Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

# Salient FlowNet and Decoupled LSTM Network for Robust Visual Odometry

Ming-yue Chen
*School of Computer Science and engineering*
*Nanjing University of Science and Technology*
Nanjing, China
117106021876@njust.edu.cn

Cai-ling Wang
*Department of Automation*
*Nanjing Univiersity of Posts and Telecommunications*
Nanjing, China
wangcl@njupt.edu.cn

Hua-jun Liu *
*School of Computer Science and engineering*
*Nanjing University of Science and Technology*
Nanjing, China
liuhj@njust.edu.cn

***Abstract*—An end-to-end CNN-LSTM network with salient feature attention and context-guided feature selection mechanism for robust visual odometry (VO) on monocular image sequence is developed in this paper. Deep learning-based visual odometry methods have drawn significant concerns comparing with traditional methods. Existing learning-based VO methods usually ignored the redundant features would increase error accumulation, and their rotational and translational motion parameters coupling would enlarge trajectory drift. A scheme on enhancing the visual salient features and decoupling motion parameters to alleviate these problems is investigated in our approach. The classical FlowNet paralleled with a VGG-based salient feature model to reinforce perceptive fields is designed based on an attention mechanism to extract the prominent geometric features from successive monocular images. Furthermore, to reduce the coupling of different motion patterns, a motion decoupled dual Long Short-Term Memory (LSTM) scheme based on guided feature selection mechanism is designed to select guided features to separately regress rotational and translational parameters. Experiments on KITTI dataset show competitive performance of the proposed approach compared with state-of-the-art deep learning based visual odometry methods.**

***Keywords—visual odometry, salient feature attention, context-guided feature selection, salient FlowNet, decoupled LSTM***

## I. INTRODUCTION

In the field of robotics and computer vision, Visual Odometry (VO) is the process to determine equivalent odometry information using sequential images to estimate the 6-DOF position and orientation of a camera or a robot. As an essential section of Visual Simultaneous Localization and Mapping (V-SLAM) [1], visual odometry provides a significant good initial trajectory estimation for back-end optimization.

For its vital role in SLAM systems and GPS-denied applications, enormous work has been done to develop robust and accurate VO algorithms. Traditional VO methods prone to follow a standard pipeline [2, 3] containing of feature extraction, feature matching, motion estimation, scale estimation and local optimization etc. Since Wang [4] first tried to solve the visual odometry tasks with a deep neural network, using learning-based methods to solve the problem for VO has become a new research hotspot.

Benefiting from the potential of deep neural network's data-driven approach, learning based monocular VO has shown apparent advantages over geometric methods in the aspect of metric scale, continuous tracking and robustness. When it comes to feature extraction, the classical methods, such as Harris [5], SIFT [6] and SURF [7] would become invalid in the weak texture scene; however, FlowNet, a classical convolutional neural network (CNN) [8], is far less restricted by the environment and is more competent to extract abundant geometric motion features from monocular image sequences. Moreover, these features along the time axis can be fed into a LSTM network for end-to-end motion modeling and camera pose recovery. Based on this idea, much work has been done to improve the robustness and accuracy of VO. On one hand, the feature outliers or feature redundancy would degrade the pose estimation accuracy, which means the error between adjacent frames will affect the subsequent pose estimation by incremental calculation. On the other hand, sequential geometric features are fed into a single recurrent neural network to simultaneously estimate translational and rotational parameters. However, classical cases on visual odometry tell us that translational motion computation in VO system usually is done by minimizing de-rotated reprojection error after the rotation has been estimated. Moreover, those features on the upper part or side part of field usually help to estimation rotational parameters, but the features near the camera help to estimate the translational parameters. So guided feature selection for motion decoupling mechanism helps to enhance the motion parameters.

Some visual odometry methods based on CNN-RNN structure, such as DeepVO [4], ESP-VO [9], where all features are involved and without features filtering strategy, could lead to excessive trajectories estimation error. In SalientDSO [10], visual saliency was first introduced to enhance salient regions and to neglect weak texture regions. And in EncNet [11], researchers proposed a context encoding module to capture global context information and highlight category information associated with the scene. Moreover, in SENet [12], authors developed a SENet block to adaptively recalibrate each channel's response to the feature map. To alleviate the problem of error accumulation, traditional VO method [13, 14, 15] usually use global map and loop closure detection to correct the prediction of poses. Some learning-based methods [16, 17] also apply the graph optimization mechanism to reduce motion reconstruction error.

Inspired by previous work on visual odometry and deep learning, based on the CNN-LSTM end-to-end deep neural network structure, we decided to introduce salient feature attention mechanism to solve the problem of feature redundancy and guided feature selection mechanism to select more contextual information for rotation and translation regression separately. So in this paper, we consider salient features by a pre-trained VGG is superimposed to a FlowNet to get salient micro motion features we would like to pay attention to, and those features are fed to dual LSTM based on



978-1-7281-6321-5/19/$31.00 ©2019 IEEE

context-guided feature selection mechanism for rotational and translational parameters decoupled estimation.

To summarize, the main contributions of this paper can be concluded as following three aspects:

1) An end-to-end CNN-LSTM network with salient feature attention and guided feature selection mechanism for monocular visual odometry;

2) The FlowNet enhanced by a VGG-based salient feature to extract predominant and effective geometric features on image sequence to reduce feature redundancy;

3) A dual decoupled LSTM network structure by guided feature selection for robust rotational and translational motion estimation separately.

The remainder of this paper is organized as follows: Section Ⅱ describes the related work, and Section Ⅲ introduces the details of our proposed end-to-end VO system. The experimental results are introduced in Section Ⅳ. Finally, the conclusion is given in Section Ⅵ.

## II. Related Work

In this section, earlier work on the monocular VO methods is summarized. There are mainly four aspects of the related work: geometry based and learning based methods, mechanism of visual saliency and decoupling of rotation and translation.

### *A. Geometry-based VO methods*

Generally, geometry based VO methods can roughly be categorized into feature-based methods and direct methods. MonoSLAM [18] is the first proposed real-time monocular vision SLAM system. MonoSLAM uses Extended Kalman Filter (EKF) as the back end to track the sparse feature points of the front end [19], and updates its mean and covariance with the camera's current state and all landmark points as state quantities. In PTAM [20], authors first put forward a rule on distinguishing the front end and the back end, which is followed by other researchers and is also the first few methods using nonlinear optimization instead of filter as the back end. For unavoidable of feature noise and outliers, most these methods are suffered from error accumulation and trajectory drift. ORB-SLAM [21, 22] used bundle adjustment for local optimization and loop closure detection for global pose graph optimization to solve this problem. Unfortunately, feature-based methods are limited to specific environment and usually fail to work with weak texture information. While DTAM [23] is a direct and dense method which minimizes the global space specification energy function to calculate the key frame to construct a dense depth map, the pose of the camera is calculated by direct image matching under the aim of the depth map. For this reason, DTAM works more robustly when features are incomplete or image blur exists. Both direct [24] and indirect methods cannot recover absolute scale without auxiliary information and deep learning is expected to play a crucial role in these aspects.

### *B. Learning-based VO methods*

At present, deep learning has been used to replace some parts of traditional VO methods, such as semantic maps, relocation, loop closure detection, feature extraction and matching.

In DeepVO [4], researchers first proposed an end-to-end structure to perform pose estimation. DeepVO is based on CNN-RNN structure, where CNN is used to extract geometric features and RNN is used to reconstruct trajectory on the basis of sequential features. Similar as DeepVO, MagicVO [25] utilized a bi-directional LSTM to obtain absolute-scale factor at each position of the camera with a sequence of successive monocular images as input. UnDeepVO [26] is also capable of estimating the 6-DOF pose of a moving monocular camera and its depth in the field of view was designed through two neural networks paralleling simultaneously. UnDeepVO has two notable characteristics: one is using unsupervised deep learning mechanisms, and the other is its ability to restore absolute scale. VINet [27] fuses visual and inertial information (IMU) as a sequence-to-sequence learning problem. Literature [28] presents a new monocular VO system which an unsupervised learning based front-end has called Neural Bundler and a graph optimization back end.

### *C. Visual saliency in visual tasks*

Visual saliency [29] has been used to improve performances in many visual task in recent years, such as tracking, recognition, segmentation and navigation.

Attention mechanism can be applied on both spatial domain or channel domain in many different ways. ViS-HuD [30] used a pre-trained DetectNet to get a visual saliency map, then the saliency map multiplied to the original images to enhance human detection results. SENet [12] focused on the relationship in temporal domain and used SENet-like mode to guide features in a contextual way. Few work utilizes salient feature in VO tasks until SalientDSO [10] first brought salient features to this area, and it proposed a direct sparse odometry framework based on a visual saliency map and also investigate a novel method to filter saliency map on scene parsing. Emilio [31] defined a soft-attention operation over the entire trajectory for each phase for visual odometry, and its attention operation allows each pose to query information over long time spans.

### *D. Decoupling of rotation and translation*

Rotation and translation are two main motion parameters of a moving rigid. Most existing learning-based VO methods [4, 32, 33] use an end-to-end network to predict 6-DOF pose simultaneously.

While some researchers assume a dominant ground plane and demonstrate that rotation and translation estimation can be decoupled. This concept of decoupling was first applied in traditional VO methods [34, 35]. For instance, in [35], researchers use a homograph formulation with decoupled rotation and translation for motion estimation. And in [35], Bazin proposed that the rotation is computed based on the detection of dominant bundles of parallel lines and the translation is calculated from a robust 2-point algorithm. In paper [36], authors designed a hybrid visual odometry algorithm which separately estimates the rotation and translation motion to achieve improved accuracy and low drift error. The main origin of drift error comes from the rotation error, so Kim estimated the translational motion by minimizing the de-rotated reprojection error. After a lot of validation experiments proving the effectiveness of motion decoupling, references [37] and [38] use this mechanism in learning-based VO methods and started some exploration.

But until now, few work uses motion decoupled deep neural network to regress 6-DOF pose.

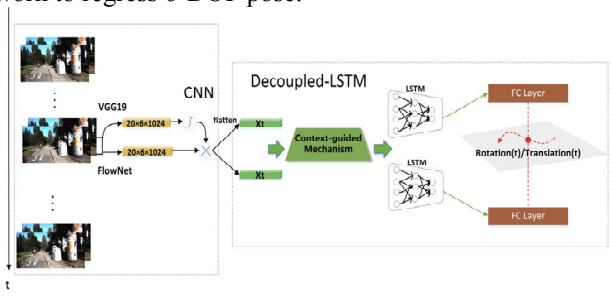



Fig.1. An overview of proposed framework. A CNN-LSTM end-to-end structure with salient feature attention and guided feature selection mechanism.

## III. METHOD

Our framework, is roughly composed of three parts seen in Fig. 1. In Section Ⅲ.1, a CNN with salient feature attention mechanism aims to encode successive image pairs into effective geometric features with pixel-level motion information. In Section Ⅲ.2, the detail of the dual LSTM network whose input at the moment are features selected by context-guided by the previous output is introduced. The dual LSTMs are fed with the guided geometric features extracted by the salient FlowNet to get separate rotational and translational parameters, which will be fed to a fully connected layer to obtain the final pose estimation. Finally, in Section Ⅲ.3, we introduce the loss function definition of the dual branches LSTM network for rotational and translational errors.

### *A. Geometric feature extraction with salient feature enhanced FlowNet*

The fundamental task of learning-based VO methods comes from geometric feature extraction. Part of CNN network can extract micro motion features hidden in the pixels of sequential images benefiting from its consciously learning and adaptation to various application scenarios. A well-known CNN architecture known as FlowNet [39] is firstly proposed for accurate dense optical flow estimation using consecutive image pair as input. Subsequently, multiple stacked convolutional networks are also designed to estimate depth and motion concurrently. To get the accurate motion features, it is necessary to concatenate two consecutive images on their channel dimension, similar as FlowNet [39], to extract the optical flow. In our structure, we retained a lightweight FlowNet as backbone, and its first 9 layers contain multiple kinds of convolutional kernels and it is fine-tuned on the pre-trained weights to fit our dataset. It is worth mentioned that there are two structures in the FlowNet, namely FlowNetS and FlowNetC. In our work, FlowNetS was first selected for its excellent performance in large displacement scenarios which is more suitable for our experiments environment.

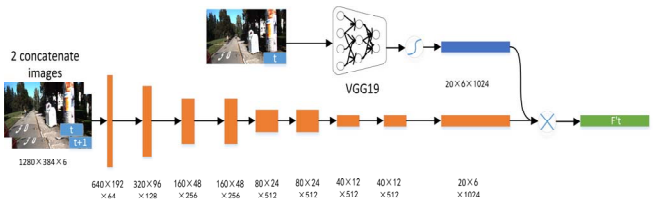



Fig.2. An illustration of a FlowNet with salient feature attention mechanism. The value of the feature map calculated by a pre-trained VGG19 is restricted by [0-1] throughout sigmoid activation function. Multiply two feature maps which calculated by FlowNetS (below) and VGG19 (above) in element-wise.

The stacked images of 6 channels are fed into the FlowNetS, and then 1024-dimensional feature maps can be output. The FlowNetS can be described as Equ. (1) and its structure is shown in Fig. 2, and it will be enhanced by a salient feature sub-network.

$$F_t = \theta_{CNN}(I_{t-1}, I_t) \tag{1}$$

where $I_{t-1}$ , $I_t$ denotes consecutive images, and $F_t = [F_t^1, F_t^2, \dots, F_t^C] \in \mathbb{R}^{C\times H\times W}$ is extracted feature with channel C and size $H \times W$. $\theta_{CNN}$ presents parameters of convolution neural network.

However, the geometric features extracted by classical FlowNet did not work well where the geometric features were not significant in part of regions such as pavement, highway or sky. Therefore, these insignificant and weak texture features should be neglected, once these redundant features were utilized equivalently, it would degrade the pose recovery performance. Therefore the visual salient attention is considered to significantly enhance the obvious geometric features, such as vehicles, buildings and trees etc. alongside the road, and weigh down the uninformative regions, such as pavement, sky and floors etc. Hence, VGG19 [40] of pre-trained weights on ImageNet is used to obtain geometric salient feature maps and is superimposed to the FlowNet features. In this way, the final feature maps are involved with visual saliency information. It is believed that deeper layers are more capable of extracting useful saliency map instead of edge information. We visualize a feature map extracted by VGG19 in Fig.3, and the geometric features of vehicles and trees are clearly shown on the feature map. The FlowNet with

Fig.3. Typical consecutive frames of KITTI dataset and the feature maps obtained by VGG19. As it is illustrated in Fig.3, geometric features of buildings, trees and vehicles are clearly shown in the feature maps.

salient feature attention mechanism is shown in Fig. 2 and its principal can be described as Equ. (2) - (4).

$$X_t = \theta_{VGG}(I_t) \tag{2}$$

$$X_t' = \sigma(X_t) \tag{3}$$

$$F_t' = X_t' \odot F_t \odot 255 \tag{4}$$

where the final output $F_t'$ are still the features with channel $C$ and size $H \times W$, $X_t$ are feature maps extracted by VGG19 with the same dimensions as $F_t$, and the value of each element of $X_t$ is in [0-255]. Meanwhile, σ (·) denotes sigmoid activation function. Due to the nature of the sigmoid activation function, the result is mapped to the [0-1]. The purpose of using a visual saliency model is to propose possible regions where the objects might have large and clear displacement. By multiplying the corresponding saliency maps, every single pixel of each feature map from VGG19 becomes magnification reducer to control the activity of features form FlowNet. At last, the model multiplies 255 to restore the original range of the features.

### B. *Decoupled and context-guided dual LSTM*

LSTM is a kind of time recurrent neural network and suitable for sequential features regression. Due to its unique structure, LSTM is suitable for processing and predicting events with very long intervals and delays in time series [41, 42]. Considering that rotation and translation are two different motion modes, it may not be wise to predict both of them simultaneously in a single network. A mount of experiments prove that the translational error is the main source of drift error because it is difficult to estimate due to scale ambiguity in monocular vision configuration. And those features on the upper parts and side parts of field usually help to estimation rotational parameters, but the features near the camera help to estimate the translational parameters. When estimating the rotational or translational parameters, these feature should be selected. And in many learning-base VO methods, the loss function of rotation component and translational component should be assigned by different ratio. Therefore, in our work, we separate them and design two different branches to learn different motion patterns of via selected features independently. In our architecture, the rotation branch and the translation branch are feed with selective features guided by the common features from CNN module. Both of them are individual LSTM network and don't share weights and parameters.

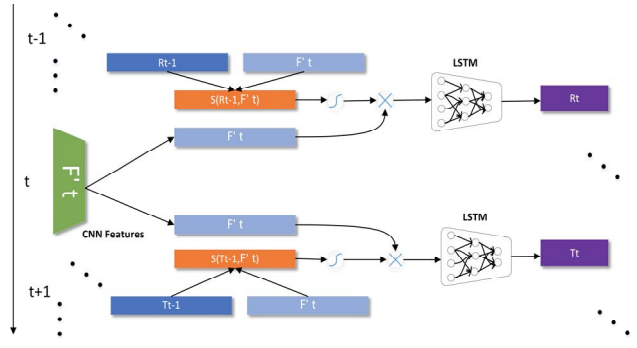



Fig.4. Decoupled LSTM module on context-guided feature selection mechanism. $F_t'$ are features from CNN. Encoded features at each time step was corrected by the previous moment and then fed into LSTM cells $S(R_{t-1}, F_t'), S(T_{t-1}, F_t')$ denotes the similarity between the output of the previous moment $R_{t-1}$, $T_{t-1}$ and the input $F_t'$ at this moment. The method of calculating the similarity is given below.

Besides, in order to make the network more streamlined, avoid redundant information may cause further deterioration and reduce the accuracy of prediction. We propose to use the state variables $c_t$ of the previous moment to let the network adaptively select and guide the input $F_t'$ at this moment. It is called context-guided feature selection mechanism by using the output of the last time to guide the current input. Since the previous output already contains visual cues that are valuable for motion estimation, it can be used to supervise $F_t'$ at this moment. Furthermore, unlike DeepVO and other learning-based methods, we use dual LSTM structure to predict rotation and translation to reduce the coupling of two type of motion. According to this rule, all input vectors $F_t'$ except the initial one must be corrected before entering the recurrent neural network. The diagram decoupled dual LSTM module on context-guided feature selection mechanism is illustrated in Fig.4. Here is the original manner for calculating output:

$$o_t, (h_t, c_t) = \theta_{LSTM}\left(F_t', (h_{t-1}, c_{t-1})\right) \tag{5}$$

Among them, $o_t, h_t, c_t$ represents output, hidden state and variable state at time t. Then a similarity criteria of the output and input is used to correct $F_t'$.

Here we take the rotation branch as an example, their similarity coefficient $S_{(F_t',c)}$ can be defined as:

$$S_{(F_t',c)} = \sigma\left(\frac{F_t' \cdot c_{t-1}}{\|F_t'\|_2 \cdot \| c_{t-1}\|_2}\right) \tag{6}$$

As is shown above, $S_{(F_t',c)}$ is the coefficient to correct the input vector $F_t'$. After correction, the rectified output is calculated as:

$$F_t^{corr} = S_{(F_t',c)} \odot F_t' \tag{7}$$

$$o_t^{corr}, (h_t, c_t) = \theta_{LSTM}\left(F_t^{corr}, (h_{t-1}, c_{t-1})\right) \tag{8}$$

Each element of the geometric feature vector $F_t'$ is multiplied by its similarity coefficient $S_{(F_t',c)}$ to guide original input and get the $F_t^{corr}$ for subsequent LSTM pose regression. This context-guided feature selection and correction strategy above has been proved to bring excellent results in our experiments.

### C. *Loss function*

In our proposed methods, we use relative pose to generate our training labels. The relative pose describes the transformation between two adjacent image pairs, and the relative pose can be calculated from the absolute pose. $T_1, T_2$ and $T_3$ represent absolute poses of the camera at three successive positions, respectively. Then $T_{12}$ and $T_{23}$ are the transformation of two adjacent positions. For two adjacent images, assuming that the homogeneous rotation matrix of the i-th picture is $T_i$ and matrix of the (i+1)-th is $T_{i+1}$, the relative pose of the two pictures is represented as:

$$T_{ri} = T_i^{-1} T_{i+1} = \begin{bmatrix} & \Delta r & & \Delta x \\ & & & \Delta y \\ & & & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{9}$$

where $\Delta r$ denotes the rotation matrix of the homogeneous matrix $T_{ri}$. As we know, rotation matrix, Euler angles, and quaternions all can describe a 6-DOF pose. In our work, Euler

angle performs better, because it has less degree of freedom and easier to regress. Besides, there is no Gimbal lock phenomenon. The relative pose labels we used are preprocessed as follows:

$$P_{ri} = [\Delta x\ \Delta y\ \Delta z\ \Delta\psi\ \Delta\chi\ \Delta\phi]^T \quad (10)$$

where $[\Delta x\ \Delta y\ \Delta z]$ denotes the translation vector of $T_{ri}$, and $[\Delta\psi\ \Delta\chi\ \Delta\phi]$ represents the Euler angle transformed by rotation $\Delta r$ of $T_{ri}$. The 6-dimensional pose obtained by the network output consists of the translation $\hat{p} = (\Delta x\ \Delta y\ \Delta z)$ and the rotation $\widehat{\Phi} = (\Delta\psi\ \Delta\chi\ \Delta\phi)$. Since the two poses of rotation and translation are existed in different spaces, two branches are used to regress them respectively, and the branch loss function is defined as Equ. (11) and Equ. (12). The total loss function is defined as their weighted sum as Equ. (13):

$$\mathcal{L}_{trans} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{p}_i - p_i\|_2^2 \quad (11)$$

$$\mathcal{L}_{rot} = \frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\Phi}_i - \Phi_i\right\|_2^2 \quad (12)$$

$$\mathcal{L}_{total} = \frac{1}{N}\sum_{i=1}^{N}(\|\hat{p}_i - p_i\|_2^2 + \beta\left\|\widehat{\Phi}_i - \Phi_i\right\|_2^2) \quad (13)$$

where $\|\cdot\|_2$ is 2-norm, N denotes the number of training samples, $\hat{p}_i\ and\ \widehat{\Phi}_i$ are the predicated rotation and translation by dual LSTM individual branches. The Mean Square Error (MSE) function is used as the criterion function. And $\beta$ is a weighted factor used to balance rotation and translation component. We empirically chose 50 and 100 as $\beta$ for experiments. Experiments show that our loss function is designed reasonably and the model training converged well.

## IV. Experiments

In this section, our method has been trained and tested on the well-known KITTI VO/SLAM benchmark [18], and it has been compared with other state-of-the-art VO methods and traditional methods. The experiment results show that the proposed method is superior to many traditional monocular methods and other state-of-the-art VO methods.

### A. Implementation details

*1) Network and Training:* The network is implemented based on a very popular deep learning framework PyTorch. The optimization algorithm is set as batch gradient descent, and the optimizer is set as Adam (Adaptive Moment Estimation), where $\beta_1$=0.9, $\beta_2$=0.99. The initial learning rate is set as $10^{-4}$. As the iterator increases, the learning rate will automatically decay, ensuring that the optimization function is closer to the optimal solution. The model was trained by using the NVIDIA TITAN XP GPU for the purpose of accelerating. Batch-size is set to 16. Early stopping strategy was used, so once the validation error has been detected rising abruptly, the model would stop training. The input of the network is a 6-channel image pair with two consecutive images concatenate. In order to expand the training dataset, we change the order of two adjacent images generating more image pairs, which means that the two images need to be input twice. Weights of LSTM are initialized by Xavier, while the CNN is based on pre-trained FlowNet [39].

*2) KITTI dataset and data augmentation:* The public KITTI VO/SLAM benchmark has 22 sequences of images, while only 11 of them have ground truth. Quantitative experiments can be performed on these sequences associated with ground truth. Specifically, training sets, validation sets, and testing sets are selected from these scenarios and are not duplicated. Train a model with the training set and validation set and then test it on the test set. All these sequences of images are recorded at a relatively low frame (10 fps) and consist of urban areas and highway environments at the speed up to 90km/h.

Since there are only 11 sequences in the KITTI dataset can be used for training, it may not be enough to validate complex VO problem. Therefore, we propose a data augmentation method by changing the order of the adjacent pictures in the image pairs to obtain the rotation and translation of the next frame relative to the previous frame. Besides, this method can also be considered to make the network more accurately predict the pose between two adjacent frames by adding constraints.

In this paper, we take sequences 00, 01, 02, 06, 08 that are relatively long as training set and other sequences for testing. The verification set comes from the training set and is sampled with no return.

### B. Ablation study

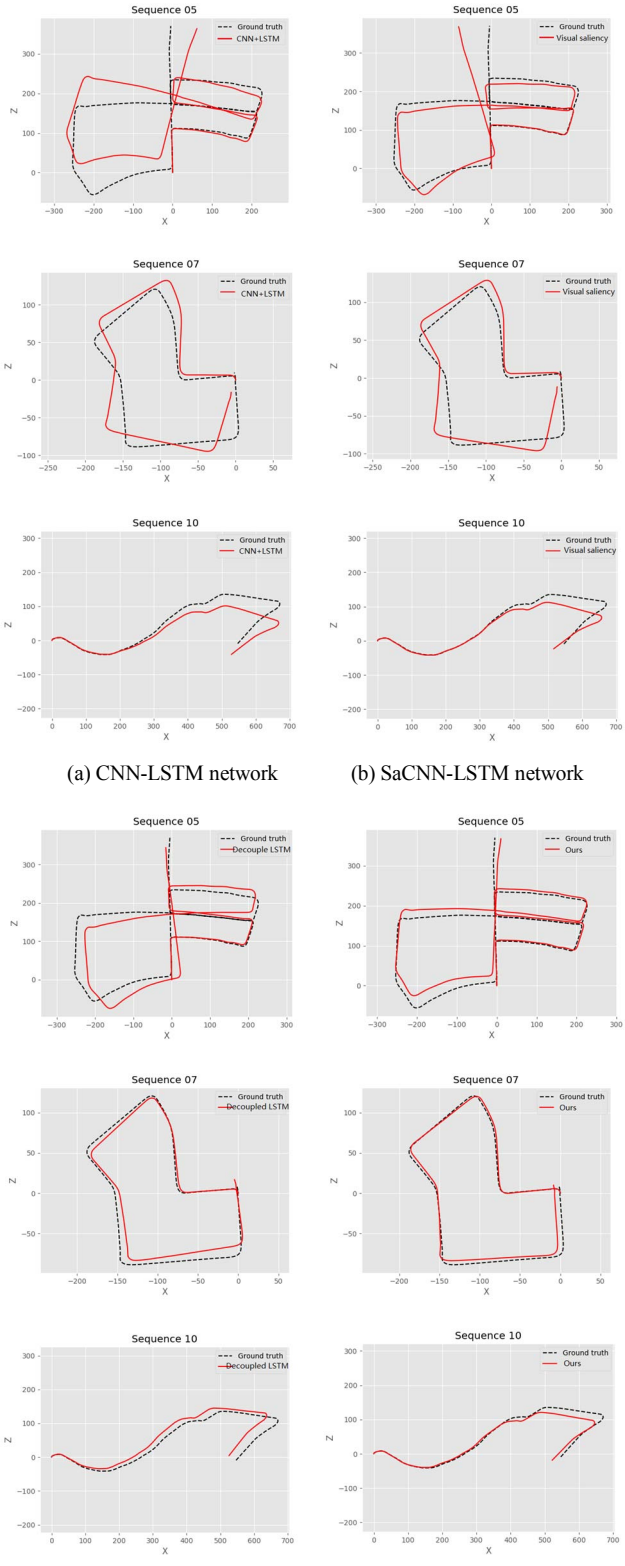
We first evaluate the efficiency of visual saliency strategy by training two net-works of the same times of iterators with or without it, marked as "CNN" or "SaCNN" repectively. The result in Fig 5. Show that for different two networks with the same numbers of training times, the network with visual saliency performs better than the one without this mechanism. It is mainly because the visual saliency mechanism is like a magnification reducer which can strengthen the effective features of the motion, making the network easier to converge and more accurate to predict. So for the same number of iterators, the network with visual saliency module has smaller error. The results demonstrate the effectiveness of the visual saliency mechanism.

Then we validate the context-guided LSTM model in the similar way. The trajectory in Fig. 5. also denotes that the network with decoupled LSTM method (marker as "DeLSTM") predicts the poses more accurately than the network without it (marked as "LSTM"). This shows that in VO tasks, our approach can successfully decouple the two patterns of motion and correct the features of each time step effectively. Quantitative comparison results are shown in the Tab.1. From the quantitative results, decoupled LSTM module brings more significant performances than CNN module, especially in the average error of rotation. We explain that the rotation error accumulated at each time step is the main cause of drift error. While LSTM module reduces the error of each step by selecting context-aware features and lets each branch learn by itself.

### C. Experiment results on KITTI dataset

The proposed method was analyzed according to the KITTI VO/SLAM error metrics, i.e., averaged Root Mean Square Errors (RMSE) of translational and rotational error, which was adopted for all sequences of lengths ranging from 100, 200 to 800. Our model was trained for more than 100 epochs on the sequence 00, 01, 02, 08 using the training method in Section 4.1.

Since two branches were applied to our model, the errors of rotation and translation both converge to a stable level as the number of iterator increases, and the decent process is shown in the Fig. 7. Besides, the average of the RMSE of the




(a) CNN-LSTM network (b) SaCNN-LSTM network

(c) CNN-DeLSTM network (d) SaCNN-DeLSTM network

Fig.5. Individual evaluation of visual saliency and decoupled LSTM.

different speeds and lengths is given in the Fig. 6. The 6-Dof pose obtained by the net-work is calculated in the SE (3) space. We visualize the trajectories in the X-axis and Z-axis spaces and compare our results with ORB-SLAM (Stereo), VISO-2 [43], and Depth-VO-Feat [44]. In order to fully evaluate our approach, the chosen method of comparison includes both traditional VO methods and learning-based methods.

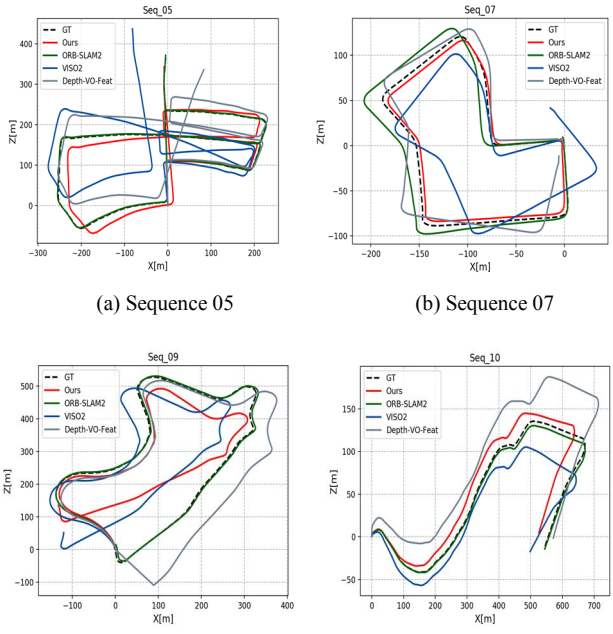
TABLE I. AVERAGE TRANSLATION AND ROTATION RMSE DRIFT OF ABLATION EXPERIMENTS

| Seq. | CNN-LSTM | | CNN-DeLSTM | | SaCNN-LSTM | | SaCNN-DeLSTM | |
|---|---|---|---|---|---|---|---|---|
| | $t_{rel}$ | $r_{rel}$ | $t_{rel}$ | $r_{rel}$ | $t_{rel}$ | $r_{rel}$ | $t_{rel}$ | $r_{rel}$ |
| 03 | 12.49 | 9.83 | 6.15 | 8.83 | 7.21 | 8.77 | **4.24** | **7.87** |
| 04 | 8.19 | 5.97 | 4.33 | 5.06 | 3.87 | 4.06 | **2.96** | **1.73** |
| 05 | 5.62 | 4.23 | 4.01 | 2.99 | 4.51 | 3.84 | **2.14** | **1.58** |
| 07 | 6.91 | 4.51 | **3.21** | 3.27 | 4.09 | 2.15 | **3.21** | **1.76** |
| 10 | 12.17 | 7.65 | 7.50 | 6.62 | 6.58 | 6.33 | **5.72** | **6.17** |

includes both traditional VO methods and learning-based methods. We selected the sequences 05, 07, 09, 10 to compare these methods quantitatively and qualitatively. Trajectories and the RMSE errors of rotation and translation are shown in the Tab. 2.

As shown in the figures and table, our method is superior to the state-of-art monocular visual odometry method VISO2_M [19] and learning-based methods, such as Depth-VO-Feat [42], ESP-VO [9] and DeepVO [4]. And the results is very close to the stereo camera configured visual odometry method VISO2_S.

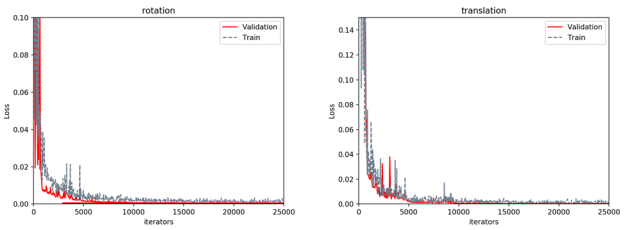
The most convincing reason is that our method strategically selects the features of the network, enhancing the effective features and eliminating the redundancy features and error. Therefore, compared with other learning-based methods, such as Depth-VO-Feat, the accumulation of trajectory errors is reduced, so the trajectory is more accurate. However, since ORB-SLAM is a stereo VO method, it is more accurate in scale than the monocular methods. Meanwhile, ORB-SLAM uses loop closure detection and other geometry algorithms for optimization, so it performs better than any other methods.




(a) Sequence 05 (b) Sequence 07

(c) Sequence 09 (d) Sequence 10

Fig.6. Trajectories of ground-truth, ORB-SLAM2, VISO2, Depth-VO-Feat and our method in sequence 05, 07, 09 and 10. Our model is trained on 00, 01, 02, 06 and 08.




(a) Rotation loss (b) Translation loss

Fig.7. Training and validation loss on rotation and translation as the time of iterators goes by.

TABLE II. RESULTS ON VISO2_M [19], DEEPVO [1] AND UNDEEP-VO [11] ON DIFFERENT SEQUENCES

| Seq. | Ours | | VISO2_M | | DeepVO | | UnDeep-VO | |
|---|---|---|---|---|---|---|---|---|
| | $t_{rel}$ | $r_{rel}$ | $t_{rel}$ | $r_{rel}$ | $t_{rel}$ | $r_{rel}$ | $t_{rel}$ | $r_{rel}$ |
| 03 | **4.24** | 7.87 | 8.47 | 8.82 | 8.49 | 6.89 | 5.00 | 6.17 |
| 04 | **2.96** | **1.73** | 4.69 | 4.49 | 7.19 | 6.97 | 5.49 | 2.13 |
| 05 | **2.14** | 1.58 | 19.22 | 17.58 | 2.62 | 3.61 | 3.40 | **1.50** |
| 07 | 3.21 | **1.76** | 23.61 | 29.11 | 3.91 | 4.60 | **3.15** | 2.48 |
| 10 | **5.72** | 6.17 | 41.56 | 32.99 | 8.11 | 8.83 | 10.63 | **4.65** |

## V. CONCLUSION

A novel end-to-end CNN-LSTM structure with salient feature attention and context-guided feature selection mechanism to enhance the monocular VO is investigated in this paper. In the part of geometric feature extraction, we use a pre-trained VGG19 to strengthen the salient features and to weaken the redundant features on a FlowNet backbone. In the part of pose recovery, a dual branch LSTM is designed based on context-guided feature selection mechanism to select and correct current features from previous iteratively. Meanwhile, dual LSTM is used to regress rotational and translational parameters separately to reduce the coupling of different motion patterns. Finally, a mount of experiments on KITTI VO/SLAM benchmark show that our method can effectively alleviate the error accumulation and predict more accurate trajectories, which outperforms many state-of-the-art methods.

## REFERENCES

[1] [24] C. Cadena et al., "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," in IEEE Transactions on Robotics, vol. 32, no. 6, pp. 1309-1332, Dec. 2016.

[2] [25] D. Scaramuzza and F. Fraundorfer, "Visual odometry: Tutorial," IEEE Robotics & Automation Magazine, vol. 18, no. 4, pp. 80–92, 2011.

[3] [26] F. Fraundorfer and D. Scaramuzza, "Visual odometry: Part II: Match- ing, robustness, optimization, and applications," IEEE Robotics & Automation Magazine, vol. 19, no. 2, pp. 78–90, 2012.

[4] S. Wang, R. Clark, H. Wen and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks," 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017, pp. 2043-2050.

[5] Mikolajczyk, Krystian and Cordelia Schmid. "Scale & Affine Invariant Interest Point Detectors." International Journal of Computer Vision 60 (2004): 63-86.

[6] Lowe, David. (2004). Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. 60. 91-. 10.1023/B:VISI.0000029664.99615.94.

[7] J. Iparraguirre, L. Balmaceda and C. Mariani, "Speeded-up robust features (SURF) as a benchmark for heterogeneous computers," 2014 IEEE Biennial Congress of Argentina (ARGENCON), Bariloche, 2014, pp. 519-524.

[8] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, 2015, pp. 730-734.

[9] Wang, Sen, et al. "End-to-End, Sequence-to-Sequence Probabilistic Visual Odometry through Deep Neural Networks." The International Journal of Robotics Research, vol. 37, no. 4–5, Apr. 2018, pp. 513–542.

[10] H. Liang, N. J. Sanket, C. Fermüller and Y. Aloimonos, "SalientDSO: Bringing Attention to Direct Sparse Odometry," in IEEE Transactions on Automation Science and Engineering.

[11] H. Zhang et al., "Context Encoding for Semantic Segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7151-7160.

[12] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7132-7141.

[13] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in Proceedings of IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2014, pp. 15–22.

[14] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," in arXiv:1607.02565, July 2016.

[15] S.Leutenegger, S.Lynen, M.Bosse, R.Siegwart, and P.Fur-gale. Keyframe-based visual-inertial odometry using non-linear optimization. The International Journal of Robotics Research, 34(3): 314-334, 2015.

[16] E. Parisotto, D. S. Chaplot, J. Zhang and R. Salakhutdinov, "Global Pose Estimation with an Attention-Based Recurrent Network," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, 2018, pp. 350-35009.

[17] G.Costante,M.Mancini,P.Valigi,andT.A.Ciarfuglia.Exploring representation learning with cnns for frame-to-frame ego-motion estimation. IEEE robotics and automation let- ters, 1(1):18–25, 2016.

[18] A. J. Davison, I. D. Reid, N. D. Molton and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1052-1067, June 2007.

[19] Liu. H, Wang. C, Yang. J, "Vanishing points estimation and road scene understanding based on Bayesian posterior probability," Industrial Robot: An International Journal, 2016, pp.12-21.

[20] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, 2007, pp. 225-234.

[21] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," in IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163, Oct. 2015.

[22] Cui. Q, Liu. H, Wang. C, "Robust multi-scale ORB algorithm in real-time monocular visual odometry," In 4th IAPR Asian Conference on Pattern Recognition, Nanjing, 2017, pp. 244-249.

[23] R. A. Newcombe, S. J. Lovegrove and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 2320-2327.

[24] Liu. H, Wang. C, Lu. J, Tang. Z, Yang, J, "Maximum Likelihood Estimation of Monocular Optical Flow Field for Mobile Robot Ego-motion," International Journal of Advanced Robotic Systems,2016.

[25] Jiao, Jian et al. "MagicVO: End-to-End Monocular Visual Odometry through Deep Bi-directional Recurrent Convolutional Neural Network," 2018.

[26] R. Li, S. Wang, Z. Long and D. Gu, "UnDeepVO: Monocular Visual Odometry Through Unsupervised Deep Learning," 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, 2018, pp. 7286-7291.

[27] Clark, Ronald et al. "VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem," 2017.

[28] Y. Li, Y. Ushiku and T. Harada, "Pose Graph optimization for Unsupervised Monocular Visual Odometry," 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 2019, pp. 5439-5445.

[29] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," in IEEE Transactions on Pattern

Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[30] Vandit Gajjar, Yash Khandhediya, Ayesha Gurnani, Viraj Mavani, Mehul S. Raval, ViS-HuD: Using Visual Saliency to Improve Human Detection With Convolutional Neural Networks, CVPR 2018.

[31] E. Parisotto, D. S. Chaplot, J. Zhang and R. Salakhutdinov, "Global Pose Estimation with an Attention-Based Recurrent Network," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, 2018, pp. 350-35009.

[32] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolu- tional network for real-time 6-dof camera relocalization. In Proceedings of the IEEE international conference on com- puter vision, pages 2938–2946, 2015.

[33] J. McCormac, A. Handa, A. Davison, and S. Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolu- tional neural networks. In Robotics and Automation (ICRA), 2017 IEEE International Conference on, pages 4628–4635. IEEE, 2017.

[34] P. Kim, B. Coltin and H. J. Kim, "Low-Drift Visual Odometry in Structured Environments by Decoupling Rotational and Translational Motion," 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, 2018, pp. 7247-7253.

[35] B. Guan, P. Vasseur, C. Demonceaux and F. Fraundorfer, "Visual Odometry Using a Homography Formulation with Decoupled Rotation and Translation Estimation Using Minimal Solutions," 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, 2018, pp. 2320-2327.

[36] Kim. Pyojin, "Visual Odometry with Drift-Free Rotation Estimation Using Indoor Scene Regularities," 2017.

[37] Xue. Fei,"Beyond Tracking: Selecting Memory and Refining Poses for Deep Visual Odometry." ArXiv abs/1904.01892 (2019): n. pag.

[38] Xue, Fei et al. "Guided Feature Selection for Deep Visual Odometry," Asian Conference on Computer Vision, pp. 293-308, 2018.

[39] A. Dosovitskiy et al., "FlowNet: Learning Optical Flow with Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 2758-2766.

[40] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, 2015, pp. 730-734.

[41] Huajun Liu, Cailing Wang, Jianfeng Lu, Zhenmin Tang, Jingyu Yang, "Maximum likelihood estimation of monocular optical flow field for mobile robot ego-motion", International Journal of Advanced Robotic Systems, 13(1): 1-12, 2016.

[42] Cailing Wang, Chunxia Zhao, Jingyu Yang, "Monocular odometry in country roads based on phase-derived optical flow and 4-DOF ego-motion model", Industrial Robot: An International Journal, 38(5): 509-520, 2011.

[43] A. Geiger, J. Ziegler and C. Stiller, "StereoScan: Dense 3d reconstruction in real-time," 2011 IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, 2011, pp. 963-968.

[44] X. Ye et al., "Unsupervised Monocular Depth Estimation Based on Dual Attention Mechanism and Depth-Aware Loss," 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 2019, pp. 169-174.