VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem

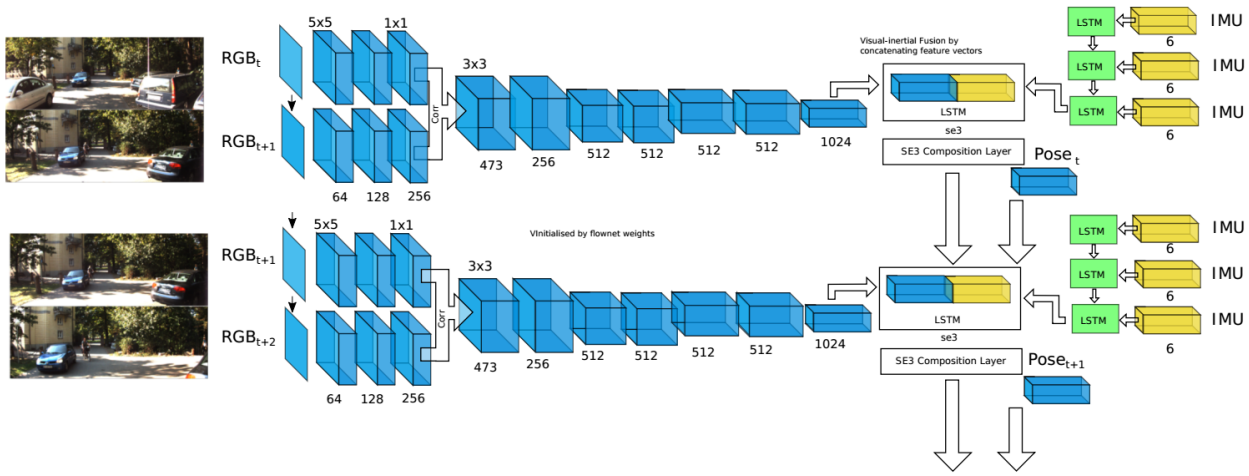
accurately navigate where no GPS signals are available.

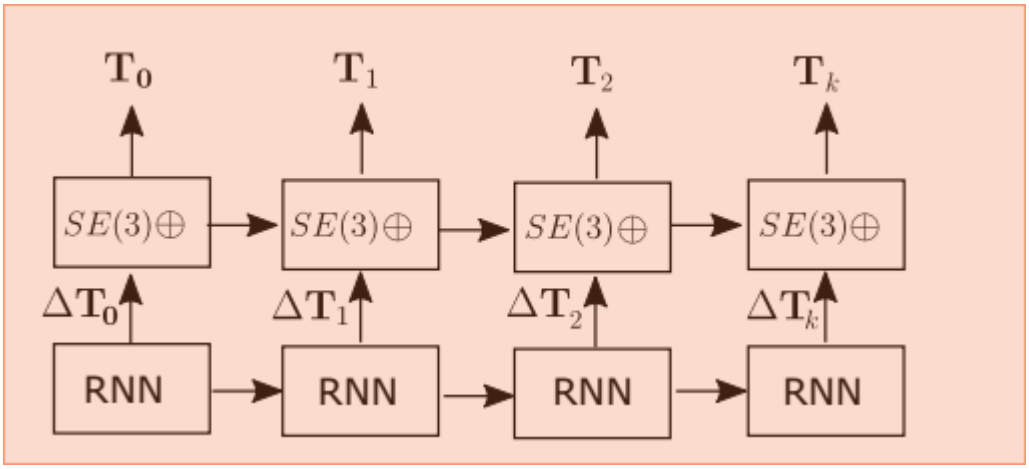
regarding VIO as a sequence-to-sequence regression problem.

input: mono RGB images and IMU data which is a 6 dimensional vector containing the x,y,z components of acceleration and angular velocity.

output: 7 dimensional vector, represent the change in pose of the rotot form the start of the sequence.

$$\mathrm{VIO}: \{(\mathcal{R}^{W\times H}, \mathcal{R}^6)_{1:N}\} \rightarrow \{(\mathcal{R}^7)_{1:N}\}$$





在传统的LSTM中，隐藏层的状态传递到下一步，但输出没有传到下一步。
在论文中，直接将输出的pose作为core LSTM的输入

Multi-rate LSTM

> IMU 100HZ, visual image 10HE. 使用一个小的LSTM处理IMU数据以IMU的接收速率。
> the final hidden-layer activation of the IMU-LSTMis then carried over to the Core LSTM.

CNN produces a single feature-vector describe the motion that the device underwent during the passing of the two frames which is used as input to the Core LSTM.

CNN模仿Flownet到 conv6层，tensor大小为1024×6×20，然后flatten, concatenate with the feature vector produced by the IMU-LSTM before being fed to the Core LSTM.

Core-LSTM fuses the internediate feature-level representation of the visual and inertial data to produce a pose estimate.

使用LSTM with 2 layers with cells of 1000 units.

与ESP-VO类似。