FROM MONOCULAR TO LEARNED vSLAM

Authors: Usama Maqbool[1], Hammad Munawar[1], Abdur Rahman[1]

[1] College of Aeronautical Engineering, Risalpur, Pakistan
usamamaqbool74 @cae.nust.edu.pk (Usama Maqbool), h.munawar@cae.nust.edu.pk (Hammad Munawar), abdurrahman@cae.nust.edu.pk (Abdur Rahman)

Received: 28-September-2021 / Revised: 09-December-2021 / Accepted: 20-December-2021
Karachi Institute of Economics and Technology || Technology Forces Journal, Issue 2, Volume 3, 2021

ABSTRACT

Size, Weight and Power (SWaP) constraints in robotics cause vSLAM strategies to prefer using monocular cameras due to their high information-to-weight ratio and miniature size. Conventional monoSLAM methodologies compete with stereo and RGB-D SLAM on the front of localization; however, 3D reconstruction of the environment is limited to sparse point clouds. In this paper, firstly, we have reviewed challenges (inherent scale ambiguity and map initialization) that have emerged in Conventional monoSLAM due to the fact that depth information of the scene is lost. This has eventually led to the development of deep learning based vSLAM strategies. Secondly, Learned vSLAM strategies (amalgam of CNNs with conventional vSLAM strategies), their eminence over conventional monoSLAM and impeding limitations of deep learning architectures have been reviewed extensively. Reviewed strategies include CNN SLAM, Scale-aware monocular SLAM, CNN SVO, DTAM (Dense Tracking And Mapping), Online Adapted Depth Prediction, and Sparse2Dense (S2D). By the end we have discussed the future prospects of Learned vSLAM which can be explored further.

Key Words: Learned vSLAM, monoSLAM, ConvNets, Predicted depth, SLAM

1. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) has been one of the most researched area in computer vision because of its wide-ranging applications in robots aiding military, space exploration, security and surveillance, disaster relief operations, underwater ridge mapping and even autonomously operating household appliances. It provides the basic building block for any autonomous strategy as it solves the localization problem in GPS denied environments and mapping the workspace for Situational Awareness of the user. SLAM is composed of several steps, it can be defined in the most generalized way as "feature extraction and association from perception through sensors for updated state estimation and information accumulation."

While these steps involve theoretical formulation of arduous mathematical equations, the basic objective behind all stays the same. Execution of each step critically influences the performance and complexity of SLAM but the most important aspect is the selection of sensors that collect workspace information and directly affect the precision level of the generated result. Therefore, several sensory options have been explored by the research community with the objective of accurate pose estimate and detailed mapping of the environment. Options include Laser scanners [1]– [3], IR sensors [4], monocular [5], [6], stereo [2], [3] and RGBD cameras [7], [8]. However, all these sensors carry limitations in their domain. Laser scanners and IR sensors provides limited information of the workspace and their performance deteriorates in cluttered environments therefore are used in conjunction with vision-based sensors [1], [2], [4], [7], [8]. SLAM that uses visual sensors (cameras) for perception goes by the name of visual SLAM (vSLAM).

Amongst visual sensors, Stereo cameras are a popular choice for accurate odometry of robots. State of the extracted features can be estimated from a single capture by the two separate cameras which present a stereo baseline between similar features. 3D feature position estimates are used for matching to subsequent captured frames for sequential update to robot position estimate and mapping. However, a sparse or semi-dense map is produced depending upon the tracked features. Depth estimates of features at large distances cannot be computed because of the relatively small offset between cameras [9], [10]. Monocular cameras have attracted lot of research due to their common availability and high information-to-weight ratio. Complexity arises in their usage, as a single image does not provide full state of the extracted features which causes scale ambiguity. Monocular SLAM strategies are required to be initialized first by defining the coordinate system of the environment before tracking and mapping starts. However, there are still limitations of calculating a metric scale and reduced stereo baseline under pure rotation [11]. With the availability of low cost RGB-D cameras like the Microsoft Kinect, depth estimation task on features has been replaced by sensed depth. IR sensors are used for depth perception through pattern projection in the Kinect v1 or time-of-flight in the Kinect v2. But depth channel has limited range (0.4-4.5 m) and performance of depth sensors deteriorates in varying lightening conditions [12].

With the advent of deep learning architectures based on Convolutional Neural Networks (CNN) and their outstanding performance in handling images for object recognition, detection, classification, localization and segmentation [14], a new class of vSLAM has emerged with the name of Learned SLAM. These strategies use CNNs for predicting depth on single image by learning monocular cues or they regress to learn the optical flow and odometry from sequence of frames to predict relative pose estimate of the platform. This has enabled the use of miniature monocular cameras on Size, Weight and Power (SWaP) constrained robots for dense map generation and helped to eliminate scale ambiguity of monocular camera-based SLAM (monoSLAM) strategies. Result comparisons like that shown in Fig. 1 help in clearly showing the advantages in scale ambiguity.

In this paper, a survey of deep learning based vSLAM strategies is presented and improvements over monoSLAM strategies are discussed. We present a summary of learned SLAM strategies presented in literature, encompassing their main modules for understanding of the readers and critically analyse them on resulting pose estimation accuracies, nature of map generation

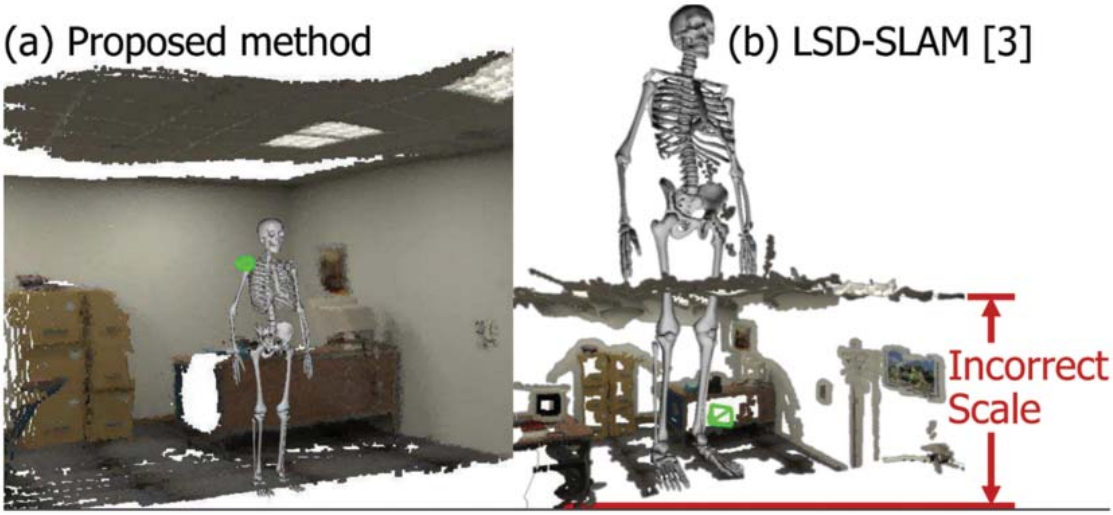

Figure-1: Mapping comparison and between Learned (CNN-SLAM)-left and Conventional Mono SLAM (LSD-SLAM)-right showing scale ambiguity

(Adapted from [13])

and sub modules added to them for purpose of effective robot navigation. Forthcoming sections in the paper are organized as follows. In Section 2, literature review is presented which highlights the advancements from conventional monocular to Learned vSLAM and works closely linked to the domain. Section 3 gives a brief overview of proposed learned vSLAM strategies however for detailed understanding we recommend consulting the referenced research papers. We have conducted an analysis of all these strategies with reference to monoSLAM strategies in Section 4. Finally, the survey is concluded in Section 5.

2. History of Visual SLAM

Miniature size and high information to weight ratio of monocular cameras has always fascinated researchers for use in Size, Weight and Power (SWaP) Constrained robots [15]. As robots are developed with the aim of being autonomous and independent of user/external aids to carry out their assigned tasks; localization and perception of surroundings is critical for their operation which SLAM readily provides for. Stereo camera equipped robots use stereo baseline for depth estimation or RGB-D sensor equipped robots make use of designated IR scanners for depth estimation. However, monocular camera equipped robots require complex algorithms to estimate third parameter of the landmark being tracked, that is depth.

In this regard, the first work was presented in 2003 with the name of MonoSLAM [16]. It uses Extended Kalman Filter (EKF) to track pose estimates of the camera and landmarks that are updated in the state vector. As the map is built over time the state vector gets larger and it becomes difficult to continue the process in real time. Also, it requires landmarks at known locations to initiate. To solve the complexity of MonoSLAM in larger environments, PTAM [17] presented the idea of tracking and mapping in different threads. Instead of initialization from known landmarks, it uses a five-point algorithm for initial mapping. ORB-SLAM [18] is one of the most complete feature-based tracking and mapping strategies which uses ORB features and employs bundle adjustment, closed loop detection and pose-graph optimization for accurate pose estimation, globally consistent mapping and drift cancellation as an all-in-one package.

Besides these feature-based tracking strategies which require hard coded feature descriptors in the environment and construct sparse maps, direct SLAM methods have also been proposed

which use photometric details in the image for tracking. This enabled the use of monoSLAM strategies in featureless environments and provided dense maps. DTAM [19] and LSDSLAM [20] are one of the most well-known strategies of Direct monoSLAM. DTAM tracks complete images by comparing them from reconstructed 3D map for pose estimation and estimates depth on all pixels. It produces dense maps, the photometric inconsistency amongst pixels that it requires for accurate mapping is hard to come by in many scenarios. Therefore, LSD-SLAM proposes to limit tracking on intensity gradients in the image and use photometric residuals among images to ascertain camera pose. It also employs pose graph optimization for consistent global maps. However, semi-dense maps are produced comprising of areas having intensity gradients. A survey covering detailed aspects of monoSLAM strategies presented from 2010 to 2016 has presented by Taketomi et al. [11].

Use of deep learning architectures for depth prediction by learning the monocular cues in presented images during training, has set a new path for monoSLAM. Predicting depth from a single image can be an ill posed problem, as the same image can be rendered from many scenes at multiple scales, but architectures can learn the scale of environments that humans live in and predict accurate enough depths. For that researchers have used deeper and deeper architectures of CNNs and crafted the loss functions to help the network focus on specific details of the training dataset. Several research works have been presented in the last decade which improve the accuracy of predicted depth and bring it closer to the sensed depth. The presented algorithms start from the use of a dual scaled network, predicting coarse outputs on the image enabling the network to learn global cues and passing to a finer scaled network for improving the details of predictions [21]. Another direction was presented by using CNN models of AlexNet [22] and VGG [23] with Conditional Random Fields (CRF), learning the depth at finer resolutions after sub-patches of the input image [24–26]. By using deeper CNNs like ResNet50 and using up-projection blocks to maintain the resolution of input image, Laina et al. reported depth prediction results of more than 80% $\delta 1$ (Number of predictions within 1.25 m of ground truth) accuracy [27]. Another notable work has been presented by Godard et al. in which instead of training the network on ground truth depth, stereo images are used to train the network. Network learns disparity by using a loss function comprising of matched appearance by left and right image, disparity smoothness and consistency in left-right image disparity. Disparity images are then used to find depth image with known camera focal length and baseline [28]. A recent survey covers the deep learning architectures designed for depth prediction from single images from 2014 to 2018 [29].

By infusing the power of deep learning architectures with monoSLAM for depth estimation, research has shown to resolve the issues of scale ambiguity, initialization to large uncertainty, failure because of pure rotation and sparse map generation. In this domain, foundation was laid by Laina et al. who along with presenting a novel Fully Connected Residual Network (FCRN) architecture showed its application to SLAM [27]. 3D reconstructed map using predicted depth was lacking shape details because of blurring effect on the borders of the object, however, it opened a new direction in vSLAM. Tateno et al. were the first to implement a complete learned SLAM strategy by infusing key-frame based depth estimation on intensity gradients with predicted depth maps [13]. To lower the trainable parameter bulk in deep learning architecture and improve on consistency of depth prediction on overlapping key-frames, Bloesch et al. propose using intensity image instead of RGB image with autoencoder-like network architecture and using photometric and geometric residuals to estimate the best pose for a key-frame [30]. Similarly, the Sparse2Dense adds surface normal along with predicted depth by using Fully Connected Dilated Residual Network (FCDRN) instead of FCRN in CNN-SLAM and report results with improved accuracy [31].

Instead of using CNNs only for depth prediction, another class of learned SLAM algorithm uses fully learned architectures to predict pose along with depth from a sequence of images. Number of research works have been proposed in this domain, few targeting learned pose estimation only, therefore are not included in this survey [32–35]. However, those which are generating depth maps from monocular images and using them for visual odometry are made part of this survey as these strategies can potentially be extended to full scale SLAM by mapping the depth map on RGB images and generating a 3D point cloud. These include CNN-SVO [36] and Scale-aware Monocular SLAM [37]. A similar work surveys the research presented in the domain up to 2017, however, valuable contribution has been made in following years which has been covered in this review.

Another closely related class that provides pose estimation and mapping is Structure from Motion (SfM). SLAM can be differentiated from SfM in that SLAM works on sequence of images for pose estimation and mapping while SfM deals with unordered images to enhance the reconstructed map, identifying the camera poses while doing so. Furthermore, SLAM is built computationally light therefore can run in real-time. Many techniques have been introduced to limit the optimization on subset of data instead of complete map like using limited key-frames, using computationally simpler feature descriptors, and tracking methodologies, de-linking tracking, and mapping threads.

On lines of SfM, many deep learning methodologies has also been built like SfM-Net [38], DeMoN [39] and Unsupervised learning of Depth and Ego-Motion [40]. These can easily be confused to SLAM, as end product of both strategies is the pose estimation and accurate depth map. However, they can be distinguished on above-mentioned points and are not kept as part of this survey to keep this survey focused on SLAM.

3. Learned vSLAM Strategies

3.1 CNN SLAM

CNN SLAM [13] proposes semantic reconstruction of scene using key-frames. Depth is predicted on key-frames using Convolutional Neural Network [27] that is infused with estimated depth. For depth estimation LSD-SLAM key-frame based tracking is used on intensity gradients [20]. Camera pose is estimated on each input frame by comparing it with the nearest key-frame and minimizing the gradient residual. CNN predicts depth only on the key-frames and predicted depth is sequentially refined using uncertainty maps. Uncertainty map is calculated by square of the difference between depth maps of reference key-frame and transformed current key-frame according to the estimated camera pose. Then the raw predicted depth is refined by weighted sum based on uncertainty map of current and reference keyframe. This enables refinement of predicted depth by giving more weightage to estimated depth in high gradient regions while plane regions in image will hold predicted depth. Global map is generated after pose graph optimization on every key frame as shown in Fig. 2. Proposed strategy not only refines the CNN predicted depth with key-frame based estimated depth but also improves upon prediction inconsistencies of same pixels from different camera poses. CNN SLAM has also proposed adjustment of predicted depth with ratio of focal lengths of the cameras (that have been used for prediction and CNN training). Authors report 20 % increase in accuracy of predicted depth after adjustment and refinement. Dense 3D reconstruction of the scene is carried out with refined depth map and monocular camera-based pose estimates. Authors also propose to use semantic labels predicted by the CNN for scene reconstruction [41]. Depth prediction CNN architecture used by the authors is trained on the NYU dataset and proposed strategy is evaluated on the ICL-NUIM [42] and TUM RGB-D SLAM datasets [43]. Targeted benchmarks for evaluation of results are absolute trajectory error and percentage of correctly estimated depth. CNN-SLAM reports low absolute trajectory error as compared to state-of-the-art monocular

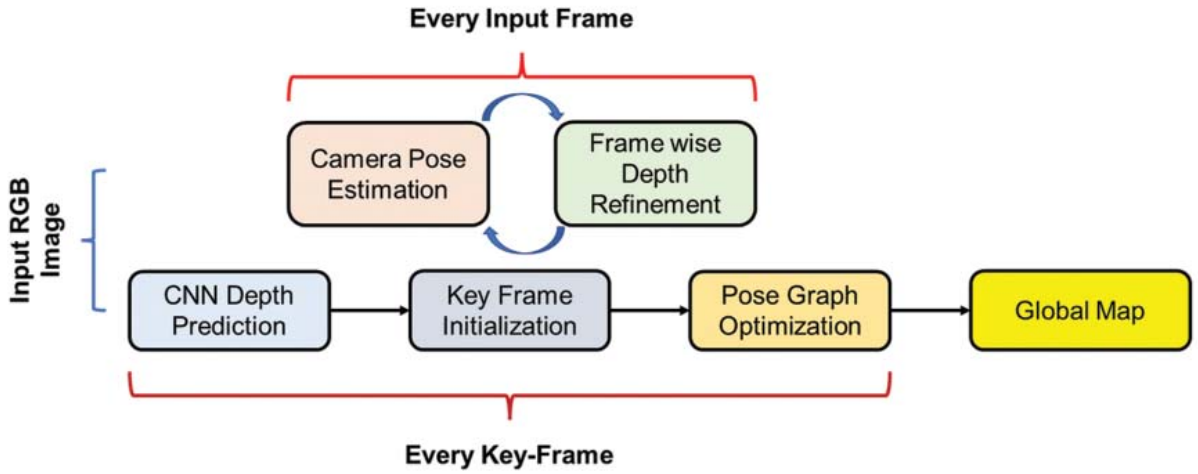

Figure-2: CNN-SLAM Flow Diagram [13]

camera-based SLAM strategies with highly dense 3D maps. Due to fusion with estimated depths and uncertainty map based refinement, blurring effect in raw depth prediction has also been reduced. Due to use of CNN predicted depth, SLAM does not fail under pure rotation of the camera and pose estimation on absolute metric scale is provided removing the scale ambiguity of monoSLAM strategies.

3.2 Scale-Aware Monocular SLAM

Scale-aware monocular SLAM [37] is composed of two modules (a) An end-to-end CNN based depth prediction module and (b) a Feature based monocular SLAM system. The depth prediction module receives consecutive monocular images as inputs and in return gives corresponding depth maps. Mono depth CNN architecture [28] is used for this purpose which is trained on stereo images to predict disparity that is used to calculate the depth map. For pose estimation ORB-SLAM is used which tracks ORB features in successive frames for initialization and then depth of particular point is calculated using triangulation. However, scale-aware monoSLAM uses predicted depth to determine 3D coordinates of the feature and therefore does not require initialization.

Authors have used the Cityscapes [44] and KITTI dataset [45] for Monodepth architecture training and evaluated the proposed strategy on the KITTI dataset. Main benchmark for result evaluation is translational and rotational drift over the distance covered. Improved results on part of pose estimation are reported as compared to monocular ORB-SLAM while results are comparable with stereo ORB-SLAM. Authors also report robustness of this strategy over monocular ORB-SLAM in pure rotational movement of the sensor. However, results of 3D mapping with estimated poses are not reported.

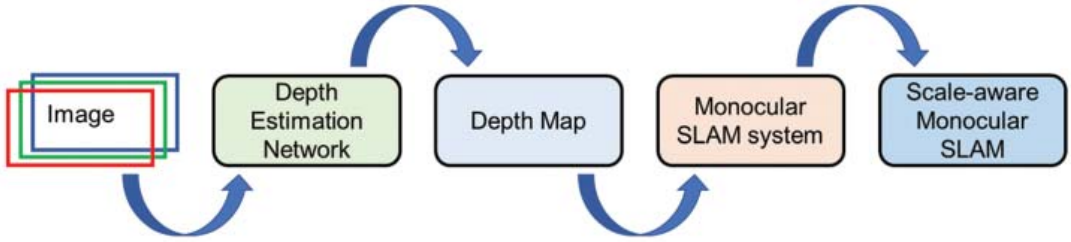

Figure-3: Scale-aware monoSLAM Flow Diagram [37]

3.3 CNN SVO

Semi direct Visual Odometry (SVO) [46] combines the strengths of direct and feature based methods. It offers an efficient probabilistic mapping method for direct camera motion estimation. One of the limitations is its initialization with large depth uncertainty which causes error in feature correspondence because of large search range along the epipolar line and a high number of measurements for converging to the true depth. CNN SVO [36] improves by initialization from depth prediction through CNN from single images. This reduces the uncertainty for identifying the corresponding features. A depth filter is used to separate good depth predictions of CNN from bad. Two consecutive image frames are used for image and feature alignment. It is followed by pose and structure refinement for tracking. Depth estimation is done on key frames whereas all the frames are used for updating depth filter. An overview of methodology is shown in Fig. 4.

CNN SVO also uses Monodepth encoder-decoder architecture [28] for depth prediction that is based on ResNet50. Pre-trained network on the Cityscapes dataset is used that is further fine tuned on the KITTI dataset [45]. To make it robust for high dynamic range (HDR) environments, random adjustments of brightness are done. Strategy is evaluated on the KITTI dataset benchmark of Absolute Trajectory Error (ATE) and compared with the results of SVO [46], DSO [47] and ORB-SLAM [48] strategies. Authors report robustness of strategy to produce pose estimate on metric scale and claims denser map generated as compared to SVO. However, mapping results are not reported on a benchmark.

3.4 DeepTAM

DeepTAM [49] is a fully learned method which uses CNNs to perform tracking and mapping. For the tracking part it follows an approach similar to that of DTAM (Dense Tracking And Mapping) [19]. CNNs are used for aligning of keyframes of both colour and depth image to current frame of colour and depth image. Refinement of the estimated camera pose is done by the network incrementally. To help in convergence of the predicted camera pose, a virtual keyframe is updated in every step. To track the camera pose, a transformation matrix is estimated that maps

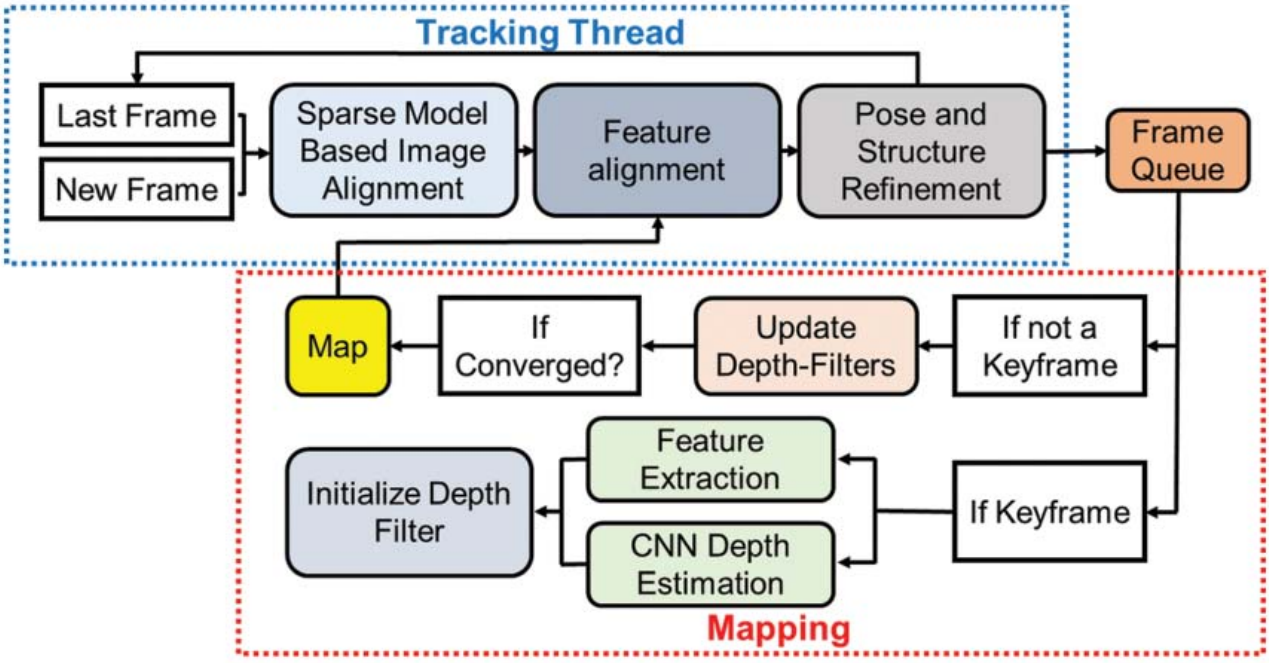

Figure-4: CNN SVO Flow Diagram [36]

a point to coordinate system of current camera frame from the coordinate system of key-frame. CNN is used to learn increment of transformation while finding 2D-3D correspondence between the key-frame and the current image. They have used an encoder decoder architecture for learning to estimate 6 DOF pose. The last part which is used for pose generation in terms of translation and rotation vectors, has 64 branches of fully connected layers sharing their weights. For tracking in real time, a coarse-to-fine strategy is used in which the pose estimation is dealt at three different tracking networks. This is done at three different resolutions and prediction is refined at these levels. For training of their neural network, they have used the SUNCG [50] and SUN3D [51] datasets.

For the mapping part the network is built upon the plane sweep stereo idea. The information from multiple images is combined together in a cost volume and then refined with all the collected depth measurements. For further improving the depth prediction, a network is appended, which, using a cost function defined on a narrow band around the previous surface estimate, iteratively refines the depth prediction. A coarse-to-fine strategy is applied along the depth axis. Mapping is subdivided into fixed band module and narrow band module. Depth labels and cost volume built by fixed band module are evenly spaced in the whole depth range, while narrow band module's cost volume is centered at current depth estimation and information is accumulated in a small band close to estimate. Training of mapping network is performed on the SUN3D [51] and SUNCG [50] datasets. Moreover, the authors also generated their own dataset called MVS.

3.5 Online Adapted Depth Prediction

Hongcheng et al. [52] proposed integration of LSD SLAM [20] with an online-updated CNN to progressively increase the completeness and accuracy of monocular mapping. To facilitate robust online adaptation of CNN model, several effective and computationally efficient mechanisms are proposed which can choose reliable and robust training data. To ensure satisfactory generalization and efficient adaptation of CNN on-the-fly, a stagewise SGD training method with selective update scheme has been used. Framework is composed of four major modules direct monocular SLAM, depth prediction through online adapted CNN, depth scale regression and fusion. Flow diagram interconnecting these modules is shown in Fig. 5.

The key-frame selected along with its camera poses is buffered into a training sample pool. Depth prediction is performed using weakly a supervised approach [53]. The first part is based on ResNet-50, excluding the fully connected layers. A convolution layer followed by FCN [54] with skip connections and three deconvolutional layers. Initially, a pretrained model is used to predict the pixel-wise depth of input images. When the depth prediction through CNN is not good enough and the training sample pool of key-frames is full, the buffered training samples are used to fine-tune the model. During the process of fine-tuning the depth prediction is still performed with the old model. Once the model is fine-tuned, it replaces the old model and starts performing depth prediction. The depth predicted and semi-dense map of key-frames are both used together to regress the absolute scale of the semi-dense map. Finally, fusion of dense and refined semi-dense map is performed.

Proposed CNN architecture was trained on the Wean Hall dataset [55] and evaluated on the ICL-NUIM [42] and TUM RGBD [43] datasets. To compare the performance of tracking accuracy, results have been compared with CNN-SLAM [13] and LSD-SLAM [20] on the metric of Absolute Trajectory Error (ATE). Moreover, CNN regressed scales of various ICL-NUIM and TUM-RGBD sequences are reported close to the ground truth scales. For generalization of camera parameters, focal length adjustment is carried out. To cater for the scale problem of Monocular scale, authors [52] have also used absolute scale regression method. However, proposed online training strategy could only be used for horizontal motion and the baseline within a batch needs to be fixed.

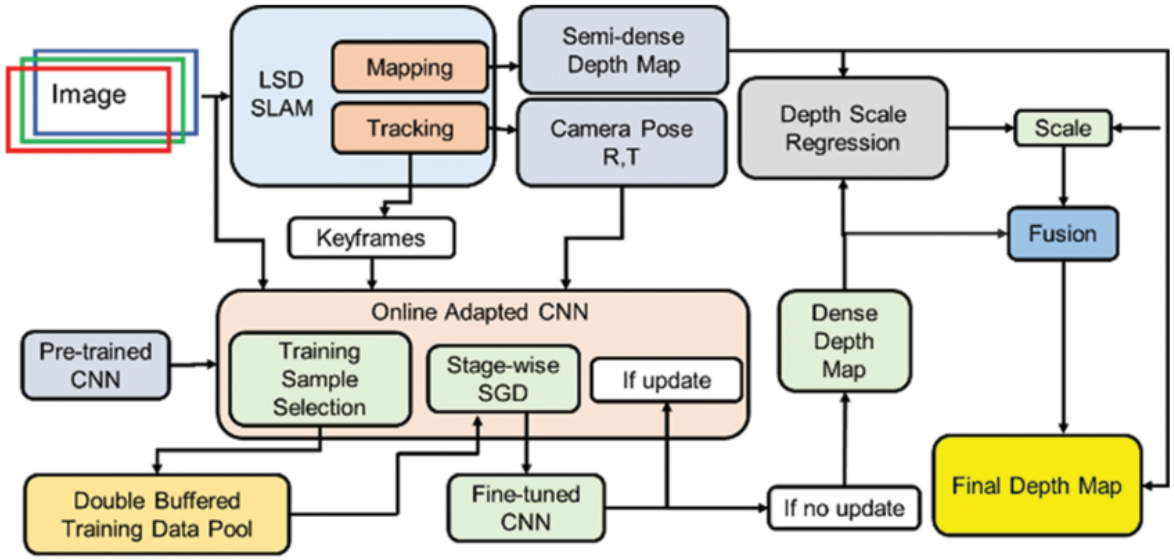

Figure-5: Online Adapted Depth Prediction Network Flow Diagram [52]

3.6 Sparse2Dense

The Sparse2Dense (S2D) [31] framework is divided into four stages (a) Learning based prior generation for depth/normal, (b) Visual tracking using direct alignment, (c) Geometrical sparse to dense reconstruction and (d) Fusion-based mapping. Flow diagram of framework is shown in Fig. 6. For depth and normal estimation, network is named as Fully Connected Deep Residual Network (FCDRN). In FCDRN the encoder part of FCRN [27] is replaced by Deep Residual Network (DRN) [56]. Rather than using up sampling residual block the network is trained to predict depth/normal at three different scales. The learned depth is used to achieve geometric optimization, reduction in scale drift and improvement in accuracy of monocular camera pose estimation. This gives optimized sparse depth estimates, which is converted into dense point cloud using surface normal based geometric reconstruction. In S2D DSO [57] is used as monocular VO (visual odometry).

Training of network is done on the SUN-3D [51] and SUN-RGBD [58] datasets. Evaluation is performed on same datasets as CNN-SLAM [13] which are TUM-RGBD and ICLNUIM datasets. S2D outperforms DSO [57], CNN-SLAM [13], LSD [20], ORB [48] and the method proposed by Laina et al. [27] on the metric of Absolute Trajectory Error (ATE).

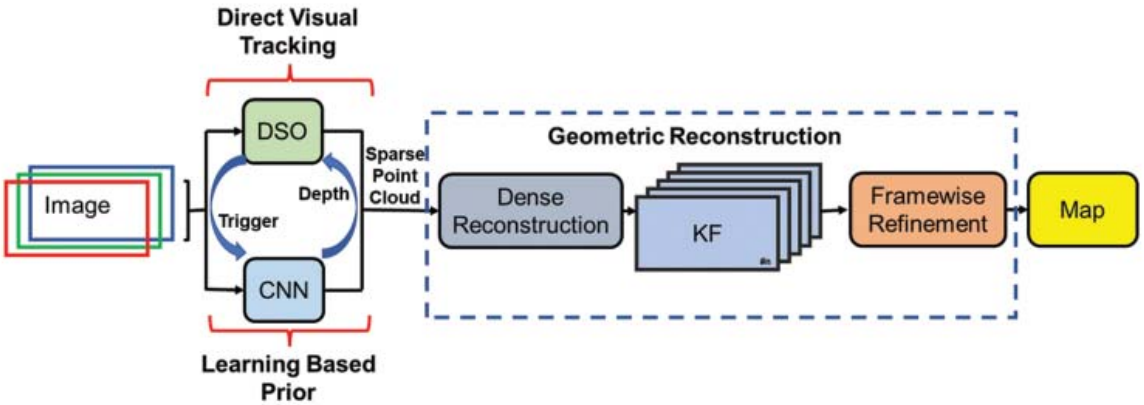

Figure-6: Sparse2Dense Flow Diagram [31]

4. Analysis

To analyze the progress made by learned vSLAM, we enumerate the limitations of conventional monoSLAM strategies and discuss how these have been tackled in surveyed methodologies. The Initialization, Scale Ambiguity, Pure Rotation and Sparse 3D Reconstruction are discussed. In each case, the disadvantages faced by conventional monoSLAM strategies are presented, followed by how Learned vSLAM tackles these issues.

4.1 Initialization

Map initialization is a well-known issue for monocular sensors as depth information is lost when a 2D image is captured. Therefore, it requires sequential images to estimate the depth on image disparity when adequate stereo baseline is achieved. Even then the map is initialized with large depth uncertainties and are converged after several key-frames are passed [20]. Alternates are initialization with known landmark states adding to the load of novice operators [16] or using epipolar geometry that requires parallax from non-planar scenes [18]. However, in learned vSLAM strategies maps are initialized with the use of CNN predicted depth on single image based on learned monocular cues.

4.2 Scale Ambiguity

From the fact that same image can be rendered by environments of infinite scales, ambiguity arises. MonoSLAM strategies are therefore initialized with features of known state [16], [59], epipolar geometry [17] or metrically scalable information from extra sensors [60] and use global Bundle Adjustment to limit scale drift [18]. However, CNNs not only learn to predict pixel wise depth but scale of the environment on which they are trained on. Discussed Learned SLAM and VO strategies initialize their pose estimation with predicted depth hence limiting the initial ambiguity of the environment scale. But this also means that such strategies will fail to initialize in environments on which they are not trained. To eliminate the issue, online depth prediction strategies propose to tune the network on-the-fly for regressing true scale of the environment [52].

4.3 Pure Rotation

During pure rotation movement of the sensor, image disparity between successive frames is lost which inhibits the depth estimation based on stereo baseline. In learned vSLAM, depth estimation in conventional monoSLAM is empowered with predicted depth fusion. Hence, depth information on frames, not providing image disparity, is still available enabling SLAM to continue pose estimation and mapping of the environment.

Table 1: CNN Parameters Summery

Learned SLAM	CNN	Parameter (million approx.)	Memory* (GB approx.)	Training dataset size (samples)
CNN SLAM	FCRN	62	0.25	96 k
Scale-aware monoSLAM	Monodepth	31	0.12	68.4 k
CNN SVO	Monodepth (Resnet-50)	48	0.19	68.4 k
DeepTAM	DeepTAM	24	0.1	-
Online Adapted Depth Pred	Resnet-50	12	0.05	42.09 k Pairs
Sparse2Dense	FCDRN	25	0.1	54.62k

*For, Batch Size: 1 and float32 variable

4.4 Sparse 3D Reconstruction

Feature based monoSLAM strategies track and map limited features like SIFT, SURF, ORB, etc and direct monoSLAM like LSD-SLAM maps pixels with intensity gradients. DTAM requires enough pixel wise difference to map the environment accurately which is not the case in most human made environments comprising of planar walls and objects [19]. Therefore, top of the line monoSLAM strategies lose clarity of 3D reconstructed maps or require multiple rolls on a scene to map it densely. Whereas, learned vSLAM uses pixel-wise predicted depths from CNNs that enables dense reconstruction of the environment.

With the use of CNNs in visual odometry and 3D reconstruction, monoSLAM performance has approached the performance of state-of-the-art stereo and RGB-D camera based vSLAM strategies (which are known for their accurate pose estimation capabilities and dense map reconstruction). However, few inherent limitations have also impeded the progress of Learned SLAM in which most glaring are lack of generalization capability, bulk of training data required to tune the millions of CNN parameters and memory utilization of the SLAM embodying multi layered deep learning architectures. CNN parameters affect training time and data required while memory utilization has been abated by low-cost Graphic Processing Units, still these parameters play an important role in steadfast applicability of the whole system in any generic environment. Therefore, an approximate calculation of the number of tuneable parameters and memory consumption (with assumed float 32 variable use) is carried out on the basis of information provided in relevant literature of surveyed methodologies and an overview of these important CNN parameters are provided in Table 1. Generalization capability of the learned vSLAM is fundamental to the operation of the system employing them. Therefore, surveyed strategies evaluate their vSLAM methodology on datasets they are not trained on. Datasets used for training and evaluation of learned vSLAM strategies along with benchmarks on which they are assessed on have been summarized in Table 2. Furthermore, we have tabulated the theme of the proposed strategies and listed their limitations. However, we recommend consulting each of the referenced works for a detailed understanding of their working.

Table 2: Learned vSLAM Strategies Summary

Learned vSLAM (Year)	Novel Concept	ConvNet	Training Dataset	Evaluation Dataset	Benchmark	Targeted Strategies	Limitation
CNN SLAM [13] (2017)	-Depth infusion of FCRN predicted and LSD-SLAM estimated -Depth adjustment from focal length ration -Depth refinement from uncertainty maps.	FCRN [27]	NYU	ICL-NUIM TUM-RGBD	ATE Correct depth (%)	LSD-SLAM ORB-SLAM REMODE	-Reconstruction of areas with low gradients -Recovery from inadequate initialization by FCRN predicted depth

Scale-aware MonoSLAM [37] (2018)	-Depth infusion of MonoDepth predicted and ORB-SLAM estimated	MonoDepth [28]	Citiscapes KITTI	KITTI	Translational / rotational drift	ORB-SLAM-S VISIO-S	-Intrinsic parameters of sensor used for training are embedded in system which limits generalization capability -Translational drift higher than targeted strategies -Trained for horizontal motion only.
CNN SVO [36] (2018)	-Initialize new map points with MonoDepth predicted depth.	MonoDepth [28]	Cityscapes KITTI	KITTI Robotcar	ATE	SVO DSO ORB-SLAM	-Dependency on pred depth to increase the corresponding features for mapping -MonoDepth pred on overexposed images -Depreciated feature matching due to illumination variation b/w key-frames -Trained for horizontal motion only
DeepTAM [49] (2018)	-Novel ConvNet architecture for pose estimation and depth prediction on multiple images	Proposed in same paper	SUN-CG SUN-3D MVS	TUM-RGBD	Trans RMSE L1-relative L1-inverse Scale-invariance	RGB-D SLAM CNN-SLAM	-Training time of 8 days for mapping network -Training time of 1 day for tracking network

Online Adapted Depth Pred [52] (2019)	-Online adapted ConvNet -Stage-wise SGD training method with a selective update scheme -Absolute scale regression on adapted ConvNet.	Proposed in same paper	Wean Hall	ICL-NUIM TUM	ATE Correct depth (%)	LSD-SLAM CNN-SLAM	-Collection of training data during fast translational and pure rotational motion -Gathered training data during operation need to be rectified for fixed stereo baseline -Trained for horizontal motion only.
Sparse2Dense [31] (2019)	-Learned depth priors for geometric optimization. -Surface normal along with depth for dense map reconstruction.	FCDRN [31]	SUN-3D SUN-RGBD	ICL-NUIM TUM-RGBD	ATE Correct depth (%)	LSD-SLAM ORB-SLAM REMODE CNN-SLAM	-Tracking at high frame rate and angular velocities -FCDRN depth prediction on textureless input images.

5. Conclusion

Learned vSLAM strategies (based on deep learning) have provided a new avenue to Simultaneous Localization and Mapping, that is one of the most researched areas in the field of robotics. By eliminating the inherent issues of conventional monoSLAM methodologies, they promise dense reconstruction of the surroundings and accurate situational awareness for robots equipped with simple monocular cameras. But deep learning methodologies employed in this field do require state-of-the-art hardware for real-time computations, which may force robots to use off-board computations. However, as hardware platforms become more powerful, and algorithms become more optimized, prospects for utilizing computationally expensive Learned vSLAM strategies on-board mini robots are bright. Future work includes optimization of Learned vSLAM implementation on mobile robots (with limited processing capacity due to SWaP constraints).

References

[1] A. Bachrach, R. He, and N. Roy, "Autonomous flight in unknown indoor environments," International Journal of Micro Air Vehicles, vol. 1, no. 4, pp. 217–228, 2009.

[2] M. Achtelik, A. Bachrach, R. He, S. Prentice, and N. Roy, "Stereo vision and laser odometry for autonomous helicopters in gps-denied indoor environments," in Unmanned Systems Technology XI, vol. 7332. International Society for Optics and Photonics, Conference Proceedings, p. 733219.

[3] A. Bachrach, S. Prentice, R. He, and N. Roy, "Range–robust autonomous navigation in gps-denied environments," Journal of Field Robotics, vol. 28, no. 5, pp. 644–666, 2011.

[4] F. Abrate, B. Bona, and M. Indri, "Experimental ekf-based slam for minirovers with ir sensors only," in EMCR, Conference Proceedings.

[5] M. Bl¨osch, S. Weiss, D. Scaramuzza, and R. Siegwart, "Vision based mav navigation in unknown and unstructured environments," in Robotics and automation (ICRA), 2010 IEEE international conference on. IEEE, Conference Proceedings, pp. 21–28.

[6] K. Celik, S.-J. Chung, and A. Somani, "Mono-vision corner slam for indoor navigation," in Electro/Information Technology, 2008. EIT 2008. IEEE International Conference on. IEEE, Conference Proceedings, pp. 343–348.

[7] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, Visual odometry and mapping for autonomous flight using an RGB-D camera. Springer, 2017, pp. 235–252.

[8] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluationofthergb-dslamsystem,"in Icra,vol.3,ConferenceProceedings, pp. 1691–1696.

[9] J. Engel, J. St¨uckler, and D. Cremers, "Large-scale direct slam with stereo cameras," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, Conference Proceedings, pp. 1935–1942.

[10] R. Mur-Artal and J. D. Tard´os, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255–1262, 2017.

[11] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: A survey from 2010 to 2016," IPSJ Transactions on Computer Vision and Applications, vol. 9, no. 1, p. 16, 2017.

[12] S. M. Abbas and A. Muhammad, "Outdoor rgb-d slam performance in slow mine detection," in ROBOTIK 2012; 7th German Conference on Robotics. VDE, Conference Proceedings, pp. 1–6.

[13] K. Tateno, F. Tombari, I. Laina, and N. Navab, "CNN-SLAM Real-time dense monocular slam with learned depth prediction," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6243– 6252.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," International journal of computer vision, vol. 115, no. 3, pp. 211–252, 2015.

[15] E. Royer, M. Lhuillier, M. Dhome, and J.-M. Lavest, "Monocular vision for mobile robot localization and autonomous navigation," International Journal of Computer Vision, vol. 74, no. 3, pp. 237–260, 2007.

[16] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in Iccv, vol. 3, Conference Proceedings, pp. 1403–1410.

[17] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. IEEE Computer Society, Conference Proceedings, pp. 1–10.

[18] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," IEEE transactions on robotics, vol. 31, no. 5, pp. 1147–1163, 2015.

[19] R.A.Newcombe, S.J.Lovegrove, and A.J.Davison,"Dtam:Densetracking and mapping in real-time," in 2011 international conference on computer vision. IEEE, Conference Proceedings, pp. 2320–2327.

[20] J. Engel, T. Schops, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in European Conference on Computer Vision. Springer, 2014, pp. 834–849.

[21] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in Advances in neural information processing systems, Conference Proceedings, pp. 2366–2374.

[22] I. Sutskever, G. E. Hinton, and A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, pp. 1097–1105, 2012.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[24] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in Proceedings of the IEEE conference on computer vision and pattern recognition, Conference Proceedings, pp. 1119–1127.

[25] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in Proceedings of the IEEE conference on computer vision and pattern recognition, Conference Proceedings, pp. 5162–5170.

[26] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Conference Proceedings, pp. 2800–2809.

[27] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 239–248.

[28] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Conference Proceedings, pp. 270–279.

[29] H. Laga, "A survey on deep learning architectures for image-based depth reconstruction," arXiv preprint arXiv:1906.06113, 2019.

[30] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, "Codeslam—learning a compact, optimisable representation for dense visual slam," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Conference Proceedings, pp. 2560–2568.

[31] J. Tang, J. Folkesson, and P. Jensfelt, "Sparse2dense: From direct sparse odometry to dense 3-d reconstruction," IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 530–537, 2019.

[32] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, Conference Proceedings, pp. 6939–6946.

[33] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, Conference Proceedings, pp. 2043–2050.

[34] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Conference Proceedings, pp. 5974–5983.

[35] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in Proceedings of the IEEE International Conference on Computer Vision, Conference Proceedings, pp. 37–45.

[36] S. Y. Loo, A. J. Amiri, S. Mashohor, S. H. Tang, and H. Zhang, "Cnn-svo: Improving the mapping in semi-direct visual odometry using single-image depth prediction," arXiv preprint arXiv:1810.01011, 2018.

[37] Y. Li, C. Xie, H. Lu, X. Chen, J. Xiao, and H. Zhang, "Scale-aware monocular slam based on convolutional neural network," 08 2018.

[38] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of structure and motion from video," arXiv preprint arXiv:1704.07804, 2017.

[39] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Conference Proceedings, pp. 5038–5047.

[40] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Conference Proceedings, pp. 1851–1858.

[41] F. T. K. Tateno and N. Nawab, "Real-time and scalable incremental segmentation on dense slam," 2015.

[42] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in 2014 IEEE international conference on Robotics and automation (ICRA). IEEE, 2014, pp. 1524–1531.

[43] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012, pp. 573–580.

[44] S. R. T. R. M. E. R. B. U. F. S. R. M. Cordts, M. Omran and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.

[45] P. L. A. Geiger and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3354–3361.

[46] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in 2014 IEEE international conference on robotics and automation (ICRA). IEEE, 2014, pp. 15–22.

[47] R. Wang, M. Schworer, and D. Cremers, "Stereo dso: Large-scale direct sparse visual odometry with stereo cameras," in Proceedings of the IEEE International Conference on Computer Vision, Conference Proceedings, pp. 3903–3911.

[48] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," IEEE transactions on robotics, vol. 31, no. 5, pp. 1147–1163, 2015.

[49] H. Zhou, B. Ummenhofer, and T. Brox, "Deeptam: Deep tracking and mapping," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 822–838.

[50] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1746–1754.

[51] J. Xiao, A. Owens, and A. Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1625–1632.

[52] H. Luo, Y. Gao, Y. Wu, C. Liao, X. Yang, and K.-T. Cheng, "Real-time dense monocular slam with online adapted depth prediction network," IEEE Transactions on Multimedia, vol. 21, no. 2, pp. 470–483, 2018.

[53] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in European Conference on Computer Vision. Springer, 2016, pp. 740–756.

[54] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[55] H. Alismail, B. Browning, and M. B. Dias, "Evaluating pose estimation methods for stereo visual odometry on robots," 2010.

[56] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 472–480.

[57] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 3, pp. 611–625, 2017.

[58] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 567–576.

[59] S. B. Knorr and D. Kurz, "Leveraging the user's face for absolute scale estimation in handheld monocular slam," in 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, Conference Proceedings, pp. 11–17.

[60] J. Mustaniemi, J. Kannala, S. Sarkka, J. Matas, and J. Heikkila, "Inertial based scale estimation for structure from motion on mobile devices," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, Conference Proceedings, pp. 4394–4401.

[61] R. Li, S. Wang, and D. Gu, "Ongoing Evolution of Visual SLAM from Geometry to Deep Learning: Challenges and Opportunities," Cognitive Computation, vol. 10, no. 6, pp. 875–889, 2018.