Indoor Relocalization in Challenging Environments with Dual-stream Convolutional Neural Networks

Ruihao Li, *Student Member, IEEE,* Qiang Liu, Jianjun Gui, *Student Member, IEEE,*
Dongbing Gu, *Senior Member, IEEE,* and Huosheng Hu, *Senior Member, IEEE,*

***Abstract*—This paper presents an indoor relocalization system using a dual-stream Convolutional Neural Network (CNN) with both color images and depth images as the network inputs. Aiming at the pose regression problem, a deep neural network architecture for RGB-D images is introduced, a training method by stages for the dual-stream CNN is presented, different depth image encoding methods are discussed and a novel encoding method is proposed. By introducing the range information into the network through a dual-stream architecture, we not only improved the relocalization accuracy by about 20% compared with the state-of-the-art deep learning method for pose regression, but also greatly enhance the system robustness in challenging scenes such as large scale, dynamic, fast movement and night-time environments. To the best of our knowledge, this is the first work to solve the indoor relocalization problems based on deep CNNs with RGB-D camera. The method is first evaluated on the Microsoft 7-Scenes dataset to show its advantage in accuracy compared with other CNNs. Large scale indoor relocalization is further presented using our method. The experimental results show that 0.3m in position and 4° in orientation accuracy could be obtained. Finally, this method is evaluated on challenging indoor datasets collected from motion capture system. The results show that the relocalization performance is hardly affected by dynamic objects, motion blur or night-time environments.**

***Note to Practitioners*–This work was motivated by the limitations of the existing indoor relocalization technology that is significant for mobile robot navigation. By using this technology, robots can infer where they are in a previously visited place. Previous visual localization methods can hardly be put into wide application for the reason that they have strict requirements for the environments. When faced with challenging scenes such as large scale environments, dynamic objects, motion blur caused by fast movement, night-time environments or other appearance changed scenes, most existing methods tend to fail. This paper introduces deep learning into the indoor relocalization problem, uses dual-stream CNN (depth stream and color stream) to realize 6-DOF pose regression in an end-to-end manner. The localization error is about 0.3m and 4° in a large scale indoor environments. And what is more important, the proposed system does not lose efficiency in some challenging scenes. The proposed encoding method of depth images can also be adopted in other Deep Neural Networks with RGB-D cameras as the sensor.**

***Keywords*—Deep learning, CNN, relocalization, pose regression, depth encoding.**

I. Introduction

INDOOR relocalization is a challenge task which is widely studied in the areas of mobile robot navigation and computer vision. It enables robots the capacity to infer where they are in a previously visited place. Nowadays, commercial RGB-D cameras are preferred in indoor robotic applications considering their low-price, good technical support, real-time performance and excellent sensing capability. They can provide not only color images which contain texture and appearance information, but also the depth of images containing range information and structure of objects which are robust and invariant when lighting condition varies.

Appearance based methods are most popular for the relocalization task. Hand-crafted features such as Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF) or Oriented FAST and Rotated BRIEF (ORB) are extracted and stored first. Then images with similar appearance are retrieved using Bag of Words (BoWs) technology. Frame-to-frame feature correspondence and pose estimation [1] are implemented for localization at last. This method can achieve satisfactory precise and real-time performance when the view is consistent. Appearance based relocalization and loop closure detection are also widely used in visual Simultaneous Localization And Mapping (SLAM) [2] [3]. However, they have strict requirements for the environment that limits their applications in practice. They would no longer be useful when appearance changes. Both moving objects and lighting variation could lead to appearance changes. [4] Motion blur caused by fast movements of the camera could also result in failures as manually designed features are fragile when encountering motion blur. In addition, accurate frame-to-frame pose estimation needs small viewpoint which limits the widespread usage of appearance based methods.

Deep Convolutional Neural Networks (CNNs) designed for image processing have achieved astonishing success in computer vision. Object recognition and detection capabilities are greatly improved due to the wide usage of CNNs [5] [6]. CNNs can learn different features according to various targets and provide an end-to-end solution to perception problems. Recently 6-DOF camera pose regression method with CNNs (PoseNet) [7] was proposed. Unlike SLAM or appearance based relocalization, the pose regression with CNNs does not need to store keyframes, match features between frames and perform pose optimization. And the storage memory and computing time using CNNs do not increase when exploring large scale area. Therefore, it can implement large scale relocalization without area limitation. Furthermore, its

Manuscript received July 27, 2016; revised November 03, 2016; accepted January 14, 2017.

Ruihao Li, Qiang Liu, Jianjun Gui, Dongbing Gu and Huosheng Hu are with Department of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK {rlig, qliui, jgui, dgu, hhu}@essex.ac.uk

performance in challenging situations could be improved as well. However, we find that PoseNet with only color images as the input could degrade the performance when working in some extremely challenging indoor environments.

In this paper, we present a novel dual-stream CNN architecture with RGB-D camera which can implement indoor relocalization even in extremely challenging environments. Depth images are introduced in a separate stream to learn reliable range features. Range features from depth stream and appearance features from color stream are jointly optimized to implement the 6-DOF camera pose regression by learning the proposed CNN. Our main contributions in this paper are summarized as follows:

- We present a dual-stream CNN to achieve indoor relocalization in challenging environments with an end-to-end manner. The CNN can learn localization features from both color and depth images, and estimate camera poses from these learned features. A network training strategy that divides the training into three stages is proposed. Approximately 20% improvement on localization accuracy is achieved compared with PoseNet.
- We study the encoding methods of depth images and propose a novel method called minimized normal + depth (MND) encoding to solve the 6-DOF pose regression problem. The MND encoding images contain both pixel orientation information and absolute depth information, and maintain the ability to leverage the transfer learning at the same time. Different network architectures with color and depth images as the input are discussed. By taking the depth information into the CNN, not only the relocalization accuracy is improved, but the system robustness in some challenging environments is enhanced.
- Robustness evaluation experiments are implemented in dynamic, night-time and fast movement environments. Experiments on a large scale indoor dataset and public datasets are also presented.

In the following section, we provide a review of the related work. In section III, we give an introduction to the architecture of the proposed CNN, present the method for pose regression and the encoding method of depth images. Section IV addresses three stages for network training with the dual-stream CNN. Section V demonstrates the experimental results on different datasets using the dual-stream CNN. In section VI, we give a summary conclusion and the future work we would like to investigate.

II. RELATED WORK

Relocalization is a significant technology in robotics which could be used for search and rescue, navigation, intelligent services and so on. Place recognition is a problem related to relocalization. When compared with place recognition, relocalization needs to solve an additional geometric transformation problem. Relocalization could also be transferred to a loop closure detector in SLAM [8].

A. *Visual geometry relocalization*

In the early years, Iterative Closest Point (ICP) [9] algorithm was usually adopted for relocalization by registering local point clouds from frames to global point clouds from global map. Steder at el. [10] firstly, extracted 3D features of images captured from RGB-D cameras, then place recognition and relocalization are implemented with 3D feature points registration using ICP. Except for feature points, high level features such as planes [11] and lines are also used for localization. Cupec at el. [12] extracted robust line segments and planar surface segments as primitive features instead of feature points to recognize places. The system performed robustly even in some appearance changed environments. A novel place recognition and scene registration method using multi-planes was addressed by Fernández-Moralc at el. [13]. High-level surface normal semantic planes, color information and other features are used. However, ICP algorithm and high-level primitives extraction are time consuming, especially in large scale environments.

In order to carry out scalable recognition within a short time, Nister at el. [14] introduced Bag of visual Words (BoWs) technology into object recognition and retrieval. Williams at el. [15] compared different loop closure detection (namely place recognition) approaches and found that image-to-image method with visual words scales best in large environment. Cummnins at el. [16] presented FAB-MAP with SURF point features as visual words to implement place recognition and relocalization in large scale environments. Afterwards they improved FAB-MAP and introduced it into SLAM [17]. Nevertheless, SURF features extraction is time consuming and limits their wide application in robotics.

Gálvez-López at el. [18] implemented faster place recognition using bag of binary words technology with Features from Accelerated Segment Test (FAST) and Binary Robust Independent Elementary Features (BRIEF). By encoding images into binary feature words and using fast extraction features, the system achieved satisfactory real-time performance in large scale environment. But both point features adopted are not rotation and scale invariant, leading to invalidation of the system in obvious appearance and viewpoint changed scenes. Based on [18], Mur-Artal at el. [19] proposed to use binary words consisting of ORB features which can be computed in 10ms with rotation and scale invariant. They then introduced the proposed relocalization and loop closing technology into ORB-SLAM [20]. However, all features used are hand-crafted and extracted from color images which limits their application when they are applied to extremely challenging environments.

B. *CNNs based relocalization*

Chen et al. [21] introduced CNNs combined with spatial and sequential checking into place recognition for the first time. By adopting robust features learned from CNNs, the system can conduct large scale place recognition and improve precise performance significantly compared with the state-of-the-art methods via hand-crafted features. In order to enhance the system robustness in challenging environments, such as severe appearance and viewpoint changes, Sunderhauf et al. [22] [23]

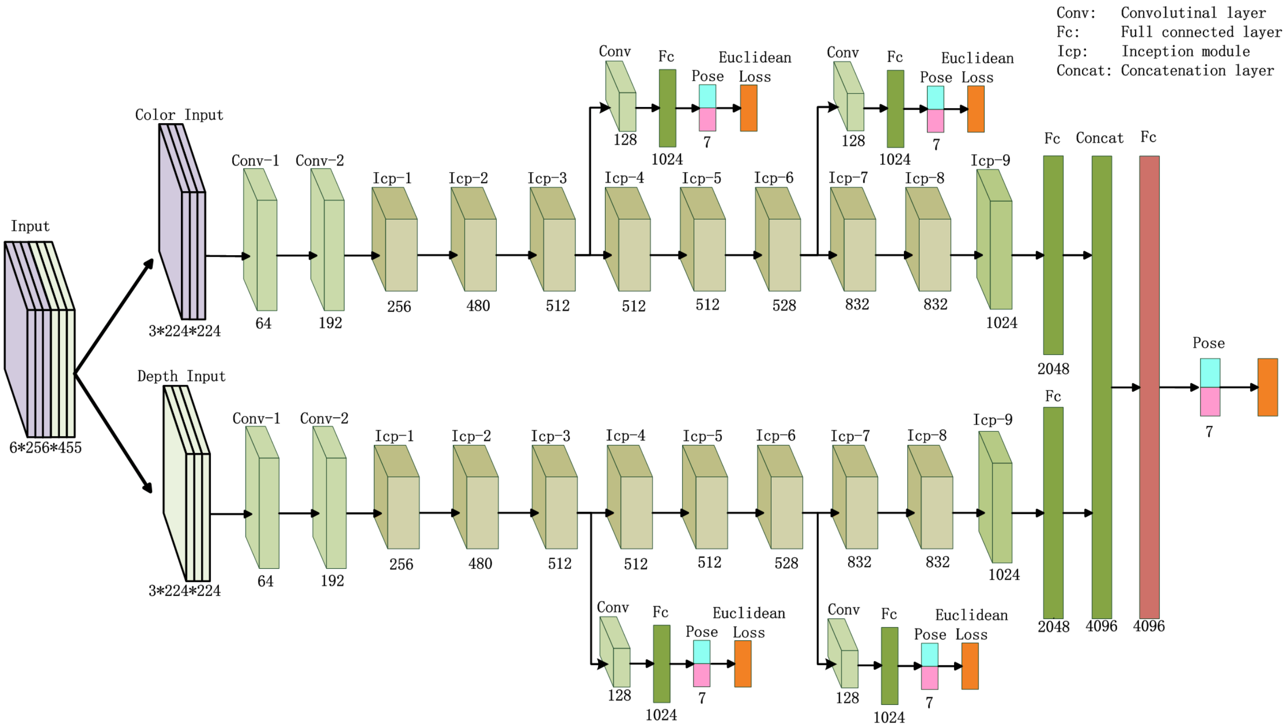


Fig. 1: Architecture overview of the proposed dual-stream CNN for indoor relocalization. Color images and depth images are fed to the network separately as shown above. Each stream of the network outputs a place feature vector with the size of 2048. A concatenation layer and a full connected layer are added to generate an information-rich feature vector with the size of 4096 which, finally converges to camera pose including position and attitude represented by quaternion.

analysed several widely used CNNs and proposed to leverage mid-layer features to cope with appearance variation and top-layer features to handle viewpoint variation. At the same time, by integrating locality-invariance hashing and semantic search space partitioning, the system achieves the real-time performance based on CNNs for the first time.

Regression forest method was introduced into indoor relocalization with Kinect by Shotton at el. [24]. The forest needs to be trained first, then 3D points in local frame could be matched to points within a global map with pre-trained forest which limits its applications in a small area. Precise 6-DOF camera position is calculated with geometry optimization at last.

Kendall at el. [7] proposed a novel convolutional neural network–PoseNet to perform relocalization in large outdoor environments with an end-to-end manner. PoseNet takes color images as the network input and achieves spectacular performance in relocalization. It particularly performs better in large scale and dynamic environments compared with traditional methods. This is the first time to solve the 6-DOF pose regression problem with CNN. The authors improved PoseNet and proposed Bayesian PoseNet [25] afterwards. By adding a Dropout layer into PoseNet and multiple randomized cropping input images, each generated position was modelled with uncertainty and the relocalization precision was improved. Li at el. [26] adopted PoseNet to infer 6-DOF pose in night-time environments with depth camera.

However all of the above methods have not paid attention to indoor complex environments and demonstrated satisfactory performance when encountering night-time, fast movement and dynamic environments.

C. *Encoding method of depth images for CNNs*

CNNs which adopt the power of convolution to learn features from color images have demonstrated spectacular capability on object recognition and detection problems. For depth images used in neural networks, preprocessing is necessary in order to achieve satisfying performance. Couprie at el. [27] introduced original depth information into CNNs to perform semantic segmentation for indoor scenes. Compared with color only images as the network input, it is better to use both color and depth images as the network input for segmentation problems. Aiming at object detection problems with RGB-D inputs, Gupta at el. [28] addressed a novel depth image encoding approach which takes horizontal disparity, height above ground and angle with gravity (HHA) as three image channels. Notice that HHA is computed from the pixels of object in image, which means that this method is not suitable for entire depth image encoding in 6-DOF pose regression. Normalized depth image which encodes scaled surface normal as three image channels is introduced by Lenz at el. [29]

and Hinterstoisser at el. [30] to solve the detection problems. Nevertheless, normalized depth image encoding approach pays much attention to object structural contrast and loses the sight of original absolute range information. In order to achieve better performance in recognition problems with RGB-D cameras, a simple and efficient encoding method that colorizes depth images was proposed. Eitel at el. [31] transformed depth images from single channel to three channels by applying jet colormap. Schwarz at el. [32] rendered depth images with color palette and achieved the colorization of depth images. Experimental results demonstrate that normalized and colorized depth images seem to carry more information than original depth images and are more suitable as the network input. In our paper, we propose a novel encoding approach that takes the advantage of normalized depth images and maintains original range information at the same time for the pose regression problem.

III. PRELIMINARIES

In this part we will introduce the network architecture we used in our paper first. Then we present the working mechanism of our dual-stream CNN for indoor relocalization. At last, a novel effective depth encoding approach for our dual-stream CNN is proposed.

A. Network architecture

The proposed dual-stream CNN is designed to perform robust indoor relocalization in challenging environments such as motion blur, dynamic environments with mobile objects or pedestrians and dark scenes. As shown in Fig. 1, each stream is a separate CNN revised from GoogLeNet [6], which achieved the state-of-the-art performance in ImageNet [33] Large-Scale Visual Recognition Challenge 2014 (ILSVRC14) for object recognition and detection.

In the dual-stream CNN, there are some inception modules which are improved from the modules introduced in Network in network [34]. By using inception modules, one can increase the width of the network without increasing the computational complexity. Compared with similar performance networks without inception modules, the networks with inception modules can obtain faster speed, or they outperform others with similar depth. Except for convolutional layers, full connected layers and inception modules shown in Fig. 1, there are also pooling layers, Rectified Linear Unit (ReLU) layers, Local Response Normalization (LRN) layers, dropout layers and Euclidean loss layers.

The input data composed of color images and depth images are sliced and fed to each stream respectively. In this way, dual-stream CNN can not only learn localization features from color images, but also can learn them from depth images which contain geometrical and structural information. A concatenation layer is used to put all features together and another full connected layer is added for better representations of localization features.

B. Dual-stream CNN for pose regression

For traditional image-to-image or image-to-map pose registration in relocalization, hand-crafted point features should be extracted and matched first, then the transformation will be estimated by minimizing the cost function below along with Random Sample Consensus (RANSAC) to remove feature outliers.

$$min\sum_{i=1}^{n} W_i \left\| \mathbf{X}_i - \mathbf{T}\mathbf{X}_i{}' \right\|^2 \tag{1}$$

Where $\mathbf{T}$ is the 4×4 transformation matrix containing rotation $\mathbf{R}$ and translation $\mathbf{t}$. $\mathbf{X}_i = (x_i, y_i, z_i, 1)^\mathsf{T}$ is the homogeneous position representation of feature point. $\mathbf{X}_i{}' = (x_i{}', y_i{}', z_i{}', 1)^\mathsf{T}$ is the matched point of $\mathbf{X}_i$ in appearance similar image or global map. W_i is the weight of corresponding point. Feature extraction and matching play a crucial role in place recognition and pose estimation especially in challenging situations. Both mismatching and failures in robust feature extraction can result in system failure.

Compared with CNNs designed for object recognition problems, the CNNs for 6-DOF pose regression use a Euclidean loss layer as the top layer instead of softmax classifier layers. Euclidean loss drives the position learning by comparing the network output to the labelled 6-DOF pose and minimizing the least-squared cost. The Euclidean loss E is computed in the network as shown below:

$$E = \frac{1}{2N}\sum_{n=1}^{N} \|\hat{x}_n - x_n\|_2^2 \tag{2}$$

Where x_n is the labelled (ground truth) vector corresponding to the input image, $\hat{x}_n$ is the estimated vector produced by the CNN in an end-to-end manner, and N is the number of vectors that the CNN produced.

In the dual-stream CNN for pose regression, camera pose is a labelled vector composed of position $\mathbf{p} = (p_x, p_y, p_z)^\mathsf{T}$ and orientation represented by unit quaternion $\mathbf{q} = (q_a, q_b, q_c, q_d)^\mathsf{T}$. Therefore the Euclidean loss $E_\mathbf{p}$ is computed as below:

$$E = E_\mathbf{p} + \lambda E_\mathbf{q} \tag{3}$$

Where $E_\mathbf{p}$ is the positional Euclidean loss, $E_\mathbf{q}$ is the orientational Euclidean loss and λ is the balancing weight between these two parts. For the reason that orientation represented by unit quaternion and position measured in meters have obvious different measurement units, a weight λ is introduced as a necessary and significant factor to balance the costs $E_\mathbf{p}$ and $E_\mathbf{q}$ in order to achieve favorable results.

Assume that each image has real robust feature representations $\{y_1, y_2, ..., y_n\}$ to be learned. $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}$ are feature representations our CNN has learned. With an inner product layer $x = \sum_{i=1}^{4096}(\mathbf{W}_i * y_i + b_i)$ as the input of Euclidean layer, the feature representation and corresponding weights can be learned by minimizing the cost below:

$$min\sum_{i=1}^{4096} W_i \|\hat{y}_i - y_i\|^2 \tag{4}$$

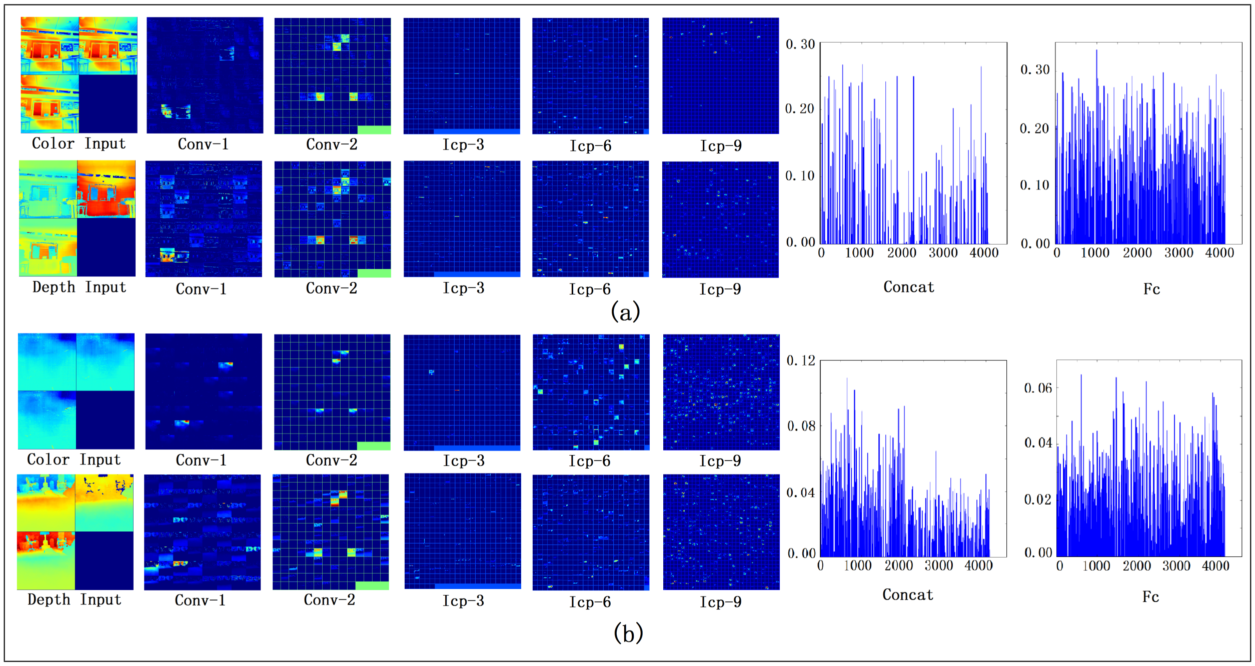


Fig. 2: Selected network layers for different network inputs. (a) Images captured in normal situation as network inputs. Both color images and depth images are in good quality. (b) Images captured at night-time as network inputs. Color images are mostly black and can hardly be recognized even by our humans. Notice that all network weights we used here are the same and pre-trained from images in normal situations.

From the cost above, we can see that the dual-stream CNN will try to learn the real localization features and automatically weight them. This demonstrates a great advantage over the methods using hand-crafted features.

The full connected layer we added after the concatenation layer in our dual-stream CNN plays an active role in balancing the weights of learned features between color inputs and range inputs. As shown in Fig. 2, the concatenation layer connects 2048 color features learned from color images and 2048 range features from depth images separately. When we use the images captured from normal situations, the mean value of texture features from color images is about 0.20, and the mean value of range features is about 0.10. By adding full connected layers and re-weighting all features, the mean value of range features is increased from 0.10 to 0.20. This improves the role of features from depth image as shown in Fig. 2(a). For night-time images, all the color images are mostly black and full of noise. However, their depth images still contain abundant structural range information. As shown in Icp-6 and Icp-9 in Fig. 2(b), the number of bright areas in the color stream are obvious more than that in the depth stream. This represents that the features learned from the color images are more noticeable than the features learned from the depth images. In this way, as shown in the concatenation layer in Fig. 2(b), the mean value of texture features from color images is about 0.08 and that from depth images is about 0.04, which has bad impact for relocalization performance and makes the results uncertain. Fortunately, the full connected layer can re-balance features and decrease the mean value of color texture features from 0.08 to 0.04 which reduces their weights. The network is smart enough to weight different features according to different situations, i.e. it strengthens the weights of texture features in normal situations and weakens the weights of structural features in challenging situations such as night-time. In this way, by combining the texture features from color images and the structural range features from depth images and weighting them automatically, the dual-stream CNN can cope with many extremely challenging environments.

C. *Encoding methods for depth images*

In the dual-stream CNN, depth images from RGB-D cameras are introduced for a separate stream CNN to learn structural range features in order to enhance the system performance. However, the convolutional layers in CNNs are particularly designed for color images in which pixel channels mainly represent light intensity. In contrast, the pixel value of depth images represents scaled distance between the optic centre of camera and objects in the environment. We can hardly achieve satisfied performance with raw depth images used as the inputs of CNNs directly. Necessary preprocessing for depth images ought to be taken before feeding them into CNNs.

The simplest preprocessing method for depth images is to rescale pixel values from 0–10000 to 0–255. The processing result of this method is the single layer grayscale image (one channel) that is shown in Fig. 3(b). To leverage the transfer learning from the network weights pretrained on ImageNet

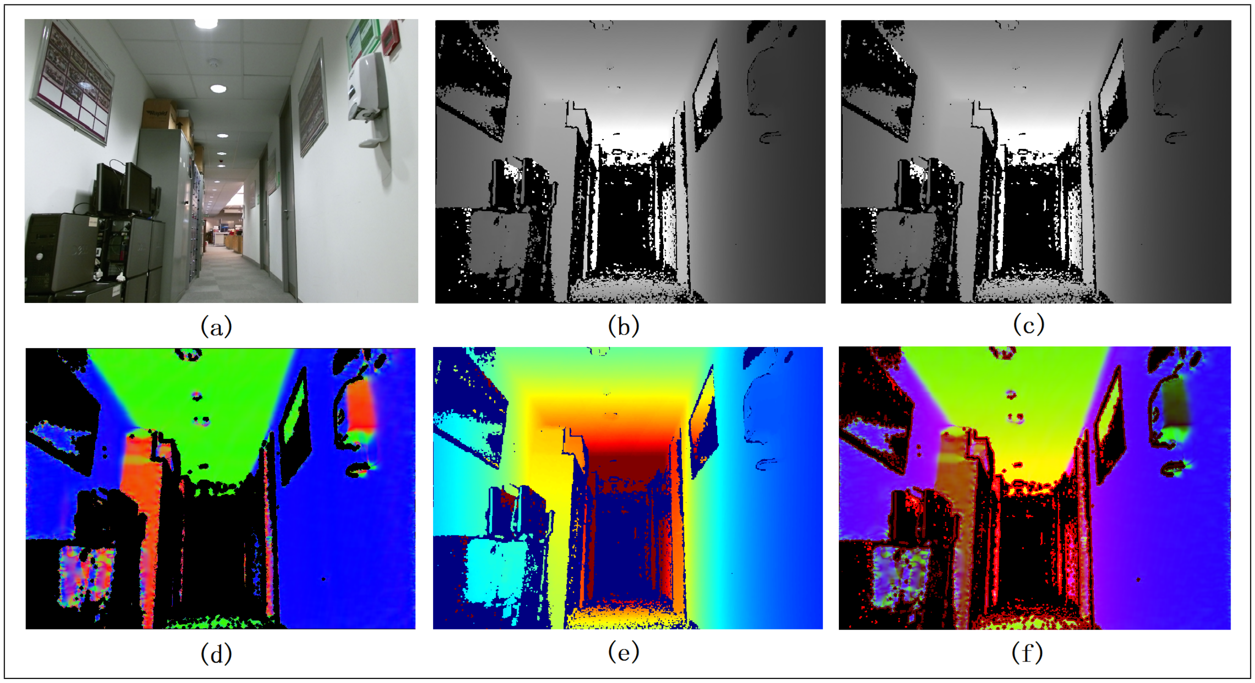


Fig. 3: Different encoding methods of depth images. (a) Color image. (b) Single layer scaled depth image. (c) Triple-layer scaled depth image. (d) Normalized depth image with computed normal parameters as three image channels. (e) Colorized image using jet color map. (f) Minimized normal+depth(MND) image.

or Places [35], we can also duplicate the grayscale image shown in Fig. 3(b) and create three copies of the depth layer. Then triple-layer scaled depth image (three channels) could be obtained that is shown in Fig. 3(c). As shown above, the appearance of processed images between these two methods has no significant difference. Another famous encoding method is to use surface normals as three channels of images [29] [30]. The unit surface normals $[n_x, n_y, n_z]$ should be computed from depth image first, then they are rescaled from 0–1 to 0–255. The normalized image is shown in Fig. 3(d). Colorization of depth images is also an easy but efficient way for depth encoding [31] [32]. Each pixel intensity value in depth image corresponds to three channel values (red, green and blue separately). Fig. 3(e) is a colorized jet map, as the intensity value increases from 0 to 255, the color changes from blue to green, then yellow to red.

From Fig. 3(d), we can see that all pixels on the wall have the same color even though they have different depth values. On the contrary, the wall pixels with discriminative depth hierarchies in colorized image are labeled with different colors as shown in Fig. 3(e). However, the box on the right part of Fig. 3(e) shows little difference on structure and the box in Fig. 3(d) is easy to distinguish. In a word, the normalized images pay more attention to relative structural information, so the objects in the normalized images could be easily distinguished. Nevertheless, the normalized images do not contain absolute range information, which is contained in original depth images and colorized images.

In this paper, we propose to use the minimized normal and depth images (MND) to encode depth images. The MND uses rescaled n_x', n_y' and depth d as three channels of image separately. For scaled surface normal $[n_x', n_y', n_z']$, we have $n_x'^2 + n_y'^2 + n_z'^2 = 255^2$. Therefore normalized depth images can be represented by n_x' and n_y' which is called minimized normal representation. The third channel we use in our approach is the scaled original depth value. As shown in Fig. 3(f), the pixels of wall with similar normal and the pixels of box with similar depth both have more obvious local contrast when using our MND encoding method. The advantage of this method is that processed depth images retain both relative structural information and absolute range information and the networks maintain the ability to leverage the transfer learning at the same time.

Another problem we need to consider when using depth images for CNNs is the input size. The dual-stream CNN we use in our paper requires an input image with a size of 224×224. The raw depth image captured from Kinect One is 960×540 and that captured from Kinect 360 or Xtion is 640×480. In our system, original images are resized to 455×256 or 341×256, and then randomly cropped to a size of 224×224. We have also tried to resize original depth images to 224×224 directly and maintain all information from depth, but found that the relocalization performance was not as good as expected.

In addition, invalid depth data is inevitable when using RGB-D sensors (Kinect v1, Kinect v2, Xtion and et al.). We briefly treat invalid depth data with zeros and let the network learn to handle this data. There could be some more elegant ways to explore for invalid depth data, such as using spatial or temporal interpolation techniques. This is left for our future research.

IV. Training Mechanism for Dual-stream CNN

The selection of training mechanism is very important for system performance. For traditional CNNs, we can train them in an end-to-end way and optimize the network weights directly. For our dual-stream CNN for relocalization, traditional training methods can hardly ensure the convergence of the dual-stream CNN. In this work, the dual-stream CNN training for relocalization is divided into three stages.

A. Training separate streams

The first training stage for our dual-stream CNN is a separate stream training stage. Different from the network architecture shown in Fig. 1, the network in this stage does not have the concatenation layer and the additional full connected layer. The stream with color images as inputs is called color stream, and the stream with depth images as inputs is called depth stream. The Euclidean Loss layer is appended to the end of both color stream and depth stream. Therefore, the network in this stage has two pose outputs, one is produced by 2048 textural features from the stream with color images as inputs, another one is generated by 2048 range features from the stream with depth images as inputs. The network here has one input and two outputs and the two streams can be trained simultaneously. Both streams take transfer learning from the same network weights pretrained on the dataset Places [35] used for training place classifiers. Although the purpose of two channels are different, both streams in our network can converge in a short time with desirable performance by using transfer learning. Different Euclidean balancing weights are adopted for the color stream and the depth stream while training. For color stream, we use 1 for positional Euclidean loss weight and λ for orientational Euclidean loss weight. Weights for depth stream are 1.2 time weights for color stream, i.e. the positional Euclidean weight is 1.2 and the orientational Euclidean weight is 1.2λ. The depth stream is endowed with heavier weights for the reason that depth images are not affected by light intensity or motion blur and are more robust than color images to some extent.

B. Fine tuning full connected layer

After the separate stream training stage, the network can learn many preliminary position features from depth images and color images respectively. In this stage, we recovery the network architecture as shown in Fig. 1. The concatenation layer and full connected layer are brought back to the network to re-weight preliminary features learned from two streams.

The network weights obtained from stage one are used for transfer learning. We only perform fine tuning operation on the full connected layer while the weights of other layers are kept invariant. This is done by setting all learning rates to zeros in the network except for the last full connected layer. By fine tuning the full connected layer, the relocalization performance of the system is improved greatly compared with single color stream or single depth stream. Additionally, for the reason that only the full connected layer is fine tuned, the number of iteration in this stage is only about 10000 to 15000 which means a short time is required.

TABLE I: Relocalization performance comparison with different depth image encoding methods using PoseNet. The unit for position error is meter (m), and the unit for orientation error is degree (°).

Input	Median position error	Median orientation error
RGB	0.55m	5.13°
Single depth	0.57m	3.83°
Triple-depth	0.52m	3.36°
Normalized depth	0.49m	3.02°
Colorized depth	0.48m	3.10°
MND	**0.46m**	**2.85°**

C. Fine tuning overall dual-stream CNN

In order to achieve the best performance for indoor relocalization, we fine tune the overall dual-stream CNN in the final stage. The network pretrained in the second stage is taken for transfer learning in the third stage. All learning rates set to zeros in the second training stage are set back to the same value in the first training stage. Base learning rate should be smaller than that in the first two stages, which is 0.8 times in our paper. This stage needs about 20000 iterations to converge. Position features and their weights are adjusted slightly. The system performance is further improved compared with that in the second stage.

V. Experimental Evaluation

In this section, we evaluate the relocalization performance of our proposed dual-stream CNN. We first compare different depth encoding methods and select the best one for 6-DOF pose regression with CNN. Then the architectures to take the advantage of both color images and depth images are discussed. A large scale relocalization experiment is presented. Following that, the quantitative experiments based on Microsoft 7-scenes benchmark dataset is given by comparing with PoseNet [7]. At last, the experiments on extremely challenging situations with our system is presented. The dual-stream CNN is designed using Caffe [36], and all experiments are performed on a desktop equipped with Nvidia GeForce Titan X GPU card and Intel Core i7-4790 4.0GHz CPU.

A. Range images encoding methods

In this part, we compare the proposed MND depth encoding method with other popular encoding methods. The dataset called Second Floor here is collected in the second floor of our network building using Kinect One. The scale of the second floor is about $40m \times 30m$. For the reason that we do not have motion capture system covering the whole floor, we use the state-of-the-art SLAM algorithm–ORB SLAM [20] to label collected images. The labelled result is also used as the groundtruth. The training dataset contains about 4300 frames and the test dataset contains about 5460 frames.

For the network training, we only perform the separate stream training in the process and take the relocalization results from color stream and depth stream, respectively. The stochastic gradient descent (SGD) is adopted as the training solver. Learning rate policy is step with 80 epochs as step size. The base learning rate is 0.00001, the gamma is 0.94 and the momentum is 0.9. Both streams take transfer learning and the number of training iteration is about 30000. For color stream, the positional weight is 1 and the orientational weight λ is 250. For depth stream, the positional weight is 1.2 and the orientational weight λ is 300. We found that it is important to give the depth stream a little bigger weight to make our system robust.

The relocalization results from the depth stream with different encoding depth images and color stream are shown in Table. I. We can see from the table that the depth stream performs better than the color stream in orientation. Normalized depth images and colorized depth images have no significant difference in relocalization as the network inputs, but both are better than single depth and triple-depth. From our experiments, the normalized encoding performs better in position and the colorized encoding performs better in orientation. Our proposed MND method which includes both structural and range information achieves the best performance in indoor relocalization since the MND has the most significant local contrast among all the depth encoding methods.

B. Network architecture for RGB-D images

After selecting the MND encoding as our depth preprocessing method, we will further discuss the network architecture taking both color images and depth images as input. The same dataset (Second Floor) as used above, is used here. As shown in Table II, four different CNNs are presented with RGB-D images as inputs. We first feed color+single-depth images (4 layers as inputs) and color+MND depth (6 layers as inputs) to PoseNet, respectively. The results show that PoseNet with color+single-depth as inputs performs better in relocalozation than that with color+MND depth. So we know that the localization performance of CNNs can not be improved with the network width if we can not find the right way. Then the dual-stream CNN taking color and MND depth as inputs is used to perform indoor relocalization. The performance of the dual-stream CNN is much better than PoseNet due to the mix of color and depth images. As mentioned above, we divide the network training into three stages. After fine tuning the full connected layer, we also fine tune the whole network in the third training stage. This step further improves the relocalization performance in both position and orientation. In addition, we also try to concatenate the full connected layers after Icp-3 (regression 1) and the full connected layers after Icp-6 (regression 2) respectively, and fine tune the new concatenated networks. The relocalization precision for these two concatenated networks is 1.19m, 5.76°and 1.07m, 4.04°, respectively. So concatenating the final features is optimal in our network.

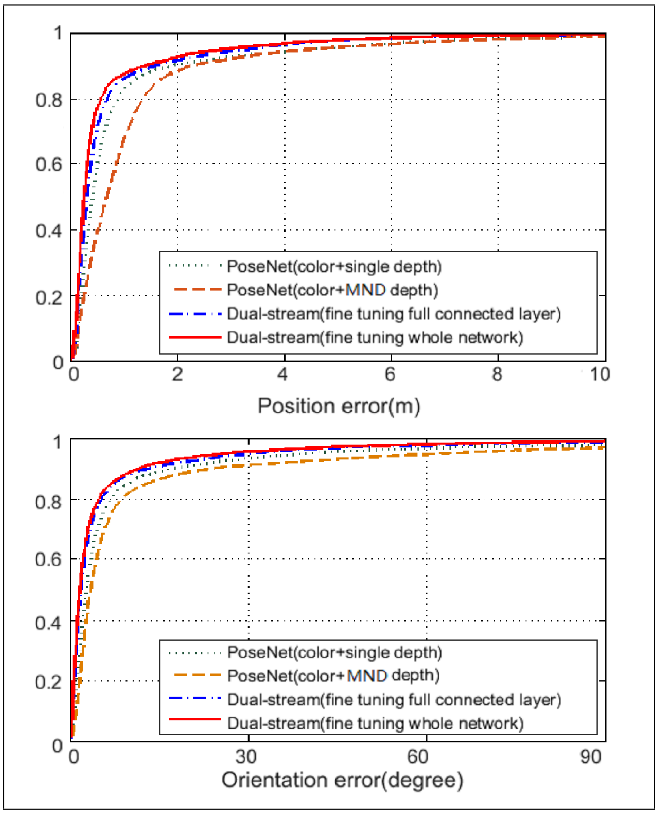


Fig. 5: Cumulative histograms of relocalization error with different network architectures for RGB-D images as input.

The cumulative histograms of indoor relocalization error produced by the above four approaches are plotted in Fig. 5. The dual-stream CNN after fine tuning the whole network performs best among them in all the aspects.

2D relocalization trajectory with Second Floor dataset is shown in Fig. 4. Fig. 4(a) shows the groundtruth trajectory of the training dataset and the test dataset. The dashed line represents the trajectory of training dataset. The solid line represents the goundtruth trajectory of test dataset. Fig. 4(b) shows the groundtruth trajectory and predicted poses of the test dataset, and we put them in the 2D floor plan of our building in order to show them clearly. Some images of the scenes are given. The dots represent the predicted positions from our dual-stream CNN. As shown in Fig. 4, most predicted camera poses are almost the same with the groundtruth. However, there are also a few predicted poses which have considerable error. This could be improved by extending our network to a Bayesian dual-stream CNN, which can model the uncertainty of poses and remove noisy data points, but the Bayesian dual-stream

TABLE II: Results of different network architectures with RGB-D images as input. The first row in the table shows the result of PoseNet [7] with color image (3 layers) as input. The second row shows the result of PoseNet with color image and single scaled depth image (4 layers) as input. The third one shows the result of PoseNet with color image and colorized depth image (6 layers) as input. The forth row and the fifth row both use the dual-stream CNN for indoor relocalization. The difference is that last two full connected layers are fine tuned for the third row and the whole network is fine tuned for the forth row.

Network architecture	Scale	Train/ Test	Input	Median position error	Median orientation error
PoseNet	40×30×5m	4300/ 5460	RGB (3)	0.55m	5.13°
PoseNet	40×30×5m	4300/ 5460	RGB+single-depth (4)	0.42m	2.77°
PoseNet	40×30×5m	4300/ 5460	RGB+MND depth (6)	0.68m	3.51°
Dual-stream CNN (Fine tuning full connected layers)	40×30×5m	4300/ 5460	RGB-MND depth (3+3)	0.33m	1.83°
Dual-stream CNN (Fine tuning whole network)	40×30×5m	4300/ 5460	RGB-MND depth (3+3)	**0.27m**	**1.65°**

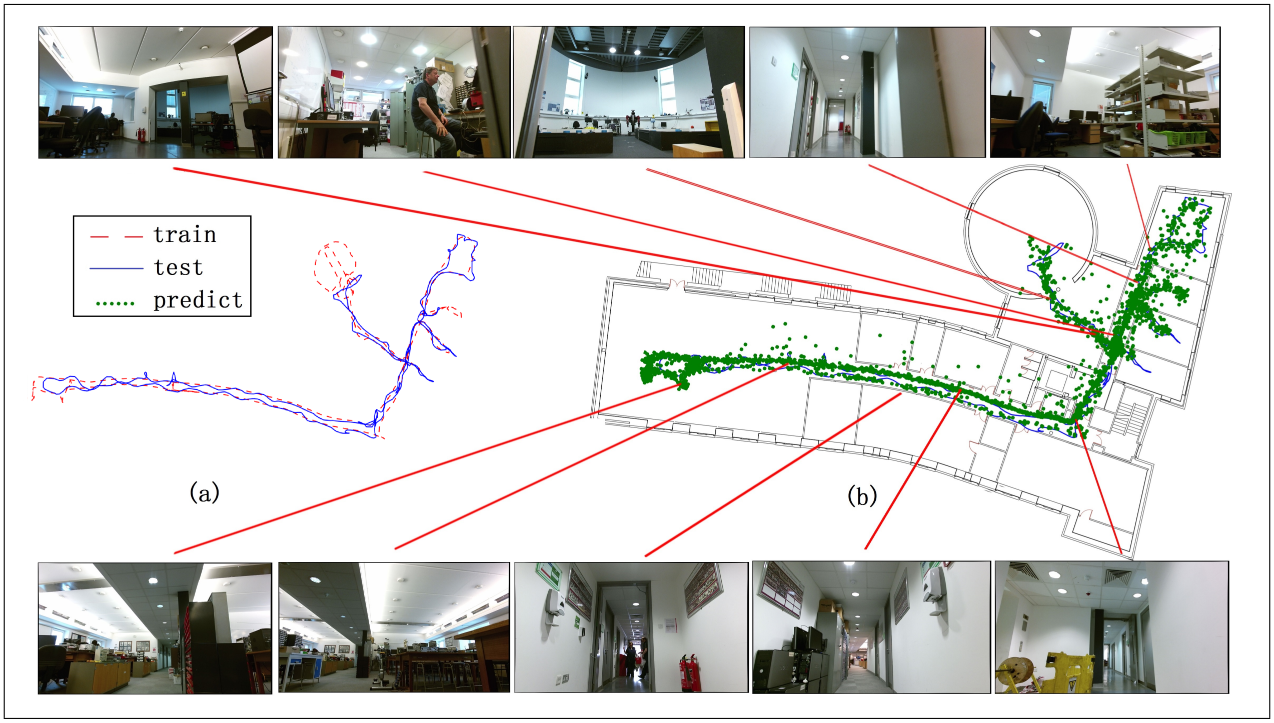


Fig. 4: Predicted trajectory using our method in our network building. The red trajectory (dashed line) is used for training, the blue one (solid line) is the ground truth of test dataset. The green one (dots) is the pose predicted by the dual-stream CNN.

will cost more time.

C. Quantitative analysis

In this part, we compare our dual-stream CNN with PoseNet. We also transform the dual-stream CNN to a Bayesian dual-stream CNN and compare it with the Bayesian PoseNet [25]. The transformation is very similar with that from PoseNet to Bayesian PoseNet. The Dropout layers are added to color stream and depth stream separately, the network inputs are randomly cropped from preprocessed images for 30 times, and the final result takes the average of all 30 network outputs.

We take the public Microsoft 7-Scenes dataset as benchmark to compare the system performance in indoor relocalization. As shown in Table III, PoseNet with color images and Posenet with depth images have no significant difference in indoor relocalization, sometimes the former with texture features performs better, sometimes the latter with range features performs better. Our proposed dual-stream CNN which takes both color images and depth images together achieves better performance than

TABLE III: Quantitative analysis. Comparison with PoseNet and Bayesian PoseNet based on the public 7-Scenes dataset downloaded from Microsoft Research. Here the median relocalization error is used to represent the system performance. The scale for Fire, Heads, Chess, Pumpkin, Office, Redktichen, Stairs dataset is 2.5×1×1m, 2×1×0.5m, 3×2×1m, 2.5×2×1m, 2.5×2×1.5m, 4×3×1.5m, 2.5×2×1.5m and 2.7×1.8×1.1m respectively. The training/test numbers for above 7-scenes dataset are 2000/2000, 1000/1000, 4000/2000, 4000/2000, 6000/4000,7000/5000 and 2000/1000 respectively.

Dataset	PoseNet (RGB-D)	PoseNet (RGB+MND)	PoseNet (RGB)	PoseNet (MND)	Dual-stream (RGB-MND)	Bayesian PoseNet (RGB)	Bayesian PoseNet (MND)	Bayesian Dual-stream (RGB-MND)
Fire	0.56m, 13.57°	0.58m, 17.06°	0.52m, **12.54**°	0.55m, 16.76°	**0.51m**, 12.88°	**0.42m**, 12.65°	0.46m, 16.83°	0.43m, **12.52**°
Heads	0.39m, 15.25°	0.33m, 15.10°	0.38m, 13.46°	0.34m, 14.77°	**0.30m, 12.73**°	0.27m, 13.06°	0.26m, 14.77°	**0.25m, 12.72**°
Chess	**0.36m, 6.91**°	0.38m, 7.65°	0.39m, 8.05°	0.45m, 9.68°	**0.36m**, , 7.79°	0.29m, 7.33°	0.32m, 9.02°	**0.28m, 7.05**°
Pumpkin	0.51m, 8.43°	0.48m, **8.16**°	0.58m, 9.20°	0.52m, 8.68°	**0.45m**, 8.30°	0.49m, 8.57°	0.37m, 8.19°	**0.36m, 7.53**°
Office	0.63m, 12.56°	0.61m, 13.22°	0.56m, **9.39**°	0.54m, 12.54°	**0.48m**, 9.68°	0.38m, **8.73**°	0.36m, 11.72°	**0.30m**, 8.92°
Redkitchen	0.94m, 18.21°	0.73m, 13.30°	0.87m, 11.40°	0.63m, 12.46°	**0.58m, 10.49**°	0.75m, 10.62°	0.51m, 11.65°	**0.45m, 9.80**°
Stairs	0.53m, **11.94**°	0.63m, 13.79°	0.54m, 13.71°	0.63m, 14.40°	**0.48m**, 13.21°	0.47m, 13.71°	0.59m, 14.01°	**0.42m, 13.06**°
Average	0.56m, 12.41°	0.53m, 12.61°	0.55m, 11.10°	0.52m, 12.75°	**0.45m, 10.72**°	0.44m, 10.67°	0.41m, 12.31°	**0.35m, 10.22**°

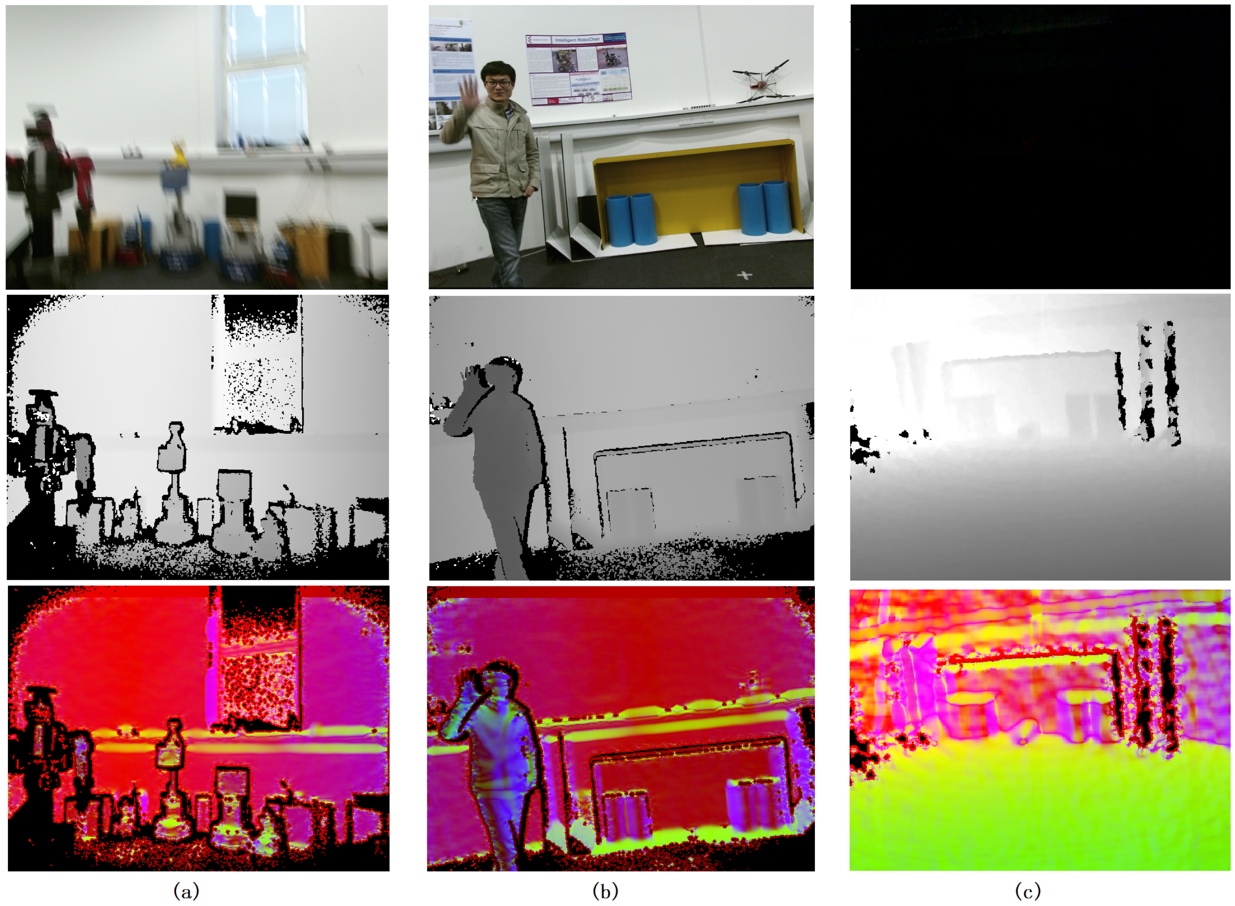
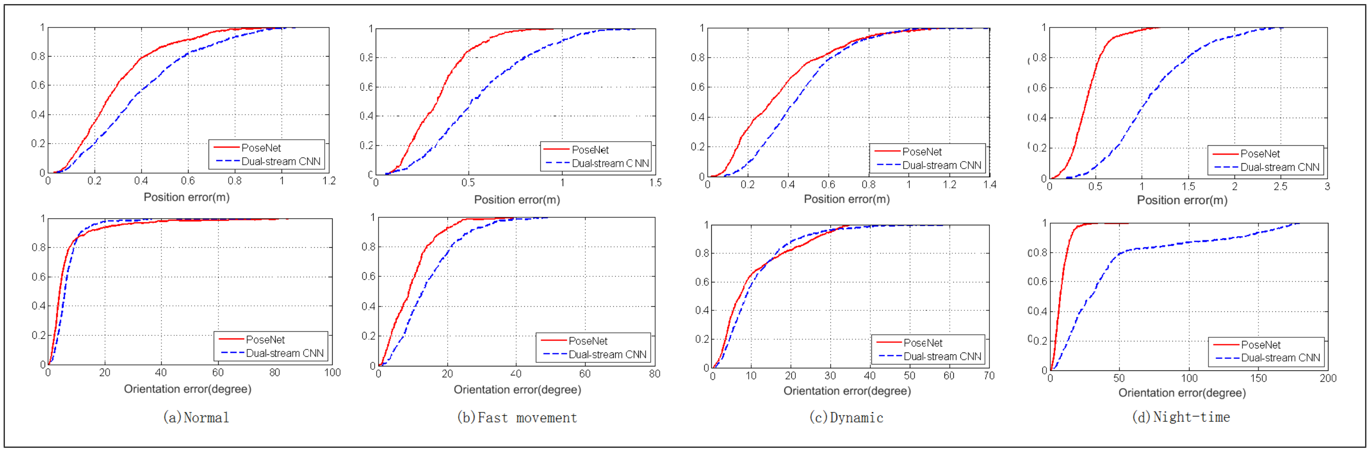


Fig. 6: Challenging scenes for indoor relocalization with the dual-stream CNN. The images in the first row are color images. The second row is the scaled depth image and the third row is the MND depth image. (a) Motion blur produced by fast movement. (b) Dynamic scenes with pedestrian or other moving objects. (c) Night-time or dimly indoor environments.

TABLE IV: Indoor relocalization results in challenging scenes with different methods. Our system shows about 30%–70% improvement in precise compared with PoseNet in these challenging but everyday situations.

Dataset	Scale	Train/ Test	Network architecture	Input	Median position error	Median orientation error
Normal	4×4×1m	5540/ 766	PoseNet	RGB	0.36m	6.08°
Normal	4×4×1m	5540/ 766	Dual-stream CNN	RGB-MND depth (3+3)	**0.26m**	**4.32**°
Fast movement	4×4×1m	5540/ 357	PoseNet	RGB	0.52m	13.03°
Fast movement	4×4×1m	5540/ 357	Dual-stream CNN	RGB-MND depth (3+3)	**0.31m**	**9.00**°
Dynamic	4×4×1m	5540/ 964	PoseNet	RGB	0.44m	9.82°
Dynamic	4×4×1m	5540/ 964	Dual-stream CNN	RGB-MND depth (3+3)	**0.32m**	**8.62**°
Night-time	4×4×1m	3473/ 656	PoseNet	RGB	1.24m	29.10°
Night-time	4×4×1m	3473/ 656	Dual-stream CNN	RGB-MND depth (3+3)	**0.39m**	**7.48**°



(a)Normal (b)Fast movement (c)Dynamic (d)Night-time

Fig. 7: Cumulative histogram of relocalization error for challenging indoor test datasets with the dual-stream CNN. The dual-stream CNN shows significant advantages over PoseNet especially in the challenging situations.

PoseNet. So does Bayesian dual-stream CNN when compared with Bayesian PoseNet. When comparing our dual stream CNN with PoseNet, both position accuracy and orientation accuracy have gained more than 15% improvement as shown in table II and table III.

D. Qualitative analysis based on challenging datasets including fast movement, night-time, dynamic scenes

This part introduces the advantage of our dual-stream CNN in indoor relocalization when encountering challenging situations which other methods can hardly deal with.

All the datasets are collected in our Robot Arena lab equipped with a motion capture system which can provide the groundtruth. For night-time scenes, we use Asus Xtion to collect RGB-D images. For other datasets, we use Kinect Xbox one. Notice that all the training datasets are collected in normal environments which means there are no fast movement of camera, moving objects and dark scenes when collecting the training dataset. The dual-stream CNN will infer the original lighting, shape, texture from images when appearance changes, and then estimate the camera poses.

Fig. 6 shows the color images, depth images and MND depth images in three challenging scenes. Fast movement of the camera, dynamic scenes including pedestrians or other moving objects, night-time scenes are really challenging for visual localization, but they happen everyday and everywhere in our life. From the figure we can see that the color images in these situations are in particularly bad quality. On the contrary, the depth images maintain all range features without much information loss. Our dual-stream CNN combining textural, structural and range features outperforms PoseNet significantly when encountering these challenging situations. From Table IV, we can see that there are approximately 27%–68% and 12%–64% improvements to position precision and orientation precision, respectively, with our system when compared with PoseNet. The cumulative histograms of indoor relocalization error in these situations are plotted in Fig. 7. Our system performs especially well when faced with fast movement of the camera and night-time scenes on account of introducing the depth stream.

We have also tried to use SLAM technology with hand-crafted features to perform localization in these challenging environments, but found that most camera poses will be lost in dynamic and motion blur situations, and the algorithm will completely lose efficacy in night-time environments.

VI. CONCLUSIONS

In this paper, we have presented a novel dual-stream CNN for 6-DOF pose regression which shows spectacular performance in large scale indoor relocalization. A novel depth encoding method that takes the minimized normal and depth (MND) as processed depth image is addressed. Compared with other encoding methods such as normalized depth or colorized depth, our MND encoding approach presents comparable performance in relocalization. Moreover, by introducing depth information with a separate stream, our network could learn both texture features from color images and structural range features from depth images. In this way, the relocalization precision is improved compared with PoseNet, and the system robustness is greatly enhanced when faced with challenging environments such as fast movement, dynamic objects and night-time which other algorithms demonstrate poor performance. However, when Kinect is faced with direct sunlight, the depth images will be in bad quality, the localization performance will degrade and be in greater uncertainty. Although our network can implement localization in many challenging situations, the localization accuracy produced by our model is not as good as geometric method in normal scenes. In the future, we would like to use features learned from the dual-stream CNN instead of hand-crafted features in SLAM to expand to the application of visual localization.

ACKNOWLEDGMENT

The authors would like to thank Robin Dowling and Ian Dukes for his support in experiments. We thank Sen Wang from University of Oxford, Ian Lenz from Cornell University and Jiaxiang Wu from Chinese Academy of Sciences for their valuable comments and discussion. We also thank Alex Kendall from University of Cambridge for sharing his code of PoseNet. The first three authors have been financially supported by scholarship from China Scholarship Council.

REFERENCES

[1] C. Ye, S. Hong, and A. Tamjidi, "6-DOF pose estimation of a robotic navigation aid by tracking visual and geometric features," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 4, pp. 1169–1180, 2015.

[2] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *Robotics & Automation Magazine, IEEE*, vol. 13, no. 2, pp. 99–110, 2006.

[3] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping: Part II," *IEEE Robotics & Automation Magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[4] Q. Liu, R. Li, H. Hu, and D. Gu, "Extracting semantic information from visual data: A survey," *Robotics*, vol. 5, no. 1, p. 8, 2016.

[5] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[7] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-DOF camera relocalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2938–2946.

[8] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *Robotics, IEEE Transactions on*, vol. 32, no. 1, pp. 1–19, 2016.

[9] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Robotics-DL tentative*. International Society for Optics and Photonics, 1992, pp. 586–606.

[10] B. Steder, G. Grisetti, and W. Burgard, "Robust place recognition for 3D range data based on point features," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1400–1405.

[11] R. Li, Q. Liu, J. Gui, D. Gu, and H. Hu, "A novel RGB-D SLAM algorithm based on points and plane-patches," in *Automation Science and Engineering (CASE), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1348–1353.

[12] R. Cupec, E. K. Nyarko, D. Filko, A. Kitanov, and I. Petrović, "Place recognition based on matching of planar surfaces and line segments," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 674–704, 2015.

[13] E. Fernández-Moral, P. Rives, V. Arévalo, and J. González-Jiménez, "Scene structure registration for localization and mapping," *Robotics and Autonomous Systems*, vol. 75, pp. 649–660, 2016.

[14] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 2161–2168.

[15] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós, "A comparison of loop closing techniques in monocular SLAM," *Robotics and Autonomous Systems*, vol. 57, no. 12, pp. 1188–1197, 2009.

[16] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[17] ——, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[18] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *Robotics, IEEE Transactions on*, vol. 28, no. 5, pp. 1188–1197, 2012.

[19] R. Mur-Artal and J. D. Tardós, "Fast relocalisation and loop closing in keyframe-based SLAM," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 846–853.

[20] R. Mur-Artal, J. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *Robotics, IEEE Transactions on*, vol. 31, no. 5, pp. 1147–1163, 2015.

[21] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *arXiv preprint arXiv:1411.1509*, 2014.

[22] N. Sunderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 4297–4304.

[23] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free," *Proceedings of Robotics: Science and Systems XII*, 2015.

[24] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.

[25] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 4762–4769.

[26] R. Li, Q. Liu, J. Gui, D. Gu, and H. Hu, "Night-time indoor relocalization using depth image with convolutional neural networks," in *Automation and Computing (ICAC), 2016 22nd International Conference on*. IEEE, 2016, pp. 261–266.

[27] C. Couprie, C. Farabet, L. Najman, and Y. Lecun, "Indoor semantic segmentation using depth information," in *First International Conference on Learning Representations (ICLR 2013)*, 2013, pp. 1–8.

[28] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 345–360.

[29] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[30] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of textureless objects in heavily cluttered scenes," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 858–865.

[31] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 681–687.

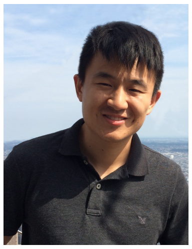
[32] M. Schwarz, H. Schulz, and S. Behnke, "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1329–1335.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[34] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[35] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[36] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.



Ruihao Li received the B.Sc. degree in automation from Beijing Institute of Technology, Beijing, China, in 2012 and the M.Sc. degree in control science and engineering in National University of Defense Technology, Changsha, China, in 2014. He is currently pursuing the Ph.D. degree in robotics at the University of Essex, UK. His research interests are in robotics, SLAM, deep learning, semantic scene understanding.

Qiang Liu received his B.Sc. and M.Sc. degrees in automation from Harbin Institute of Technology, Harbin, China, in 2012 and 2014 respectively. Currently, he is pursuing his Ph.D. studies in robotics at the University of Essex, UK. His research interests are in robotics, deep learning, semantic mapping and relocalization, place and object recognition.

Jianjun Gui received the B.Sc. degree in simulation engineering from the National University of Defense Technology, Changsha, China in 2011 and continued the M.Sc. degree in control science and engineering until 2013. He is currently pursuing the Ph.D. degree in robotics at the University of Essex, Colchester, UK. His research interests include robotics, computer vision, SLAM, visual inertial odometry, robot pose estimation and visual servoing.

Dongbing Gu received the B.Sc. and M.Sc. degrees in control engineering from Beijing Institute of Technology, Beijing, China, and the Ph.D. degree in robotics from University of Essex, Essex, UK. He was an Academic Visiting Scholar with the Department of Engineering Science, University of Oxford, Oxford, UK, from October 1996 to October 1997. In 2000, he joined the University of Essex as a Lecturer. Currently, he is a Professor with the School of Computer Science and Electronic Engineering, University of Essex. His current research interests include robotics, multiagent systems, cooperative control, model predictive control, visual SLAM, wireless sensor networks, and machine learning.

Huosheng Hu received the M.Sc. degree in industrial automation from Central South University, Changsha, China, in 1982 and the Ph.D. degree in robotics from the University of Oxford, Oxford, UK, in 1993. He is currently a Professor with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK, leading the Human Centred Robotics Group. His research interests include autonomous robots, human-robot interaction, multi-robot collaboration, pervasive computing, sensor integration, intelligent control, cognitive robotics, and networked robots. He is a Fellow of Institute of Engineering & Technology and Institution of Measurement & Control in the UK, a senior member of IEEE and ACM, and a Chartered Engineer. He is currently an Editor-in-Chief for the International Journal of Automation and Computing, Founding Editor-in-Chief for Robotics Journal and an Executive Editor for International Journal of Mechatronics and Automation.