

Recurrent Neural Network for (Un-)supervised Learning of Monocular Video Visual Odometry and Depth

Rui Wang, Stephen M. Pizer, Jan-Michael Frahm
University of North Carolina at Chapel Hill

Abstract

Deep learning-based, single-view depth estimation methods have recently shown highly promising results. However, such methods ignore one of the most important features for determining depth in the human vision system, which is motion. We propose a learning-based, multi-view dense depth map and odometry estimation method that uses Recurrent Neural Networks (RNN) and trains utilizing multi-view image reprojection and forward-backward flow-consistency losses. Our model can be trained in a supervised or even unsupervised mode. It is designed for depth and visual odometry estimation from video where the input frames are temporally correlated. However, it also generalizes to single-view depth estimation. Our method produces superior results to the state-of-the-art approaches for single-view and multi-view learning-based depth estimation on the KITTI driving dataset.

1. Introduction

The tasks of depth and odometry (also called ego-motion) estimation are longstanding tasks in computer vision providing valuable information for a wide variety of tasks, e.g. autonomous driving, AR/VR applications, and virtual tourism.

Recently, convolutional neural networks (CNN) [20, 4, 8, 41, 32] have begun to produce results of comparable quality to traditional geometric computer vision methods for depth estimation in measurable areas and achieve significantly more complete results for ambiguous areas through the learned priors. However, most methods treat depth estimation as a single view task and thus ignore the important temporal information in monocular or stereo videos. The underlying rationale of these single view depth estimation methods is the possibility of human depth perception from a single image. However, they neglect the fact that motion is actually more important for the human to infer distance [28]. We are constantly exposed to moving scenes, and the speed of things moving in the image is related to the combination of their relative speed and effect inversely propor-

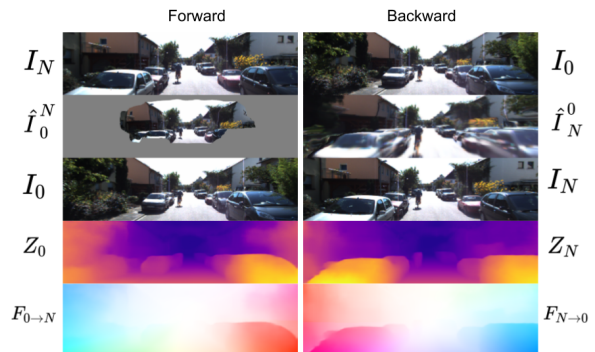


Figure 1: Example results from our method. The first row shows the source image. The second row illustrates the projection of the source image into the target image. The third row shows the target image. The fourth row illustrates the estimated depth map and the last row illustrates the estimated optical flow.

tional to their depth.

In this work, we propose a framework that simultaneously estimates the visual odometry and depth maps from a video sequence taken by a monocular camera. To be more specific, we use convolutional Long Short-Term Memory (ConvLSTM) [34] units to carry temporal information from previous views into the current frame’s depth and visual odometry estimation. We have improved upon existing deep single- and two-view stereo depth estimation methods by interleaving ConvLSTM units with the convolutional layers to effectively utilize multiple previous frames in each estimated depth maps. Since we utilize multiple views, the image reprojection constraint between multiple views can be incorporated into the loss, which shows significant improvements for both supervised and unsupervised depth and camera pose estimation.

In addition to the image reprojection constraint, we further utilize a forward-backward flow-consistency constraint [37]. Such a constraint provides additional supervision to image areas where the image reprojection is ambiguous. Moreover, it improves the robustness and generalizability of the model. Together these two constraints can even allow satisfactory models to be produced when groundtruth

is unavailable at training time. Figure 1 shows an example of forward-backward image reprojection and optical flow as well as the resulting predicted depth maps.

We summarize our innovations as follows: 1) An RNN architecture for monocular depth and odometry estimation that uses multiple consecutive views. It does so by incorporating LSTM units, as used in natural language processing, into depth and visual odometry estimation networks. 2) These LSTM units importantly allow the innovation of using depth and camera motion estimation to benefit from the richer constraints of a multi-view process. In particular, they use multi-view image reprojection and forward-backward flow-consistency constraints to produce a more accurate and consistent model. 3) This design allows two novel capabilities: a) it can be trained in both supervised and unsupervised fashion; b) it can continuously run on arbitrary length sequences delivering a consistent scene scale.

We demonstrate on the KITTI [10] benchmark dataset that our method can produce superior results over the state-of-the-art for both supervised and unsupervised training. We will release source code upon acceptance.

2. Related work

Traditionally, the 3D reconstruction and localization are mostly solved by pure geometric reasoning. SfM and SLAM are the two most prevalent frameworks for sparse 3D reconstruction of rigid geometry from images. SfM is typically used for offline 3D reconstruction from unordered image collections, while visual SLAM aims for a real-time solution using a single camera [3, 26]. More recent works on SLAM systems include ORB-SLAM [25] and DSO [5]. Schönberger and Frahm [30] review the state-of-the-art in SfM and propose an improved incremental SfM method.

Recently, CNNs are increasingly applied to 3D reconstruction, in particular, to the problem of 3D reconstruction of dense monocular depth, which is similar to the segmentation problem and thus the structure of the CNNs can be easily adapted to the task of depth estimation [21].

Supervised methods. Eigen *et al.* [4] and Liu *et al.* [20] proposed end-to-end networks for single-view depth estimation, which opened the gate for deep learning-based supervised single-view depth estimation. Following their work, Laina *et al.* [18] proposed a deeper residual network for the same task. Qi *et al.* [27] jointly predicted depth and surface normal maps from a single image. Fu *et al.* [6] further improved the network accuracy and convergence rate by learning it as an ordinal regression problem. Li *et al.* [19] used modern structure-from-motion and multi-view stereo (MVS) methods together with multi-view Internet photo collections to create the large-scale MegaDepth dataset providing improved depth estimation accuracy via bigger training dataset size. We improve upon these single-view methods by utilizing multiple views through an RNN

architecture to generate more accurate depth and pose.

Two-view or multi-view stereo methods have traditionally been the most common techniques for dense depth estimation. For the interested reader, Scharstein and Szeliski [29] give a comprehensive review on two-view stereo methods. Recently, Ummenhofer *et al.* [32] formulated two-view stereo as a learning problem. They showed that by explicitly incorporating dense correspondences estimated from optical flow into the two-view depth estimation, they can force the network to utilize stereo information on top of the single view priors. There is currently a very limited body of CNN-based multi-view reconstruction methods. Choy *et al.* [2] use an RNN to reconstruct the object in the form of a 3D occupancy grid from multiple viewpoints. Yao *et al.* [36] proposed an end-to-end deep learning framework for depth estimation from multiple views. They use differentiable homography warping to build a 3D cost volume from one reference image and several source images. Kumar *et al.* [16] proposed an RNN architecture that can learn depth prediction from monocular videos. However, their simple training pipeline, e.g., no explicit temporal constraints, failed to explore the full capability of the network. Our method is trained with more sophisticated multi-view reprojection losses and can perform both single-view and multi-view depth estimation.

Unsupervised methods. Recently, by incorporating elements of view synthesis [42] and Spatial Transformer Networks [14], monocular depth estimation has been trained in an unsupervised fashion. This was done by transforming the depth estimation problem into an image reconstruction problem where the depth is the intermediate product that integrates into the image reconstruction loss. Godard *et al.* [11], and Garg *et al.* [8] use stereo pairs to train CNNs to estimate disparity maps from single views. Luo *et al.* [22] leverage both stereo and temporal constraints to generate improved depth at known scale. Zhou *et al.* [41] further relax the needs of stereo images to monocular video by combining a single view depth estimation network with a multi-view odometry estimation network. Following Zhou *et al.* [41]’s work, Mahjourian *et al.* [23] further enforced consistency of the estimated 3D point clouds and ego-motion across consecutive frames. In addition to depth and ego-motion, Yin *et al.* [37] also jointly learn optical flow in an end-to-end manner which imposed additional geometric constraints. However, due to scale ambiguity and the lack of temporal constraints, these methods cannot be directly applied for full trajectory estimation on monocular videos. By leveraging recurrent units, our method can run on arbitrary length sequences delivering a consistent scene scale.

3. Method

In this section we introduce our method for multi-view depth and visual odometry estimation. We first describe our

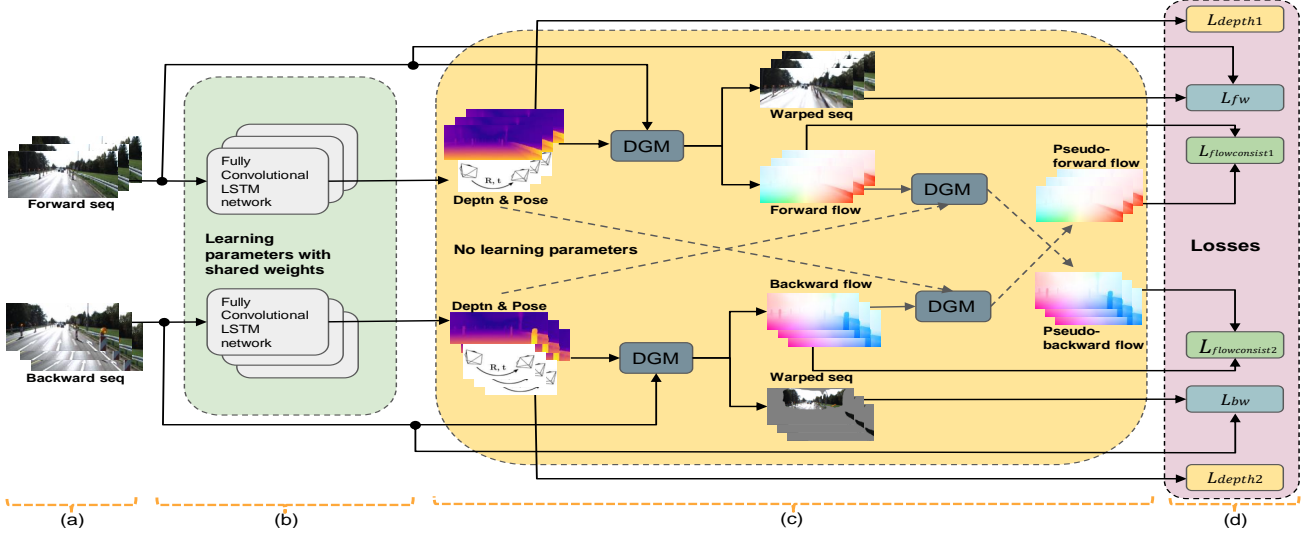


Figure 2: Training pipeline of our proposed RNN-based depth and visual odometry estimation network. During training our framework takes forward and backward 10-frame subsequences as input and uses multi-view image reprojection, flow-consistency, and optionally groundtruth depth to train our depth and visual odometry networks. DGM is a differentiable geometric module.

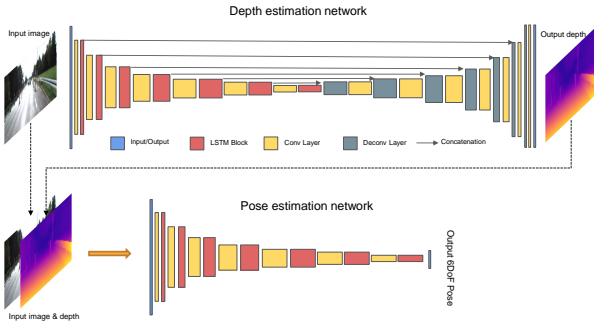


Figure 3: Overall network architecture of our RNN-based depth and visual odometry estimation framework. The height of each rectangle represents the size of its feature maps, where each smaller feature map is half the size of the preceding feature map.

recurrent neural network architecture and then the multi-view reprojection and forward-backward flow-consistency constraints for the network training.

3.1. Network Architecture

Our architecture, shown in Figure 3, is made up of two networks, one for depth and one for visual odometry.

Our depth estimation network uses a U-shaped network architecture similar to DispNet [24]. Our main innovation is to interleave recurrent units into the encoder which allows the network to leverage not only spatial but also temporal information in the depth estimation. The spatial-temporal features computed by the encoder are then fed into the decoder for accurate depth map reconstruction. The ablation study in Secion 4.5 confirms our choice for the place-

ments of the ConvLSTM [34] units. Table 1 shows the detailed network architecture. The input to the depth estimation network is a single RGB frame I_t and the hidden states h_{t-1}^d from the previous time-step (h_{t-1}^d are initialized to be all zero for the first time-step). The hidden states are transmitted internally through the ConvLSTM units. The output of our depth estimation network are the depth map Z_t and the hidden states h_t^d for the current time-step .

Type	Filters	Output size
Input		$128 \times 416 \times 3$
Conv+ConvLSTM	$32 @ 3 \times 3 \times 3$	$64 \times 208 \times 32$
Conv+ConvLSTM	$64 @ 3 \times 3 \times 32$	$32 \times 104 \times 64$
Conv+ConvLSTM	$128 @ 3 \times 3 \times 64$	$16 \times 52 \times 128$
Conv+ConvLSTM	$256 @ 3 \times 3 \times 128$	$8 \times 26 \times 256$
Conv+ConvLSTM	$256 @ 3 \times 3 \times 256$	$4 \times 13 \times 256$
Conv+ConvLSTM	$256 @ 3 \times 3 \times 256$	$2 \times 7 \times 256$
Conv+ConvLSTM	$512 @ 3 \times 3 \times 256$	$1 \times 4 \times 512$
Deconv+Concat+Conv	$256 @ 3 \times 3 \times 512$	$2 \times 7 \times 256$
Deconv+Concat+Conv	$128 @ 3 \times 3 \times 256$	$4 \times 13 \times 128$
Deconv+Concat+Conv	$128 @ 3 \times 3 \times 128$	$8 \times 26 \times 128$
Deconv+Concat+Conv	$128 @ 3 \times 3 \times 128$	$16 \times 52 \times 128$
Deconv+Concat+Conv	$64 @ 3 \times 3 \times 128$	$32 \times 104 \times 64$
Deconv+Concat+Conv	$32 @ 3 \times 3 \times 64$	$64 \times 208 \times 32$
Deconv	$16 @ 3 \times 3 \times 32$	$128 \times 416 \times 16$
Conv (output)	$1 @ 3 \times 3 \times 16$	$128 \times 416 \times 1$

Table 1: Detailed depth estimation network architecture. Every convolution in the encoder uses stride 2 for downsampling. Before the output a sigmoid activation function is used to ensure the output is in range $[0, 1]$; All the other convolutions and deconvolutions are followed by batch norm and LeakyRELU activation.

Our visual odometry network uses a VGG16 [31] architecture with recurrent units interleaved. Table 2 shows

Type	Filters	Output size
Input		128×416×4
Conv+ConvLSTM	32@3×3×3	64×208×32
Conv+ConvLSTM	64@3×3×32	32×104×64
Conv+ConvLSTM	128@3×3×64	16×52×128
Conv+ConvLSTM	256@3×3×128	8×26×256
Conv+ConvLSTM	256@3×3×256	4×13×256
Conv+ConvLSTM	256@3×3×256	2×7×256
Conv+ConvLSTM	512@3×3×256	1×4×512
Conv (output)	6@1×1×512	1×1×6

Table 2: Detailed visual odometry network architecture. Every convolution (except for output layer) is followed by batch normalization and RELU as activation.

the detailed network architecture. The input to our visual odometry network is the concatenation of I_t and Z_t together with the hidden states h_{t-1}^p from the previous time-step. The output is the relative 6DoF camera pose $P_{t \rightarrow t-1}$ between the current view and the immediately preceding view. The main differences between our visual odometry network and most current deep learning-based visual odometry methods are 1) At each time-step, instead of a stack of frames, our visual odometry network only takes the current image as input; the knowledge about previous frames is in the hidden layers. 2) Our visual odometry network also takes the current depth estimation as input, which ensures a consistent scene scale between depth and camera pose (important for unsupervised depth estimation, where the scale is ambiguous). 3) Our visual odometry network can run on a full video sequence while maintaining a single scene scale.

3.2. Loss Functions

3.2.1 Multi-view Reprojection Loss

Zhou *et al.* [41] showed that the learning of depth and visual odometry estimation can be formulated as an image reconstruction problem using a differentiable geometric module (DGM). Thus we can use the DGM to formulate an image reconstruction constraint between I_t and I_{t-1} using the estimated depth Z_t and camera pose $P_{t \rightarrow t-1}$ as introduced in the previous subsection. However, such a pairwise photometric consistency constraint is very noisy due to illumination variation, low texture, occlusion, etc. Recently, Iyer *et al.* [13] proposed a composite transformation constraint for self-supervised visual odometry learning. By combining the pairwise image reconstruction constraint with the composite transformation constraint, we propose a multi-view image reprojection constraint that is robust to noise and provides strong self-supervision for our multi-view depth and visual odometry learning. As shown in Figure 2(c), the output depth maps and relative camera poses together with the input sequence are fed into a differentiable geometric module (DGM) that performs differentiable image warping of every previous view of the sub-sequence into the current

view. Denote the input image sequence (shown in Figure 2(a)) as $\{I_t | t = 0 \dots N-1\}$, the estimated depth maps as $\{Z_t | t = 0 \dots N-1\}$, and the camera poses as the transformation matrices from frame t to $t-1$: $\{P_{t \rightarrow t-1} | t = 0 \dots N-1\}$. The multi-view reprojection loss is

$$L_{fw} = \sum_{t=0}^{N-1} \sum_{i=0}^{t-1} \sum_{\Omega} \lambda_t^i \omega_t^i |I_t - \hat{I}_t^i| \quad (1)$$

where \hat{I}_t^i is the i^{th} view warped into t^{th} view, Ω is the image domain, ω_t^i is a binary mask indicating whether a pixel of I_t has a counterpart in I_i , and λ_t^i is a weighting term that decays exponentially based on $t-i$. Image pairs that are far away naturally suffer from larger reprojection error due to interpolation and moving foreground so we use λ_t^i to reduce the effect of such artifacts. ω_t^i and \hat{I}_t^i are obtained as

$$\omega_t^i, \hat{I}_t^i, F_{t \rightarrow i} = \phi(I_i, Z_t, P_{t \rightarrow i}, K) \quad (2)$$

where $F_{t \rightarrow i}$ is a dense flow field for 2D pixels from view t to view i , which is used to compute flow consistency. K is the camera intrinsic matrix. The pose change from view t to i , $P_{t \rightarrow i}$ can be obtained by a composite transformation as

$$P_{t \rightarrow i} = P_{i+1 \rightarrow i} \cdot \dots \cdot P_{t-1 \rightarrow t-2} \cdot P_{t \rightarrow t-1} \quad (3)$$

The function ϕ in Equation 2 warps image I_i into I_t using Z_t and $P_{t \rightarrow i}$. The function ϕ is a DGM [38], which performs a series of differentiable 2D-to-3D, 3D-to-2D projections, and bi-linear interpolation operations [14].

In the same way, we reverse the input image sequence and perform another pass of depth $\{Z_t | t = N-1 \dots 0\}$ and camera pose $\{P_{t \rightarrow t+1} | t = N-1 \dots 0\}$ estimation, obtaining the backward multi-view reprojection loss L_{bw} . This multi-view reprojection loss can fully exploit the temporal information in our ConvLSTM units from multiple previous views by explicitly putting constraints between the current view and every previous view.

A trivial solution to Equation 1 is ω_t^i to be all zeros. To prevent the network from converging to the trivial solution, we add a regularization loss L_{reg} to ω_t^i , which gives a constant penalty to locations where ω_t^i is zero.

3.2.2 Forward-backward Flow Consistency Loss

A forward-backward consistency check has become a popular strategy in many learning-based tasks, such as optical flow [12], registration [40], and depth estimation [37, 11, 33], which provides additional self-supervision and regularization. Similar to [37, 33] we use the dense flow field as a hybrid forward-backward consistency constraint for both the estimated depth and pose. We first introduce a forward-backward consistency constraint on a single pair of frames and then generalize to a sequence. Let us denote a pair of

consecutive frames as I_A and I_B , and their estimated depth maps and relative poses as $Z_A, Z_B, P_{A \rightarrow B}$, and $P_{B \rightarrow A}$. We can obtain a dense flow field $F_{A \rightarrow B}$ from frame I_A to I_B using Equation 2. Similarly we can obtain $F_{B \rightarrow A}$ using $Z_B, P_{B \rightarrow A}$. Using $F_{B \rightarrow A}$ we can compute a pseudo-inverse flow $\hat{F}_{A \rightarrow B}$ (due to occlusion and interpolation) as

$$\omega_A^B \cdot \hat{F}_{A \rightarrow B}, F_{A \rightarrow B} = \phi(-F_{B \rightarrow A}, Z_A, P_{A \rightarrow B}, K) \quad (4)$$

This is similar to Equation 2 except that we are interpolating $F_{A \rightarrow B}$ from $-F_{B \rightarrow A}$ instead of I_t from I_i . Therefore, we can formulate the flow consistency loss as

$$L_{flowconsist} = \omega_A^B \cdot |F_{A \rightarrow B} - \hat{F}_{A \rightarrow B}| + \omega_B^A \cdot |F_{B \rightarrow A} - \hat{F}_{B \rightarrow A}| \quad (5)$$

This is performed for every consecutive pair of frames in the input sequence. Unlike the multi-view reprojection loss we only compute flow-consistency on pairs of consecutive frames given the fact that the magnitude of the flow increases, for frame pairs that are far apart, leading to inaccurate pseudo-inverses due to interpolation.

3.2.3 Smoothness Loss

Local smoothness is a common assumption for depth estimation. Following Zhan *et al.* [38], we use an edge-aware smoothness constraint which is defined as

$$L_{smooth} = \sum_{t=0}^{N-1} \sum_{\Omega} |\nabla Z_t| \cdot e^{-|\nabla I_t|} \quad (6)$$

3.2.4 Absolute depth loss

The combination of multi-view reprojection loss L_{fw} , L_{bw} defined in Equation 1, forward-backward flow-consistency loss $L_{flowconsist}$ defined in Equation 5, and smoothness loss L_{smooth} defined in Equation 6 can form an unsupervised training strategy for the network. This manner of training is suitable for cases where there is no groundtruth depth available, which is true for the majority of real world scenarios. However, the network trained in this way only produces depth at a relative scale. So optionally, if there is groundtruth depth available, even sparsely, we can train a network to estimate depth at absolute scale by adding the absolute depth loss defined as

$$L_{depth} = \sum_{t=0}^{N-1} \sum_{\Omega} |\xi_t - \hat{\xi}_t| \quad (7)$$

where ξ_t is the inverse depth. In addition, we can replace the local smoothness loss in Equation 6 by a gradient similarity to the groundtruth depth, which can be defined as

$$L_{smooth} = \sum_{t=0}^{N-1} \sum_{\Omega} |\nabla \xi_t - \nabla \hat{\xi}_t| \quad (8)$$

3.3. Training Pipeline

The full training pipeline of our method is shown in Figure 2. Every N consecutive key frames (we use $N = 10$ in all our experiments) are grouped together as an input sequence S_{fw} . The frames are grouped in a sliding window fashion such that more training data can be generated. Here the key frame selection is based on the motion between successive frames. Because the image reprojection constraints are ambiguous for very small baselines, we discard frames with baseline motion smaller than σ . Before passing the sequence to the network for training, we also reverse the sequence to create a backward sequence S_{bw} , which not only serves as a data augmentation but also is used to enforce the forward-backward constraints. The input sequence S_{fw} is generated offline during the data preparation stage while the backward sequence S_{bw} is generated online during the data preprocessing stage. S_{fw} and S_{bw} are fed into two networks with shared weights; each generates a sequence of depth maps and camera poses as shown in Figure 2. The estimated depth maps and camera poses are then utilized to generate dense flows to warp previous views to the current view through a differentiable geometric module (DGM) [42, 37]. Furthermore, we utilize DGMs to generate the pseudo-inverse flows for both the forward and backward flows. By combining image warping loss, flow-consistency loss, and optionally absolute depth loss, we form the full training pipeline for our proposed framework.

Once trained, our framework can run on arbitrary length sequences without grouping frames into fixed length sub-sequences. To bootstrap the depth and pose estimation, the hidden states for the ConvLSTM units are initialized by zero for the first frame. All following estimations will then depend on the hidden states from the previous time-step.

4. Experiments

In this section we show a series of experiments using the KITTI driving dataset [9, 10] to evaluate the performance of our RNN-based depth and visual odometry estimation method. As mentioned in Section 1, our architecture can be trained in a supervised or unsupervised mode. Therefore, we evaluated both supervised and unsupervised versions of our framework. In the following experiments we named the supervised version as *ours-sup* and the unsupervised version as *ours-unsup*. We also performed detailed ablation studies to show the impact of the different constraints, architecture choices, and estimations at different time-steps.

4.1. Implementation Details

We set the weights for depth loss, smoothness loss, forward-backward consistency loss, and mask regularization to 1.0, 1.0, 0.05, and 0.05, respectively. The weight for the image reprojection loss is $\frac{1}{2\sigma^2 - 1}$, where δ is the num-

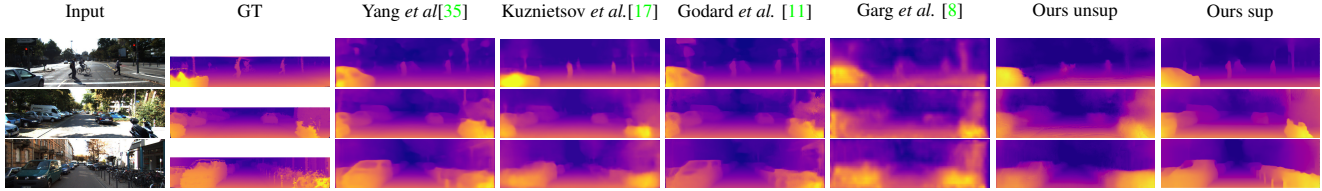


Figure 4: Visual comparison between the state-of-the-art methods. For visualization the groundtruth depth is interpolated. Our method captures more details in thin structures, such as the motorcycle and columns in the lower right corner of figure rows 2 and 3.

Methods	Dataset	Supervised		Error metric				Accuracy metric		
		depth	pose	RMSE	RMSE log	Abs Rel	Sq Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou <i>et al.</i> [41]	CS+K			6.709	0.270	0.183	1.595	0.734	0.902	0.959
Liu <i>et al.</i> [20]	K	✓		6.523	0.275	0.202	1.614	0.678	0.895	0.965
Eigen <i>et al.</i> [4]	K	✓		6.307	0.282	0.203	1.548	0.702	0.890	0.958
Yin <i>et al.</i> [37]	K			5.857	0.233	0.155	1.296	0.806	0.931	0.931
Zhan <i>et al.</i> [39]	K		✓	5.585	0.229	0.135	1.132	0.820	0.933	0.971
Zou <i>et al.</i> [43]	K			5.507	0.223	0.150	1.124	0.793	0.933	0.973
Godard <i>et al.</i> [11]	CS+K		✓	5.311	0.219	0.124	1.076	0.847	0.942	0.973
Atapour <i>et al.</i> [1]	K+S*	✓		4.726	0.194	0.110	0.929	<u>0.923</u>	0.967	0.984
Kuznetsov <i>et al.</i> [17]	K	✓	✓	4.621	0.189	0.113	0.741	0.875	0.964	0.988
Yang <i>et al.</i> [35]	K	✓		4.442	0.187	0.097	0.734	0.888	0.958	0.980
Fu <i>et al.</i> (VGG) [7]	K	✓		3.056	0.132	0.081	0.376	0.915	0.980	0.993
Ours-unsup (multi-view)	K			2.320	0.153	0.112	0.418	0.882	0.974	0.992
Ours-sup (single-view)	K	✓		<u>1.949</u>	<u>0.127</u>	0.088	<u>0.245</u>	0.915	<u>0.984</u>	<u>0.996</u>
Ours-sup (multi-view)	K	✓		1.698	0.110	0.077	0.205	0.941	0.990	0.998

Table 3: Quantitative comparison of our network with other state-of-the-art CNN-based methods on KITTI [10] dataset using the Eigen Split [4]. *Ours sup* (Single-view) is the evaluation of single view depth estimation result. Ours sup (multi-view) is the result generated with the assistance of nine previous views. Even though our method is not restricted to a fixed number of frames per sequence during prediction or evaluation, we still use 10-frame sequence here for the consistency with the training. We discuss continuous estimation results in the ablation study Section 4.5. The bold numbers are results that rank first and the underlined results those that rank second. All results are capped at 80m depth.

ber of frame intervals between source and target frame. We use the Adam [15] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is 0.0002. The training process is very time-consuming for our multi-view depth and odometry estimation network. One strategy we use to speed up our training process, without losing accuracy, is first to pretrain the network with the consecutive view reprojection loss for 20 epochs. Then we fine-tune the network with the multi-view reprojection loss for another 10 epochs.

4.2. Training datasets

We used the KITTI driving dataset [10] to evaluate our proposed framework. To perform a consistent comparison with existing methods, we used the Eigen Split approach [4] to train and evaluate our depth estimation network. From the 33 training scenes, we generated 45200 10-frame sequences. Here we used the stereo camera as two monocular cameras. A sequence of 10 frames contains either 10 left-camera or 10 right-camera frames. We resized the images from 375×1242 to 128×416 for computational efficiency and to be comparable with existing methods. The image re-

projection loss is driven by motion parallax, so we discarded all static frames with baseline motion less than $\sigma = 0.3$ meters during data preparation. 697 frames from the 28 test scenes were used for quantitative evaluation. For odometry evaluation we used the KITTI Odometry Split [10], which contains 11 sequences with ground truth camera poses. We follow [41, 38], which use sequences 00-08 for training and 09-10 as evaluation.

4.3. Depth Estimation

To evaluate the depth estimation component of our multi-view depth and odometry network, we compare to the state-of-the-art CNN-based depth estimation methods. Our network takes advantage of previous images and depths through recurrent units and thus achieves best performance when running on a continuous video sequence. However, it would be unfair to compare against single view methods when our method uses multiple views. On the other hand, if we also use only a single view for our method, then we fail to reveal the full capacity of our framework. Therefore, in order to present a more comprehensive depth evalu-

ation, we report both our depth estimation results with and without previous views’ assistance. *Ours-sup (single-view)* is the single view (or first view) depth estimation result of our framework, which also shows the bootstrapping performance of our approach. *Ours-sup (multi-view)* is the tenth view depth estimation result from our network. As shown in Table 3, *ours-sup (multi-view)* performs significantly better than all of the other supervised [20, 4, 1, 35, 7, 17] and unsupervised [41, 37, 43, 38, 11] methods. The unsupervised version of our network outperforms the state-of-the-art unsupervised methods as well as several supervised methods. Both the supervised and unsupervised version of our network outperform the respective state-of-the-art by a large margin. Figure 4 shows a visual comparison of our method with other methods. Our method consistently captures more detailed structures, e.g., the motorcycle and columns in the lower right corner of the figures in rows 2 and 3.

4.4. Pose Estimation

We used the KITTI Odometry Split to evaluate our visual odometry network. For pose estimation we directly ran our method through the whole sequence instead of dividing into 10-frame sub-sequences. We compared to the state-of-the-art learning-based visual odometry methods [38, 41, 37] as well as a popular monocular SLAM method: ORB-SLAM [25]. We used the KITTI Odometry evaluation criterion [10], which computes the average translation and rotation errors over sub-sequences of length (100m, 200m, ... , 800m).

Methods	Seq 09		Seq 10	
	$t_{err}(\%)$	$r_{err}(\text{deg/m})$	$t_{err}(\%)$	$r_{err}(\text{deg/m})$
ORB-SLAM [25]	15.30	0.003	3.68	0.005
GeoNet [37]	43.76	0.160	35.60	0.138
Zhan <i>et al.</i> [38]	11.92	0.036	12.62	0.034
Our unsupervised	9.88	0.034	12.24	0.052
Our supervised	9.30	0.035	7.21	0.039

Table 4: Quantitative comparison of visual odometry results on the KITTI Odometry dataset. t_{err} is the percentage of average translational error and r_{err} is the average degree per meter rotational error.

Both the monocular ORB-SLAM and the unsupervised learning-based visual odometry methods are suffering from scale ambiguity, so we aligned their trajectories with groundtruth prior to evaluation using *evo*¹. The supervised version of our method (absolute depth supervision) and the stereo supervised method [38] are able to estimate camera translations at absolute scale, so there is no post-processing for these two methods.

Table 4 shows quantitative comparison results based on the KITTI Visual Odometry criterion. Figure 5 shows a visual comparison of the full trajectories for all the methods.

¹github.com/MichaelGrupp/evo

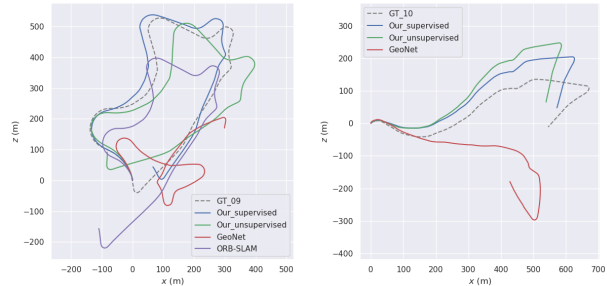


Figure 5: Visual comparison of full trajectories on Seq 09 (left) and 10 (right). Our predictions are closest to groundtruth (GT_09 and GT_10).

Including our method, all the full trajectories of learning-based visual odometry methods are produced by integrating frame-to-frame relative camera poses over the whole sequence without any drift correction.

The methods [41, 37] take a small sub-sequence (5 frames) as input and estimate relative poses between frames within the sub-sequence. There is no temporal correlation between different sub-sequences and thus the scales are different between those sub-sequences. However, our method can perform continuous camera pose estimation within a whole video sequence for arbitrary length. The temporal information is transmitted through recurrent units for arbitrary length and thus maintains a consistent scale within each full sequence.

4.5. Ablation study

In this section we investigate the important components: placements of the recurrent units, multi-view reprojection and forward-backward consistency constraints in the proposed depth and visual odometry estimation network.

Placements of recurrent units. Convolutional LSTM units are essential components for our framework to leverage temporal information in depth and visual odometry estimation. Thus we performed a series of experiments to demonstrate the influence of these recurrent units as well as the choice for the placements of recurrent units in the network architecture. We tested three different architecture choices which are shown in Figure 6. The first one is interleaving LSTM units across the whole network (full LSTM). The second one is interleaving LSTM units across the encoder (encoder LSTM). The third one is interleaving LSTM units across the decoder (decoder LSTM). Table 5 shows the quantitative comparison results. It can be seen that the encoder LSTM performs significantly better than the full LSTM and the decoder LSTM. Therefore, we chose the encoder LSTM as our depth estimation network architecture.

Multi-view reprojection and forward-backward con-

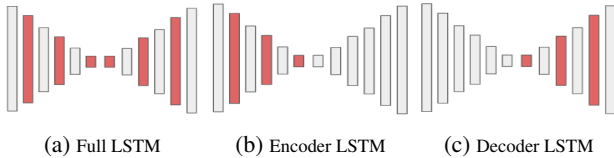


Figure 6: Three different architectures depend on the placements of recurrent units. (a) We put a convolutional LSTM after every convolution or deconvolution layer. (b) We only place convolutional LSTM in the encoder. (c) We only place convolutional LSTM in the decoder.

Method	RMSE	RMSE log	Abs Rel	Sq Rel
full LSTM	1.764	0.112	0.079	0.214
decoder LSTM	1.808	0.117	0.082	0.226
encoder LSTM	1.698	0.110	0.077	0.205

Table 5: Ablation study on network architectures. The evaluation data and protocol are the same as table 3.

sistency constraints. To investigate the performance gain from the multi-view reprojection and forward-backward consistency constraints, we conducted another group of experiments. Table 6 shows the quantitative evaluation results. We compared among three methods: with only the consecutive image reprojection constraint (*Ours-d*), with the consecutive image reprojection constraint and the forward-backward consistency constraint (*Ours-dc*), and with the multi-view reprojection constraint and the forward-backward consistency constraint (*Ours-mc*). The multi-view reprojection loss is more important in the unsupervised training, which is shown by the results of the last two rows in Table 6. Figure 7 shows a qualitative comparison between networks trained using consecutive image reprojection loss and those using multi-view reprojection loss. It can be seen that multi-view reprojection loss provides better supervision to areas that lack groundtruth depth.

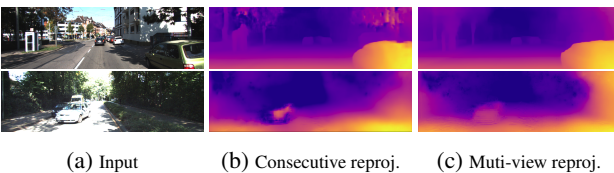


Figure 7: Visual examples between networks trained using consecutive image reprojection loss and those using multi-view reprojection loss. Results in the first row are from *ours-sup*, and results in the second row are from *ours-unsup*.

Estimation with different temporal-window sizes. Table 7 shows a comparison between depth estimation with different temporal-window sizes, i.e., the number of frames forming the temporal summary. Here we use the Eigen Split 697 testing frames for these sliding-window-based evalua-

Method	RMSE	RMSE log	Abs Rel	Sq Rel
Ours-d	1.785	0.116	0.081	0.214
Ours-dc	1.759	0.113	0.079	0.215
Ours-mc	1.698	0.110	0.077	0.205
Ours-dc unsup	2.689	0.184	0.138	0.474
Ours-mc unsup	2.361	0.157	0.112	0.416

Table 6: Ablation study on multi-view reprojection and forward-backward flow consistency constraints. *d* stands for consecutive image reprojection. *m* stands for multi-view image reprojection. *c* stands for forward-backward flow consistency constraint. The first three rows are comparison between supervised training and the last two rows are unsupervised.

Window size	RMSE	RMSE log	Abs Rel	Sq Rel
1	1.949	0.127	0.088	0.245
3	1.707	0.110	0.077	0.206
5	1.699	0.110	0.077	0.205
10	1.698	0.110	0.077	0.205
20	1.711	0.117	0.077	0.208
Whole seq.	1.748	0.119	0.079	0.214

Table 7: Depth estimation with different time-window sizes.

tions. In addition, we also ran through each whole testing sequence and again performed evaluation on those 697 testing frames. The result demonstrates that 1) the performance of the depth estimation is increasing with the number of depth estimations performed before the current estimation; 2) the performance of the depth estimation is not increasing after 10 frames; 3) even though our network is trained on 10-frame based sub-sequences, it can succeed on an arbitrary length sequences.

4.6. Conclusion

In this paper we presented an RNN-based, multi-view method for depth and camera pose estimation from monocular video sequences. We demonstrated that our method can be trained either supervised or unsupervised and that both produce superior results compared to the state-of-the-art in learning-based depth and visual odometry estimation methods. Our novel network architecture and the novel multi-view reprojection and forward-backward consistency constraints let our system effectively utilize the temporal information from previous frames for current frame depth and camera pose estimation. In addition, we have shown that our method can run on an arbitrary length video sequences while producing temporally coherent results.

References

- [1] A. Atapour-Abarghouei and T. P. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings*

- of the *IEEE Conference on Computer Vision and Pattern Recognition*, volume 18, page 1, 2018. 6, 7
- [2] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 2
- [3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1052–1067, 2007. 2
- [4] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015. 1, 2, 6, 7
- [5] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2018. 2
- [6] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [7] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 6, 7
- [8] 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*, pages 740–756. Springer, 2016. 1, 2, 6
- [9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 5
- [10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2, 5, 6, 7
- [11] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 2, 4, 6, 7
- [12] J. Hur and S. Roth. Mirrorflow: Exploiting symmetries in joint optical flow and occlusion estimation. In *Proc. of Intl Conf. on Computer Vision (ICCV)*, 2017. 4
- [13] G. Iyer, J. K. Murthy, K. Gunshi Gupta, and L. Paull. Geometric consistency for self-supervised end-to-end visual odometry. *arXiv preprint arXiv:1804.03789*, 2018. 4
- [14] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2, 4
- [15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [16] A. C. Kumar, S. M. Bhandarkar, and P. Mukta. Depthnet: A recurrent neural network architecture for monocular depth prediction. In *1st International Workshop on Deep Learning for Visual SLAM (CVPR)*, volume 2, 2018. 2
- [17] Y. Kuznetsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6647–6655, 2017. 6, 7
- [18] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016. 2
- [19] Z. Li and N. Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [20] 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*, pages 5162–5170, 2015. 1, 2, 6, 7
- [21] 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*, pages 3431–3440, 2015. 2
- [22] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single view stereo matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [23] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [24] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 3
- [25] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 2, 7
- [26] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtm: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011. 2
- [27] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [28] B. Rogers and M. Graham. Motion parallax as an independent cue for depth perception. *Perception*, 8(2):125–134, 1979. 1
- [29] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002. 2
- [30] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

- [31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3
- [32] 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 *IEEE Conference on computer vision and pattern recognition (CVPR)*, volume 5, page 6, 2017. 1, 2
- [33] 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. 4
- [34] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 1, 3
- [35] N. Yang, R. Wang, J. Stückler, and D. Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *European Conference on Computer Vision*, pages 835–852. Springer, 2018. 6, 7
- [36] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2
- [37] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 4, 5, 6, 7
- [38] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4, 5, 6, 7
- [39] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018. 6
- [40] J. Zhang. Inverse-consistent deep networks for unsupervised deformable image registration. *arXiv preprint arXiv:1809.03443*, 2018. 4
- [41] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. 1, 2, 4, 6, 7
- [42] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016. 2, 5
- [43] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *The European Conference on Computer Vision (ECCV)*, September 2018. 6, 7