 describes the likelihood of image observations at time $t$ made through a measurement system. The system models how $I_t$ is derived from the latent states $\theta_t$, formally $I_t = h(\theta_t)$. However, the high nonlinearity of $h(\cdot)$ makes the following computation intractable. Alternatively, we map $I_t$ to $z_t \in \mathbb{R}^{N \times 3}$ via a nonlinear function inspired by [12], so that the system can be approximately expressed by a linear measurement equation:

$$z_t = \theta_t + v_t,$$  \hspace{1cm} (5)

where $v_t \sim \mathcal{N}(0, V_t)$. $V_t \in \mathbb{S}^{N}_{++}$ denotes the measurement noise, and $z_t$ can be interpreted as the noisy observed scene coordinates. In this way, the likelihood can be rewritten as $P(z_t | \theta_t, \bar{I}_{t-1})$ by substituting $z_t$ for $I_t$.

Let $e_t$ denote the residual of predicting $z_t$ from time $t - 1$; thus

$$e_t = z_t - \hat{\theta}_t^- = z_t - G_t \hat{\theta}_{t-1}.$$  \hspace{1cm} (6)

Since $G_t$ and $\hat{\theta}_{t-1}$ are all known, observing $z_t$ is equivalent to observing $e_t$. Hence, the likelihood $P(z_t | \theta_t, \bar{I}_{t-1})$ can be rewritten as $P(e_t | \theta_t, \bar{I}_{t-1})$. Substituting Eq. 5 into Eq. 6, we have $e_t = \theta_t - \theta_t^- + v_t$, so that the likelihood can be described by

$$(e_t | \theta_t, \bar{I}_{t-1}) \sim \mathcal{N}(0, V_t).$$  \hspace{1cm} (7)

Based on the theorems in multivariate statistics [1, 37], combining the two distributions 4 & 7 gives the bivariate normal distribution:

$$\begin{bmatrix} \theta_t \\ e_t \end{bmatrix} | \bar{I}_{t-1} \sim \mathcal{N} \begin{bmatrix} \hat{\theta}_t^- \\ 0 \end{bmatrix}, \begin{bmatrix} R_t & R_t \\ R_t & R_t + V_t \end{bmatrix}.$$  \hspace{1cm} (8)

Making $e_t$ the conditioning variable, the filtering system gives the posterior distribution that writes

$$\theta_t^+ \triangleq (\theta_t | \bar{I}_t) = (\theta_t | e_t, \bar{I}_{t-1}) \sim \mathcal{N}(\hat{\theta}_t^+, \Sigma_t)$$

$$\sim \mathcal{N}(\hat{\theta}_t^+, K_t e_t, R_t (I - K_t),$$

where $K_t = \frac{R_t}{V_t + R_t}$ is conceptually referred to as the Kalman gain and $e_t$ as the innovation$^2$ [37, 19].

As shown in Fig. 1, the inference of the posterior scene coordinates $\hat{\theta}_t$ and covariance $\Sigma_t$ for image pixels proceeds recursively as the time $t$ evolves, which are then used for online pose determination. Specifically, the pixels with variances greater than $\lambda$ are first excluded as outliers. Then, a RANSAC+P3P [18] solver is applied to compute the initial camera pose from the 2D-3D correspondences, followed by a nonlinear optimization for pose refinement.

### 4. The Measurement System

The measurement system is basically a generative model explaining how the observations $z_t$ are generated from the latent scene coordinates $\theta_t$, as expressed in Eq. 5. Then, the remaining problem is to learn the underlying mapping from $I_t$ to $z_t$. This is similar to the SCORe task [52, 5, 6], but differs in the constraint about $z_t$ imposed by Eq. 5. Below, the

---

$^1\mathbb{S}^{N}_{++}$ denotes the set of $N$-dimensional positive definite matrices.

$^2$The derivation of Eqs. 8 & 9 is shown in the supplementary material.
architecture of SCoordNet is first introduced, which outputs the scene coordinate predictions, along with the uncertainties, to model the measurement noise $\mathbf{v}_t$. Then, we define the probabilistic loss based on the likelihood $P(\mathbf{z}_t | \theta_t, I_{t-1})$ of the measurement system.

### 4.1. Architecture

SCoordNet shares the similar fully convolutional structure to [6], as shown in Fig. 1. However, it is far more lightweight, with parameters fewer than one eighth of [6]. It encompasses twelve $3 \times 3$ convolution layers, three of which use a stride of 2 to downsize the input by a factor of 8. ReLU follows each layer except the last one. To simplify computation and avoid the risk of over-parameterization, we postulate the isotropic covariance of the multivariate Gaussian measurement noise, i.e., $\mathbf{V}(i) = v(i)^2 \mathbb{I}_3$ for each pixel $\mathbf{p}_i$, where $\mathbb{I}_3$ denotes the $3 \times 3$ identity matrix. The output thus has a channel of 4, comprising 3-d scene coordinates and a 1-d uncertainty measurement.

### 4.2. Loss

According to Eq. 5, the latent scene coordinates $\theta(i)$ of pixel $\mathbf{p}_i$ should follow the distribution $\mathcal{N}(\mathbf{z}(i), v(i)^2 \mathbb{I}_3)$. Taking the negative logarithm of the probability density function (PDF) of $\theta(i)$, we define the loss based on the likelihood which gives rise to the maximum likelihood (ML) estimation for each pixel in the form [23]:

$$ L_{\text{likelihood}} = \sum_{i=1}^{N} \left( 3 \log v(i) + \frac{||\mathbf{z}(i) - \mathbf{y}(i)||^2}{2v(i)^2} \right), $$

where $\mathbf{y}(i)$ being the groundtruth label for $\theta(i)$. For numerical stability, we use logarithmic variance for the uncertainty measurements in practice, i.e., $\sigma^2(i) = \log v(i)^2$.

Including uncertainty learning in the loss formulation allows one to quantify the prediction errors stemming not just from the intrinsic noise in the data but also from the defined model [14]. For example, at the boundary with depth discontinuity, a sub-pixel offset would cause an abrupt coordinate shift which is hard to model. SCoordNet would easily suffer from a significant magnitude of loss in such cases. It is sensible to automatically downplay such errors during training by weighting with the uncertainty measurements. Fig. 2(a) illustrates the uncertainty predictions in such cases.

### 5. The Process System

The process system models the transition process of pixel states from time $t - 1$ to $t$, as described by the process equation of Eq. 3. Herein, first, we propose a cost volume based network, OFlowNet, to predict the optical flows and the process noise covariance jointly for each pixel. Once the optical flows are determined, Eq. 3 is equivalent to the flow-guided warping from time $t - 1$ towards $t$, as commonly used in [41, 20, 28, 53, 39]. Second, after the warping, the prior distribution of the states, i.e., $\theta_t \sim \mathcal{N}(\hat{\theta}_t, \mathbf{R}_t)$ of Eq. 4, can be evaluated. We then define the probabilistic loss based on the prior to train OFlowNet.

### 5.1. Architecture

OFlowNet is composed of two components: the cost volume constructor and the flow estimator.

The cost volume constructor first extracts features from the two input images $I_{t-1}$ and $I_t$ respectively through seven $3 \times 3$ convolutions, three of which have a stride of 2. The output feature maps $F_{t-1}$ and $F_t$ have a spatial size of one-eighth of the inputs and a channel number of $c$. Then, we build up a cost volume $C_t \in \mathbb{R}^{c \times w \times w}$ for each pixel $\mathbf{p}_i$ of the feature map $F_t$, so that

$$ C_t(o) = \frac{\mathbf{F}_t(\mathbf{p}_i) - \mathbf{F}_{t-1}(\mathbf{p}_i + o)}{\mathbf{F}_{t-1}(\mathbf{p}_i + o)}, $$

where $w$ is the size of the search window which corresponds to $8w$ pixels in the full-resolution image, and $o \in \{ -w/2, \ldots, w/2 \}^2$ is the spatial offset. We apply L2-normalization to the feature maps along the channel dimension before differentiation, as in [62, 34].

The following flow estimator operates over the cost volumes for flow inference. We use a U-Net with skip connections [44] as shown in Fig. 1, which first subsamples the cost volume by a factor of 8 for an enlarged receptive field and then upsamples it to the original resolution. The output is a $w \times w \times 1$ unbounded confidence map for each pixel. Related works usually attain flows by hard assignment based on the matching cost encapsulated by the cost volumes [62, 54]. However, it would cause non-differentiability in later steps where the optical flows are to be further used for spatial warping. Thus, we pass the confidence map through the differentiable spatial softmax operator [16] to compute the optical flow as the expectation of the pixel offsets inside the search window. Formally,

$$ \hat{o} \overset{\text{def}}{=} \mathbb{E}(o) = \sum_o \text{softmax}(f(o)) \cdot o, $$

4922
where \( f_o \) is the confidence at offset \( o \). To fulfill the process noise modeling, i.e., \( w_i \) in Eq. 3, we append three fully connected layers after the bottleneck of the U-Net to regress the logarithmic variance, as shown in Fig. 1. Sample optical flow predictions are visualized in Fig. 3.

5.2. Loss

Once the optical flows are computed, the state transition matrix \( G_t \) of Eq. 3 can be evaluated. We then complete the linear transition process of Eq. 3 by warping the scene coordinate map and uncertainty map from time \( t-1 \) towards \( t \) through bilinear warping [64]. Let \( \hat{\theta}_i \) and \( \sigma_i^{-2} \) be the warped scene coordinates and Gaussian variance, and \( w_i^{-2} \) be the Gaussian variance of the process noise of pixel \( p_i \) at time \( t \). Then, the prior coordinates of \( p_i \), denoted as \( \theta_i \), should follow the distribution

\[
\theta_i \sim \mathcal{N}(\hat{\theta}_i, r_i^{-2}I_3),
\]

where \( r_i^{-2} = \sigma_i^{-2} + w_i^{-2} \). Taking the negative logarithm of the PDF of \( \theta_i \), we get the loss of the process system as

\[
L_{\text{prior}} = \sum_{i=1}^{N} \left( 3 \log r_i + \frac{\| \theta_i - y_i \|^2}{2r_i^2} \right). \quad (14)
\]

It is noteworthy that the loss definition uses the prior distribution of \( \theta_i \) to provide the weak supervision for training OFlowNet, with no recourse to the optical flow labeling.

One issue with the proposed process system is that it assumes no occurrence of occlusions or dynamic objects which are two outstanding challenges for tracking problems [27, 67]. Our process system partially addresses the issue by giving the uncertainty measurements of the process noise. As shown in Fig. 2(b), OFlowNet generally produces much larger uncertainty estimations for the pixels from occluded areas and dynamic objects. This helps to give lower weights to these pixels that have incorrect flow predictions in the loss computation.

6. The Filtering System

The measurement and process systems in the previous two sections have derived the likelihood and prior estimations of the scene coordinates \( \theta_i \), respectively. The filtering system aims to fuse both of them based on Eq. 9 to yield the posterior estimation.

6.2. Consistency Examination

In practice, the filter could behave incorrectly due to the outlier estimations caused by the erratic scene coordinate regression or a failure of flow tracking. This would induce accumulated state errors in the long run. Therefore, we use the statistical assessment tool, Normalized Innovation Squared (NIS) [4], to filter the inconsistent predictions during inference.

Normally, the innovation variable \( e_i \in \mathbb{R}^3 \) follows the Gaussian distribution \( \mathcal{N}(0, S_i) \) as shown by Eq. 8, where

\[
L_{\text{full}} = \tau_1 L_{\text{likelihood}} + \tau_2 L_{\text{prior}} + \tau_3 L_{\text{posterior}}. \quad (18)
\]
we train SCoordNet for each scene with the likelihood loss and exponential decay. The training procedure has 3 stages. First, we use the critical value of χ²(9) as the acceptance region of the pixel state as an outlier if its NIS value locates outside. The uncertainties of the pixels failing the statistical evidence to regard one pixel as an outlier. The ADAM optimizer is used with β₁ = 0.9 and β₂ = 0.999. We use an initial learning rate of γ = 0.0001 and then drop it with exponential decay. The training procedure has 3 stages. First, we train SCoordNet for each scene with the likelihood loss L_{likelihood} (Eq. 10). The iteration number is set to be proportional to the surface area of each scene and the learning rate drops from γ to γ / 2. In particular, we use SCoordNet as the one-shot version of the proposed approach. Second, OFlowNet is trained using all the scenes for each dataset with the prior loss L_{prior} (Eq. 14). It also experiences the learning rate decaying from γ to γ / 2. Each batch is composed of two consecutive frames. The window size of OFlowNet in the original images is set to 64, 128, 192 and 256 for the four datasets mentioned above, respectively, due to the increasing ego-motion through them. Third, we fine-tune all the parameters of KFNet jointly by optimizing the full loss L_{full} (Eq. 18) with a learning rate going from γ / 2 to γ / 2. Each batch in the third stage contains four consecutive frames.

7.2. Results

7.2.1 The Relocalization Accuracy

Following [5, 6, 11, 56], we use two accuracy metrics: (1) the median translation and translation error of poses (see Table 2); (2) the 5cm-5deg accuracy (see Table 3), i.e., the mean percentage of the poses with translation and rotation errors less than 5 cm and 5°, respectively. The uncertainty threshold λ (Sec. 3) is set to 5 cm for 7scenes and 12scenes and 50 cm for DeepLoc and Cambridge.

One-shot relocalization. Our SCoordNet achieves the lowest pose errors on 7scenes and Cambridge, and the highest 5cm-5deg accuracy on 12scenes among the one-shot methods, surpassing CamNet [13] and MapNet [8] which are the state-of-the-art relative and absolute pose regression methods, respectively. Particularly, SCoordNet outperforms the state-of-the-art structure-based methods DSAC++ [6] and ESAC [7], yet with fewer parameters (24M vs. 210M vs. 28M, respectively). The advantage of SCoordNet should be mainly attributed to the uncertainty modeling, as we will analyze in the supplementary material. It also surpasses Active Search (AS) [49] on 7scenes and Cambridge, but underperforms AS on DeepLoc. We find that, in the experiments

<table>
<thead>
<tr>
<th>Scenes</th>
<th>One-shot Relocalization</th>
<th>Temporal Relocalization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Active Search [49]</td>
<td>DSAC ++ [57]</td>
</tr>
<tr>
<td></td>
<td>VILOCNet++ [12]</td>
<td>LSG [63]</td>
</tr>
<tr>
<td></td>
<td>KFNet (Ours)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>98.9%</td>
<td>97.9%</td>
</tr>
<tr>
<td></td>
<td>98.9%</td>
<td>97.9%</td>
</tr>
<tr>
<td></td>
<td>98.9%</td>
<td>97.9%</td>
</tr>
<tr>
<td></td>
<td>98.9%</td>
<td>97.9%</td>
</tr>
</tbody>
</table>

Table 3: The 5cm-5deg accuracy of one-shot and temporal relocalization methods on 12scenes [57].

S_{i}\left(\tau_{i}\right) = \left(\frac{v_{i}}{\tau_{i}} + r_{i}\right)^{2}. Then, NIS = e_{i}^{T}S_{i}^{-1}e_{i} is supposed to follow the Chi-squared distribution with three degrees of freedom, denoted as χ²(3). It is thus reasonable to see a pixel state as an outlier if its NIS value locates outside the acceptance region of χ²(3). As illustrated in Fig. 4, we use the critical value of 0.05 in the NIS test, which means we have at least 95% statistical evidence to regard one pixel state as negative. The uncertainties of the pixels failing the test, i.e., ŷ_{i}, are reset to be infinitely large so that they will have no effect in later steps.

7.2.1 The Relocalization Accuracy

Following [5, 6, 11, 56], we use two accuracy metrics: (1) the median translation and translation error of poses (see Table 2); (2) the 5cm-5deg accuracy (see Table 3), i.e., the mean percentage of the poses with translation and rotation errors less than 5 cm and 5°, respectively. The uncertainty threshold λ (Sec. 3) is set to 5 cm for 7scenes and 12scenes and 50 cm for DeepLoc and Cambridge.

One-shot relocalization. Our SCoordNet achieves the lowest pose errors on 7scenes and Cambridge, and the highest 5cm-5deg accuracy on 12scenes among the one-shot methods, surpassing CamNet [13] and MapNet [8] which are the state-of-the-art relative and absolute pose regression methods, respectively. Particularly, SCoordNet outperforms the state-of-the-art structure-based methods DSAC++ [6] and ESAC [7], yet with fewer parameters (24M vs. 210M vs. 28M, respectively). The advantage of SCoordNet should be mainly attributed to the uncertainty modeling, as we will analyze in the supplementary material. It also surpasses Active Search (AS) [49] on 7scenes and Cambridge, but underperforms AS on DeepLoc. We find that, in the experiments

To adapt this one-shot version to the full loss L_{full} (Eq. 18) with a learning rate going from γ / 2 to γ / 2. Each batch in the third stage contains four consecutive frames.
Figure 5: The point clouds predicted by different relocalization methods. Our SCoordNet and KFNet increasingly suppress the noise as highlighted by the red boxes and produce much neater point clouds than the state-of-the-art DSAC++ [6]. The KFNet-filtered panel filters out the points of KFNet of which the uncertainties are too large and gives rather clean and accurate mapping results.

Table 4: The mean and standard deviation of predicted scene coordinate errors in centimeters.

<table>
<thead>
<tr>
<th></th>
<th>7scenes</th>
<th>12scenes</th>
<th>DeepLoc</th>
<th>Cambridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>stddev</td>
<td>mean</td>
<td>stddev</td>
</tr>
<tr>
<td>DSAC++ [6]</td>
<td>28.8</td>
<td>33.1</td>
<td>28.8</td>
<td>41.1</td>
</tr>
<tr>
<td>SCoordNet</td>
<td>16.8</td>
<td>23.3</td>
<td>9.8</td>
<td>20.0</td>
</tr>
<tr>
<td>KFNet</td>
<td>15.3</td>
<td>21.7</td>
<td>7.3</td>
<td>13.7</td>
</tr>
</tbody>
</table>

Table 4: The mean and standard deviation of predicted scene coordinate errors in centimeters.

of AS on DeepLoc [50], AS is tested on a SfM model built with both training and test images. This may explain why AS is surprisingly more accurate on DeepLoc than on other datasets, since the 2D-3D matches between test images and SfM tracks have been established and their geometry has been optimized during the SfM reconstruction.

Temporal relocalization. Our KFNet improves over SCoordNet on all the datasets as shown in Tables 2 & 3. The improvement on Cambridge is marginal as the images are over-sampled from videos sparsely. The too large motions between frames make it hard to model the temporal correlations. KFNet obtains much lower pose errors than other temporal methods, except that it has a larger translation error than VLocNet++ [43] on 7scenes. However, the performance of VLocNet++ is inconsistent across different datasets. On DeepLoc, the dataset collected by the authors of VLocNet++, VLocNet++ has a much larger pose error than KFNet, even though it also integrates semantic segmentation into learning. The inconsistency is also observed in [50], which shows that VLocNet++ cannot substantially exceed the accuracy of retrieval based methods [55, 2].

7.2.2 The Mapping Accuracy

Relocalization methods based on SCoRe [52, 6] can create a mapping result for each view by predicting per-pixel scene coordinates. Hence, relocalization and mapping can be seen as dual problems, as one can be easily resolved once the other is known. Here, we would like to evaluate the mapping accuracy with the mean and the standard deviation (stddev) of scene coordinate errors of the test images.

As shown in Table 4, the mapping accuracy is in accordance with the relocalization accuracy reported in Sec. 7.2.1. SCoordNet reduces the mean and stddev values greatly compared against DSAC++, and KFNet further reduces the mean error over SCoordNet by 8.9%, 25.5%, 77.3% and 11.4% on the four datasets, respectively. The improvements are also reflected in the predicted point clouds, as visualized in Fig. 5. SCoordNet and KFNet predict less noisy scene points with better temporal consistency compared with DSAC++. Additionally, we filter out the points of KFNet with uncertainties greater than $\lambda$ as displayed in the KFNet-filtered panel of Fig. 5, which helps to give much neater and more accurate 3D point clouds.

7.2.3 Motion Blur Experiments

Although, in terms of the mean scene coordinate error in Table 4, SCoordNet outperforms DSAC++ by over 41.6% and KFNet further improves SCoordNet by a range from 8.9% to 77.3%, the improvements in terms of the median
pose error in Table 2 are not as significant. The main reason is that the RANSAC-based PnP solver diminishes the benefits brought by the scene coordinate improvements, since only a small subset of accurate scene coordinates selected by RANSAC matters in the pose accuracy. Therefore, to highlight the advantage of KFNet, we conduct more challenging experiments over motion blur images which are quite common in real scenarios. For the test image sequences of 7scenes, we apply a motion blur filter with a kernel size of 30 pixels for every 10 images as shown in Fig. 6(a). In Fig. 6(b)&(c), we plot the cumulative distribution functions of the pose errors before and after applying motion blur. Thanks to the uncertainty reasoning, SCoordNet generally attains smaller pose errors than DSAC++ whether motion blur is present. While SCoordNet and DSAC++ show a performance drop after motion blur is applied, KFNet maintain the pose accuracy as shown in Fig. 6(b)&(c), leading to a more notable margin between KFNet and SCoordNet and demonstrating the benefit of the temporal modelling used by KFNet.

7.3. Ablation studies

Evaluation of Temporal Aggregation. This section studies the efficacy of our Kalman filter based framework in comparison with other popular temporal aggregation strategies including ConvLSTM [61, 28], temporal pooler (TPooler) [41] and similarity weighting (SWeight) [65, 66]. KFNet is more related to TPooler and SWeight which also use the flow-guided warping yet within an n-frame neighborhood. For equitable comparison, the same feature network and probabilistic losses as KFNet are applied to all. We use a kernel size of 8 for ConvLSTM to ensure a window size of 64 in images. The same OFlowNet structure and a 3-frame neighborhood are used for TPooler and SWeight for flow-guided warping.

Table 5 shows the comparative results on 7scenes. ConvLSTM largely underperforms SCoordNet and other aggregation methods in pose accuracy, which manifests the necessity of explicitly determining the pixel associations between frames instead of implicit modeling. Although the flow-guided warping is employed, TPooler and SWeight only achieve marginal improvements over SCoordNet compared with KFNet, which justifies the advantage of the Kalman filtering system. Compared with TPooler and SWeight, the Kalman filter behaves as a more disciplined and non-heuristic approach to temporal aggregation that ensures an optimal solution of the linear Gaussian state-space model [17] defined in Sec. 3.

Evaluation of Consistency Examination Here, we explore the functionality of the consistency examination which uses NIS testing [4] (see Sec. 6.2). Due to the infrequent occurrence of extreme outlier predictions among the well-built relocalization datasets, we simulate the tracking lost situations by trimming a sub-sequence off each testing sequence of 7scenes and 12scenes. Let \( I_q \) and \( I_0 \) denote the last frame before and the first frame after the trimming. The discontinuous motion from \( I_q \) to \( I_0 \) would cause outlier scene coordinate predictions for \( I_q \) by KFNet. Fig. 7 plots the mean pose and scene coordinate errors of frames around \( I_q \) and visualizes the poses of a sample trimmed sequence. With the NIS test, the errors revert to a normal level promptly right after \( I_q \), whereas without the NIS test, the accuracy of poses after \( I_q \) is affected adversely. NIS testing stops the propagation of the outlier predictions of \( I_q \) into later steps by giving them infinitely large uncertainties, so that \( I_{q+1} \) will leave out the prior from \( I_q \) and reinitialize itself with the predictions of the measurement system.

8. Conclusion

This work addresses the temporal camera relocalization problem by proposing a recurrent network named KFNet. It extends the scene coordinate regression problem to the time domain for online pose determination. The architecture and the loss definition of KFNet are based on the Kalman filter, which allows a disciplined manner of aggregating the pixel-level predictions through time. The proposed approach yields the top accuracy among the state-of-the-art relocalization methods over multiple benchmarks. Although KFNet is only validated on the camera relocalization task, the immediate application alongside other tasks like video processing [20, 28] and segmentation [59, 39], object tracking [34, 66] would be anticipated.

9. Acknowledgements

This work is supported by Hong Kong RGC GRF 16206819 & 16203518 and T22-603/15N.
References


Fei Xue, Xin Wang, Zike Yan, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Local supports global: Deep camera relocalization with sequence enhancement. In ICCV, 2019.


