DepthNet: A Recurrent Neural Network Architecture for Monocular Depth Prediction

Arun CS Kumar Suchendra M. Bhandarkar
The University of Georgia
aruncs@uga.edu suchi@cs.uga.edu

Mukta Prasad
Trinity College Dublin
prasadm@tcd.ie

Abstract

Predicting the depth map of a scene is often a vital component of monocular SLAM pipelines. Depth prediction is fundamentally ill-posed due to the inherent ambiguity in the scene formation process. In recent times, convolutional neural networks (CNNs) that exploit scene geometric constraints have been explored extensively for supervised single-view depth prediction and semi-supervised 2-view depth prediction. In this paper we explore whether recurrent neural networks (RNNs) can learn spatio-temporally accurate monocular depth prediction from video sequences, even without explicit definition of the inter-frame geometric consistency or pose supervision. To this end, we propose a novel convolutional LSTM (ConvLSTM)-based network architecture for depth prediction from a monocular video sequence. In the proposed ConvLSTM network architecture, we harness the ability of long short-term memory (LSTM)-based RNNs to reason sequentially and predict the depth map for an image frame as a function of the appearances of scene objects in the image frame as well as image frames in its temporal neighborhood. In addition, the proposed ConvLSTM network is also shown to be able to make depth predictions for future or unseen image frame(s). We demonstrate the depth prediction performance of the proposed ConvLSTM network on the KITTI dataset and show that it gives results that are superior in terms of accuracy to those obtained via depth-supervised and self-supervised methods and comparable to those generated by state-of-the-art pose-supervised methods.

1. Introduction

Scene reconstruction is one of the fundamental problems in computer vision research. In recent times, learning-based approaches to depth estimation have been explored and exploited widely for 3D scene reconstruction in a wide range of applications including simultaneous localization and mapping (SLAM) for self-driving cars and virtual reality (VR)-based and motion capture (MOCAP)-based gaming, to cite a few. For a given input image, infinite depth maps can be conjured up and determining the correct one is very difficult. However, by understanding the underlying scene semantics and employing suitable priors one can narrow down the possibilities to obtain realistic depth maps in a reasonable time frame. For example, for a continuous video of a slowly changing scene, the corresponding depth map also exhibits low temporal variation. Consequently, temporal smoothness is an important prior that is exploited in almost all current SLAM techniques.

Significant recent progress has been made in single- and multi-view 3D scene reconstruction, deriving 3D scene structure from motion (SfM) and simultaneous localization and mapping (SLAM) [9, 22]. However, accurate monocular depth prediction through deep learning is considered the ultimate test of the efficacy of modern learning- and prediction-based 3D scene reconstruction techniques. The ready availability of RGBD sensors (such as Kinect and LiDAR) in recent times has made acquiring pairs of images with accompanying depth maps considerably easier, at least for the purpose of learning, even if too expensive for perpetual deployment. Pre-calibrated stereo rigs also provide an effective substitute for RGBD sensors, but require reasoning about scene disparity. Monocular prediction using pairs of frames is the toughest, as one needs to reason about the relative camera pose as well as disparity/flow and there is an inherent ambiguity in scale, unless we resort to a consistent SLAM like reconstruction pipeline.

Learning to predict depth, even if only approximately, provides an opportunity to inject valuable information into the 3D reconstruction, 3D pose estimation and inference procedures in SLAM. Learning complex semantic scene relationships by capturing the spatio-temporal relationships between image entities (such as regions and textures) across different imaging modalities (such as RGB images and depth maps) calls for the formulation of complex learning models accompanied by large datasets. In recent times, aided by the rapid progress in deep learning methods and availability of large datasets [11], learning-based techniques for some of

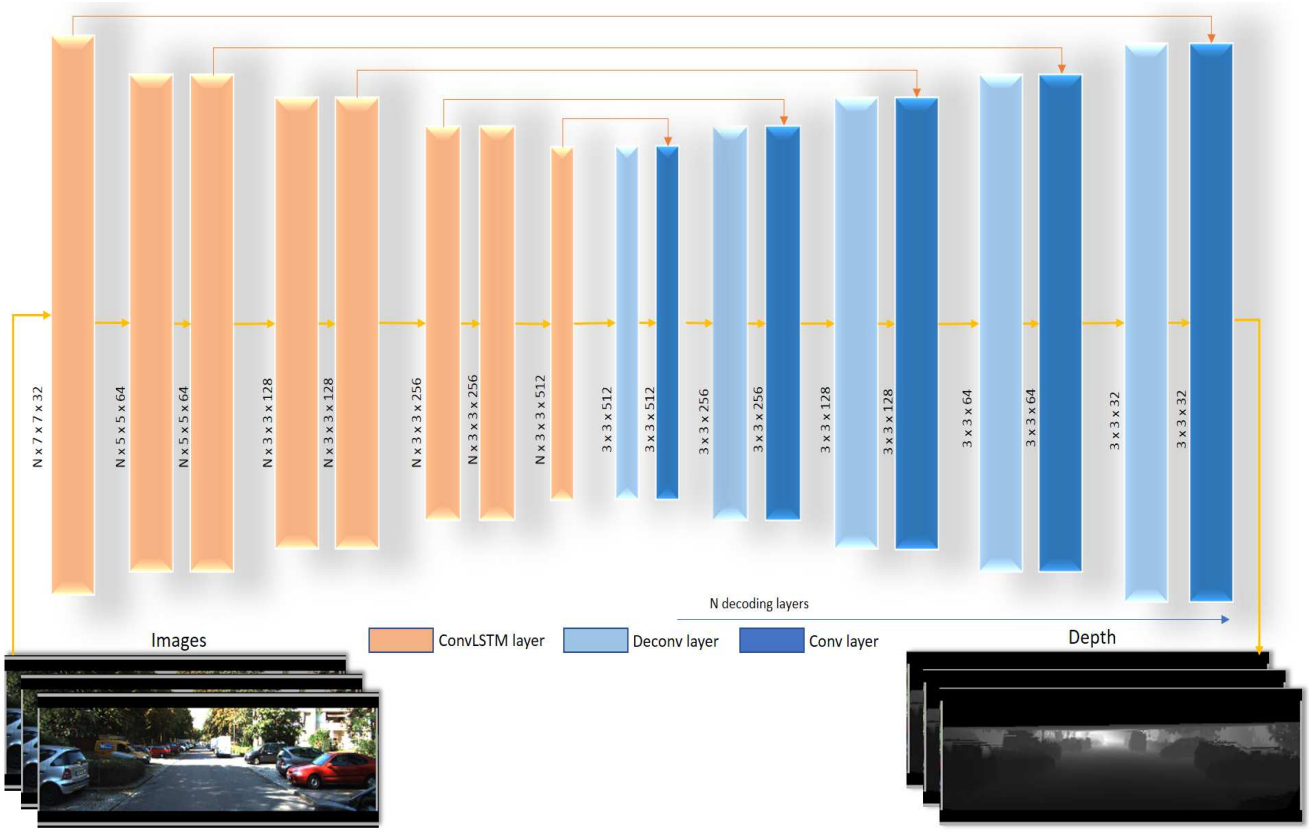

Figure 1: Proposed network architecture: The encoding layer consisting of multiple ConvLSTM [31] layers (orange blocks) takes a single image or image sequence as input at test time. The decoding layer consisting of an alternating sequence of deconvolutional and convolutional layers ((*blue blocks*) reconstructs the depth maps.

the sub-problems in SLAM has become viable, *e.g.*, [6, 19, 12, 32, 28]. Methods in predictive reconstruction also need to account for uncertainty and noise in the video sequence data, especially when distinct objects in a scene have motion parameters that are independent of the global camera motion parameters (*e.g.* multiple cars moving in different directions on a road). Additionally, unlike depth maps which can be readily obtained using depth sensors, getting accurate ground truth pose data for objects moving independently in a scene is much more difficult.

In this paper, we propose a scheme for learning object pose implicitly from sequences of image and depth map pairs for training. We demonstrate that we are able to effectively learn and predict depth as a function of image appearance over time using an LSTM-based deep learning model [27] The proposed model is shown to capture inter-frame dependencies and variations, without explicit modeling of the object pose. Given a sequence of images, the proposed spatio-temporal approach predicts the depth map using both the current image frame and its predecessors. We also demonstrate the performance of the proposed approach on predicting the depth maps for future or unseen image frames given the current image frame and/or previous image frames by harnessing the sequential reasoning capability of the LSTM.

While use of depth as supervision or priors based on reasoning about forward-backward image consistency have been substantially explored, temporal smoothness as a prior is relatively unexploited when using a video sequence as input. In this paper, we propose a novel convolutional LSTM-based recurrent neural network architecture that learns depth as a function of appearance while implicitly learning the object pose and its smooth temporal variation. The goal of the paper is to demonstrate that the use of temporal information is particularly effective in estimating depth. The proposed convolutional LSTM-based recurrent neural network architecture is depicted in Figure 1.

2. Related Work

Traditional *structure-from-motion* (SFM) and SLAM techniques jointly estimate structure and motion parameters, either using point correspondences [9, 22, 29, 30] or direct methods [7, 23]. In recent times, CNN-based ap-

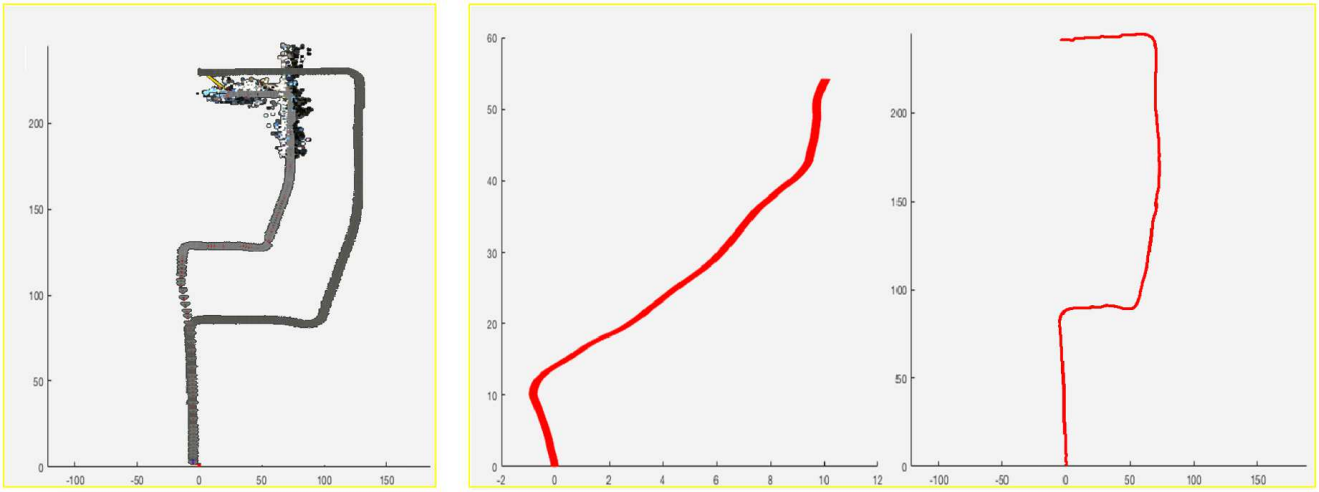

Figure 2: Comparison of pose predictions: Traditional SLAM-based camera calibration versus deep learning-based pose regression. The 3D coordinates obtained using camera matrices are plotted for an example image sequence from the KITTI dataset [11]. (a) Coordinates plotted in *dark gray* represent the ground truth path or trajectory whereas the path plotted in *light gray* represents the output of a traditional SLAM algorithm [1] on a Kitti sequence, without inertial measurements; *(b)* Path plotted using camera parameters estimated via deep learning techniques [32] (left) versus the ground-truth plot (right). We observe considerable drift between the *dark gray* and *light gray* plots in the case of the traditional SLAM algorithm (a) especially in the absence of inertial measurements. However, the traditional SLAM algorithm is able to estimate the object pose over longer image sequences with higher accuracy than the deep learning based pose prediction technique (b).

proaches [6, 18, 12, 28, 17] have been seen to achieve good performance on well constrained subsets of the general monocular reconstruction problem, *e.g*. predicting depth and pose given a single or 2-3 consecutive images in space and/or time. The ability of CNN-based approaches to capture complex relationships between the depth maps and the corresponding image textures along with other scene semantics has been demonstrated in [6, 12, 28]. These CNN-based approaches are trained either in a depth supervised manner [6, 18] from a single view to predict corresponding depth maps, or in a pose/self-supervised mode employing photo-consistency with input stereo images [12, 17] or consecutive images in time and their capabilities are often demonstrated on single-view depth prediction [28, 32]. The *FlowNet* architecture proposed by Dosovitskiy *et al.* [5] for optical-flow estimation is based on an encoder-decoder CNN where channel-concatenated image pairs are provided as input to the network that learns optical flow in a supervised fashion. A CNN variant termed as *FlowNetCorr*, merges two different convolutional networks from pairs of adjacent images, by correlating the tensors to learn the disparity map, mimicking the traditional point correspondence-based optical flow methods.

Since the formulation of *FlowNet* and *FlowNetCorr* architectures, several approaches have proposed encoder-decoder CNN architectures for computing disparity maps, which are subsequently used for depth prediction [12, 21, 28]. Additionally, the architecture proposed by Godard *et al.* [12] learns to minimize the left-right consistency between adjacent image pairs to improve the pose estimation accuracy in a pose-supervised setting. The stereo problem formulated in [12] assumes a known pose, making it equivalent to one of estimating depth through disparity. On the other hand, Zhou *et al.* [32] propose a joint pose and depth prediction technique that learns reconstruction up to scale in an unsupervised setting using video frames as input. In addition, explainabilty masks are used to isolate individual motions of objects that do not agree with the predicted motion parameters of the scene [32]. Vijayanarasimhan *et al.* [28] extend the work of Zhou *et al.* [32] to model the individual motions of the objects isolated using the explainability maps. In their more recent work, DeTone *et al.* [4] propose a CNN-based approach that identifies isolated and evenly distributed feature points from image pairs, which are then fed to another neural network that learns to compute the homography between them.

While the use of object pose information has been shown to improve depth prediction accuracy considerably [12], object pose predicted using optical flow-based approaches on real-world images [28], is far from accurate, and in fact often

falls short in comparison to traditional point correspondence-based pose estimation approaches. Figure 2 displays visual plots of traditional SLAM-based object pose predictions and deep learning-based pose estimations where the former are shown to outperform the latter in terms of pose estimation accuracy. Eigen *et al*. [6] propose a simple, but effective scheme for monocular depth prediction from image appearance, albeit with considerable supervision. Their network consists of two components, the first component is a traditional convolutional network that captures coarse global scene structure, followed by the refinement of the estimated coarse depth map using the image color/texture information. In addition, they also propose a scale-invariant error measure to address the global scale ambiguity. In this paper, we extend the scheme proposed by Eigen *et al.* [6], to predict depth from an image explicitly, while modelling pose information implicitly via temporal reasoning using LSTMs.

2.1. Recurrent Neural Networks for Temporal Learning

Recurrent neural networks (RNNs) [14] are a class of neural networks that models the temporal behavior of sequential data using hidden states with cyclic connections. In feed-forward convolutional networks the gradient is back-propagated through the network; an RNN additionally back-propagates through time (across multiple instances of the network) that enables them to learn dependencies across time. The long short-term memory (LSTM) [14] is an extension of the traditional RNN, that is capable of learning long-term dependencies within an input sequence.

Recently several approaches have used the LSTM for learning temporal dependencies across image frames in a video sequence [8, 20, 26]. Srivastava *et al*. [26] learn a video representation using an encoder-decoder framework which is then used for future frame prediction. Similarly, Lotter *et al*. [20] propose a predictive neural network that is inspired by predictive coding and trained in an unsupervised fashion for the purpose of video prediction. Choy *et al.* [3] present a scheme for learning the mapping between images and the corresponding 3D object shapes using a gated recurrent unit GRU [2], a variant of LSTM with fewer parameters. At test time Choy *et al.* [3] reconstruct a 3D occupancy grid for the underlying scene using one or more images.

In our paper, we employ a convolutional LSTM (ConvLSTM) [31], to model the spatio-temporal dependencies between video frames for the purpose of predicting depth maps. The use of the ConvLSTM instead of the traditional fully-connected LSTM allows us to jointly exploit the ability of the multiple convolutional layers to capture appearance cues at multiple spatial scales along with the ability of the LSTM to reason about the temporal variations in the input data, without losing any spatial information.

The major contributions of our paper can be summarized as follows:

- We adapt the convolutional LSTM (ConvLSTM)-based encoder-decoder architecture for scene depth prediction from monocular video sequences. Given temporally adjacent image frames and their corresponding coarse ground truth depth maps, the proposed ConvLSTM network has the opportunity to learn a spatio-temporal mapping between the image and depth data. At test time, the network can predict depth maps from both, image sequences and single images.

- We demonstrate the ability of the network to reason sequentially, by extrapolating the current depth maps for the future (or unseen) image frames, without explicitly training it to do so.

- We present new results for monocular depth prediction, on the KITTI dataset [11], that outperforms other depth(only) [6, 19], pose/stereo-supervised [12, 10] and other self-supervised [32] methods, and are comparable to some state-of-the-art [17] that use depth+pose/stereo.

3. Proposed Approach

We propose using a convolutional LSTM (ConvLSTM)-based network architecture for depth prediction from a monocular video sequence. In contrast to traditional depth prediction models that process a single input image, the proposed ConvLSTM network learns depth maps from a set of N consecutive video frames in a depth-supervised setting, allowing the ConvLSTM network to perform spatio-temporal reasoning about the image-depth map relationship. In addition, unlike the traditional LSTM-based approaches [3], where a *fully-connected* LSTM layer is introduced between *encoder* and *decoder* networks, we stack a set of ConvLSTM layers [31] on top of each other to construct the encoding phase. The ConvLSTM is a variant of the traditional *fully connected* LSTM that has convolutional structures underneath. The traditional LSTM layer unfolds the input tensor into a vector, thus does not take into consideration the spatial correlations between the grid cells [31]. The advantage of stacking multiple ConvLSTM layers is that, the multiple LSTM layers allows the network to better learn the temporal information whereas the underlying convolutional structure helps retain the spatial relationships between the grid cells. Moreover, the use of the ConvLSTM also reduces the total number of trainable network parameters significantly [31]. This is in contrast to the fully connected LSTM layers which unfold to generate a densely connected vector with much spatial data redundancy. In the proposed network, the ConvLSTM layers can be shown to effectively capture the spatio-temporal information with much higher accuracy than the traditional LSTM layers.

In the proposed network, the encoder layer is comprised of ConvLSTM layers, each layer holding N states where N is the total number of timestamps. The decoder layer reconstructs the depth maps learned for each of the states separately. The decoder layer follows an architecture similar to that of the U-Net [24], with N separate deconvolutional layers and skip connections between the encoder and decoder layers. This decoder architecture, which has been shown to work well for several reconstruction tasks [32], also allows for more accurate reconstruction of the depth map. In the proposed network, the encoder layer learns the spatio-temporal relationships between N image frames for the purpose of predicting the depth maps whereas the decoder layer learns to reconstruct the N individual depth maps.

4. Network Architecture

The proposed network architecture illustrated in Figure 1, consists of an encoding (contraction) phase followed by a decoding (expansion) phase. The encoding (contraction) phase takes as input a set of N consecutive video frames and computes an intermediate depth representation towards the end. The decoding (expansion) phase reconstructs the depth maps from the intermediate depth representation. The encoding (contraction) phase comprises of a stack of K ConvLSTM layers that takes as input an image sequence across N time points to learn the depth representation as a function of time. As mentioned above, in comparison with a *Fully-connected* LSTMs (FC-LSTM) [27], ConvLSTM [31] has fewer connections, with shared weights, that make them easier to learn than the dense connections of the traditional FC-LSTMs. In addition FC-LSTM as opposed to ConvLSTM, does not take spatial correlation into consideration [31]. Thus the ConvLSTM well suited for learning the underlying representation from spatio-temporal data. A more comprehensive description of the network architecture and details pertaining to its training are provided in Section 5.

Unlike traditional ConvNets where the network learns and predicts from a single view and the error is propagated from the bottom layer of the decoder to the top layer of encoder, in our case the error propagates temporally thereby capturing the time-dependent progression of the scene depth. While the contraction (encoding) phase learns to encode depth as a spatio-temporal function of the image sequences, the expanding (decoding) phase learns to reconstruct the depth map from the intermediate representation. We use deconvolutional layers with skip connections for the purpose of depth reconstruction. The architecture for the decoding phase is detailed in Section 5. Towards the end of the expansion phase, we use a 1×1 convolutional layer with sigmoid activation to obtain the depth map. For training, we use a scale-invariant error metric proposed by Eigen *et al*. [6] as the loss function. Given a predicted depth map y_i and its ground truth depth map y_i^*, the loss function [6] is given by:

$$L(y, y^*) = \frac{1}{n}\sum_i d_i^2 - \frac{\lambda}{n^2}\left(\sum_i d_i\right)^2 \quad (1)$$

where $d_i = log(y_i) - log(y_i^*)$ for the i^{th} pixel and n corresponds to the total number of pixels, with the value of λ set to 0.5 [6].

5. Implementation Details

Encoding Phase: The encoding phase consists of a series of 3×3 ConvLSTM layers consisting of {32, 64, 64, 128, 128, 256, 256, 512} filters respectively, with alternating strides of 2's and 1's, except for the first two layers that use filters of size 7×7 and 5×5 respectively. While the *relu* activation function is used in each convolutional step of the ConvLSTM layer, the recurrent step uses the *hard sigmoid* activation function. The padding is set to be the same for all layers. The first ConvLSTM layer takes an input I of size $\{B \times N \times H \times W \times C\}$, where H and W are the height and width of the image respectively, C represents the number of image channels (C is 3 as we use standard RGB images), N is the total number of time steps and B is the batch size. In our experiments we set the value of N to 3. Each ConvLSTM layer is designed to return the entire sequence (comprising of all states) instead of just the final state, so that it can be used in the decoding phase.

Decoding Phase: The decoding layer consists of alternating sequence of deconvolutional and convolutional layers. The deconvolutional or transposed convolutional layer takes as input the output sequence the last ConvLSTM layer and performs a deconvolution operation on it. We also use skip connections across the encoding and decoding phases, by concatenating the deconvolved tensor with the output of the corresponding original convolutional layer. Since concatenating the tensors doubles the number of channels, we affix the concatenation layer with an additional convolutional layer to reduce the tensor size. In recent times, the use of skip connections has shown to work well, especially when dealing with the vanishing gradient problem [13] thereby effectively allowing the exploration of deeper network architectures. In a manner similar to the encoding phase, we use a series of deconvolutional layers of sizes $\{512, 256, 256, 128, 128, 64, 64, 32\}$ respectively (which are then followed by a concatenation and convolutional layer for the skip connections) with alternating strides of 2's and 1's. The filter sizes are set to 3×3 for all deconvolutional layers. Towards the end, we use a 1×1 convolutional layer with a $sigmoid$ activation function that converts the tensor to a depth map.

Training: The network is trained with inputs as temporally concatenated image sequences of size N, with the corre-

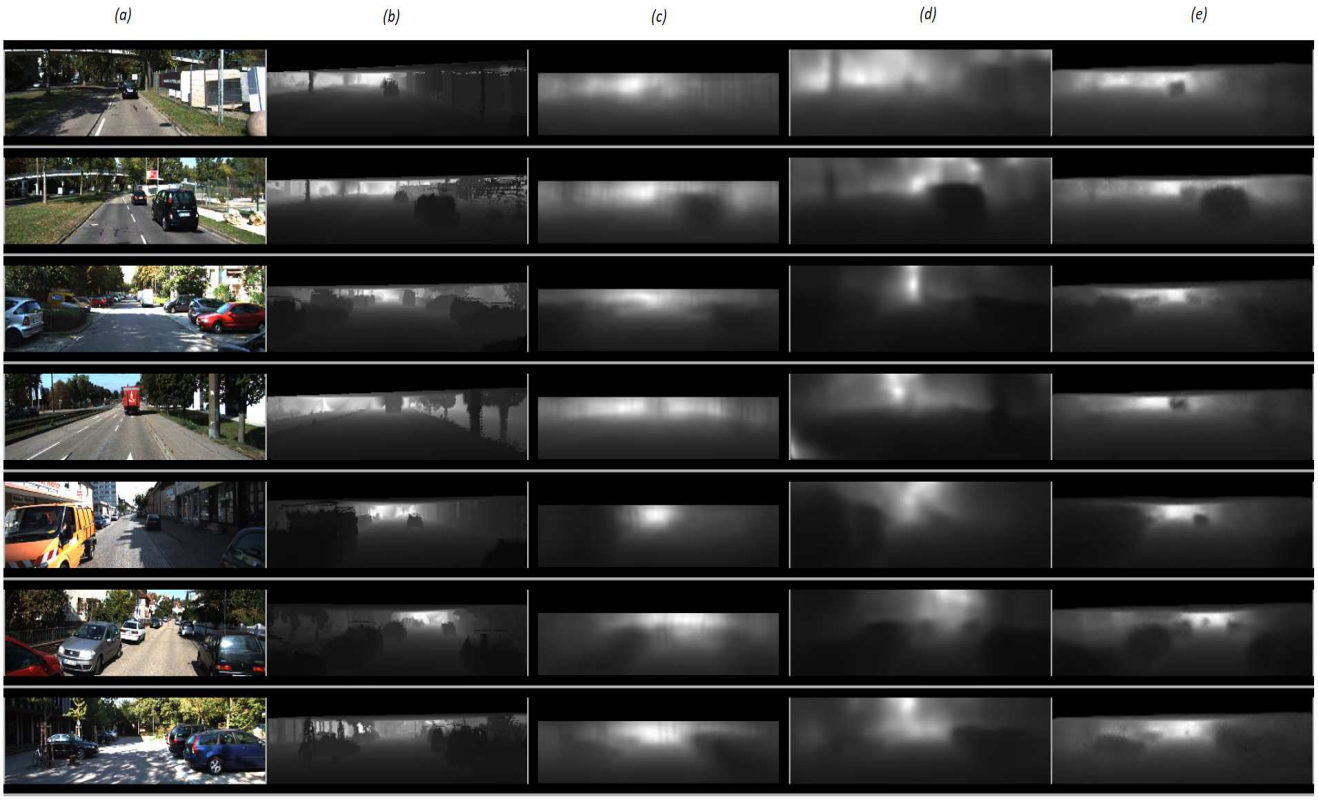

Figure 3: Qualitative results (*good*). (a) Image (t) (b) Corresponding ground truth depth map (c) Depth predictions from Eigen *et al.* [6] (depth-supervised) (d) Depth predictions from Zhou *et al*. [32] (e) Depth predictions from the proposed scheme (*note*: the proposed scheme uses $N-1$ preceding images in addition to the test image (a), for predicting the depth map)

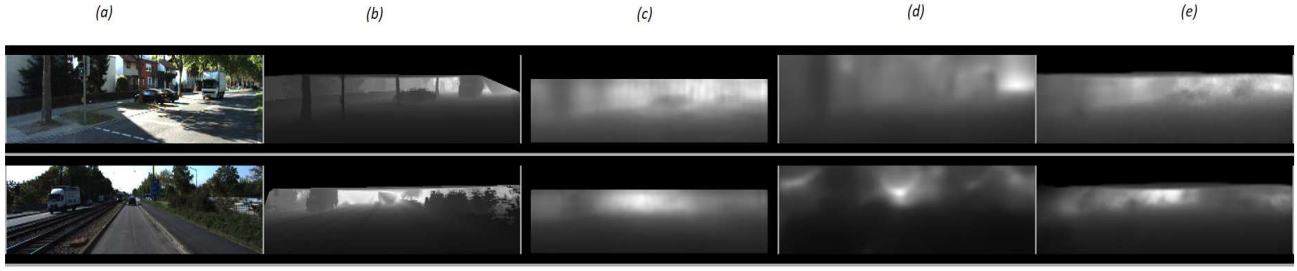

Figure 4: Qualitative results (*bad*). (a) Image (t) (b) Corresponding ground truth depth map (c) Depth predictions from Eigen *et al.* [6] (depth-supervised) (d) Depth predictions from Zhou *et al*. [32] (e) Depth predictions from the proposed scheme (*note*: the proposed scheme uses $N-1$ preceding images in addition to the test image (a), for predicting the depth map)

sponding ground truth depth map as output. We use *batch normalization* [15] layers after each pair of convolution or deconvolutional layers, and train the network using the Adam optimizer [16], with a learning rate of 1×10^{-4} and the loss function described in Section 4. In most cases, the validation loss is observed to converge within 20 epochs. Details regarding the dataset split are provided in Section 6.

6. Evaluation

We train and evaluate our model on the KITTI dataset [11]. The KITTI dataset consists of video sequences of outdoor scenes along with their corresponding depth maps, procured using car-mounted cameras and Velodyne LiDAR sensors. We use the train/test split described in [6], where we train on 28 sequences and test on the 697 images provided

Table 1: Comparison of monocular depth prediction results on KITTI dataset [11].

θ	Supervision			Error Metric				Accuracy Metric		
	Depth	Pose	Unsupervised	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al. [6]* (Coarse)	✓			0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen *et al. [6]* (Fine)	✓			0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al. [19]*	✓			0.202	1.614	6.523	0.275	0.678	0.895	0.965
(**Ours**–*image sequence*)	✓			**0.137**	**1.019**	**5.187**	**0.218**	**0.809**	**0.928**	**0.971**
(**Ours**–*single image*)	✓			0.176	1.3711	5.971	0.265	0.740	0.896	0.959
(**Ours** t_{n+1} *frame*)	✓			0.296	3.251	9.849	0.469	0.535	0.749	0.855
(**Ours**–*CNN*)	✓			0.145	1.062	5.424	0.273	0.754	0.904	0.969
Godard *et al.* [12]		✓		0.148	1.344	5.927	0.247	0.803	0.922	0.964
Garg *et al.* [10] (50m cap)		✓		0.169	1.080	5.104	0.273	0.740	0.904	0.962
Zhou *et al.* [32](w/ exp. mask)			✓	0.221	2.226	7.527	0.294	0.676	0.885	0.954
Zhou *et al.* [32]			✓	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou *et al.* [32] (50m cap)			✓	0.208	1.551	5.452	0.273	0.695	0.900	0.964
Kuznietsov et al. [17]	✓	✓(stereo)		**0.113**	**0.741**	**4.621**	**0.189**	**0.875**	**0.964**	**0.988**
Kuznietsov et al. [17]		✓(stereo)		0.308	9.367	8.700	0.367	0.752	0.904	0.952

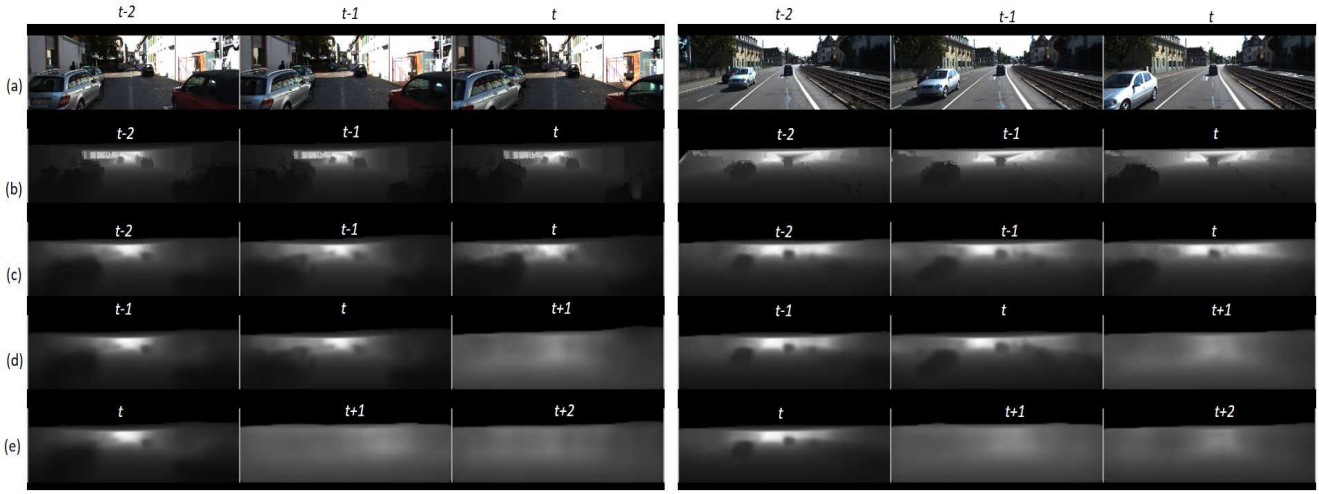

Figure 5: Qualitative demonstration of depth prediction results for the current image frame and future image frames obtained by unfolding the LSTM layers (a) Images at times t_{n-2}, t_{n-1}, t_n, (b) The corresponding ground truth depth maps (c) Depth predictions of the proposed ConvLSTM network for image frames at time steps t_{n-2}, t_{n-1}, t_n (which is the way the network is actually trained to predict) (d) Predictions of the proposed ConvLSTM network, for future frames at time t_{n-1}, t_n, t_{n+1} (e) predictions of the proposed ConvLSTM network for future frames at time t_n, t_{n+1}, t_{n+2}. For (d) & (e), it must be noted that, the proposed ConvLSTM network is not trained to predict future frames; instead we mask the inputs for specific time steps and force the network to predict the frames, thereby exploiting its recurrent nature. Qualitative analysis of the results over several images showed that the proposed network was able to reliably estimate the layout of the scene, but failed to interpolate accurately the motion of the scene objects into the future.

in [6]. Throughout our experiments, the number of time steps for training the ConvLSTM is set to 3. We evaluate our approach by using the standard metrics proposed by [6].

6.1. Depth Prediction from Monocular Sequences and Single Images

Although the proposed network is trained using fixed-length image sequences, at test time we evaluate its performance on both monocular image sequences and single images. In addition, we also evaluate, qualitatively and quantitatively, the accuracy of the extrapolated depth maps corresponding to future (or) unseen image frame(s).

Monocular Image Sequences: For predicting depth on image sequences, we gather image sequences of of size N —of the 697 images provided by [6], 23 images are the first of

their respective sequences, which we had to omit for this experiment as they have no preceding images. We tabulated the results of our approach against other state-of-the-art in Table 1. The proposed approach is observed to outperform depth-supervised approaches while yielding results comparable to those of pose-supervised techniques [10, 12]. The qualitative depth prediction results are shown in Figures 3 and 4, where Figure 4 shows instances where the proposed approach fails to predict the scene depth reasonably. In order to demonstrate the improvement due to the use of the ConvLSTM component over the CNN baseline, we trained a CNN with a similar architecture and reported the results in Table 1. Our results outperforms most (reported) state-of-the-art methods that are depth-supervised [6, 19], pose/stereo-supervised [12, 10] supervised and other self-supervised approaches [32]. While our numbers are inferior to [17] (Table 1, *row* 13), it must also be noted that [17], uses both depth and stereo as supervision, along with more sophisticated and deeper network architecture (ResNet [13]) with pre-trained weights, as opposed to our less sophisticated network (in terms of depth) that does not rely on pre-trained weights [25].

Single Images: Although the network is trained using image sequences, the decoding layer is designed to individually reconstruct each state of the encoding phase. Doing so allows us to use a single image at test time, and still get reliable depth reconstruction, although the network is trained using image sequences only. Using a single image at test time would mean that only the first recurrent layer will receive input (others will get empty placeholders), in which case the network will act like an end-to-end ConvNet or CNN instead of a recurrent network. The quantitative results for depth prediction using single images are presented in Table 1. The results are comparable to the predictions obtained using monocular sequences and are even better than those of most other approaches.

Future Depth Prediction: As an attempt to exploit the ability of the LSTM to reason temporally, we analyze quantitatively its ability to predict depth maps of future frames. The goal of the experiment is to see how well the network is able to learn inter-frame dependencies. For that purpose, in a manner similar to our previous experiment, we replace images in the image sequence with empty placeholders, and force the network to predict depth maps. The quantitative results are presented in Table 1. Though the prediction results are not comparable, the future prediction results show, both qualitatively and quantitatively, how the information propagates over time, and how well the network is able to learn inter-frame dependencies. The qualitative results for future depth prediction are depicted in Figure 5. The results suggest that the future frame predictions, though not quite accurate especially when modeling individual objects, are still able to estimate the layout of the scene reasonably well. Also, it has to be noted that we do not train the network explicitly for predicting future image frames, instead we simply force the network to predict by masking the input(s).

7. Conclusion

In this paper we explored whether recurrent neural networks (RNNs) can learn spatio-temporally accurate monocular depth prediction from video sequences, even without explicit definition of the inter-frame geometric consistency or pose supervision. To this end, we proposed a novel convolutional LSTM (ConvLSTM)-based network architecture for depth prediction from a monocular video sequence. In the proposed ConvLSTM network architecture, we harnessed the ability of long short-term memory (LSTM)-based RNNs to reason sequentially and predict the depth map for an image frame as a function of the appearances of scene objects in the image frame as well as image frames in its temporal neighborhood. We demonstrated quantitatively and qualitatively that the proposed ConvLSTM is able to perform better at depth prediction than traditional CNN models, by obtaining convincing state-of-the-art results on the KITTI dataset compared to current depth-supervised approaches. Although our network is trained to make depth predictions for image sequences, it can predict depth maps, at test time, on single images as well with high accuracy. Also, we have demonstrated the network's ability to reason temporally, by extrapolating depth maps for future/unseen frames, without the network being explicitly trained to do so. In the future, we plan to automatically learn explainability masks, that would model individually each independently moving object in the scene. The explainability masks could then be used for predicting the depth map for each individual object, in the current image frame and in future image frames, more accurately.

References

[1] Bailey, T., & Durrant-Whyte, H. (2006). Simultaneous localization and mapping (SLAM): Part II. *IEEE Robotics & Automation Magazine*, 13(3), 108-117. 3

[2] Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. 4

[3] Choy, C. B., Xu, D., Gwak, J., Chen, K., & Savarese, S. (2016, October). 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In *Proc. European Conference on Computer Vision* (pp. 628-644). Springer International Publishing. 4

[4] DeTone, D., Malisiewicz, T., & Rabinovich, A. (2017). Toward Geometric Deep SLAM. arXiv preprint arXiv:1707.07410. 3

[5] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., & Brox, T. (2015) Flownet: Learning optical flow with convolutional networks. In *Proc. IEEE International Conference on Computer Vision*, (pp. 2758-2766). 3

[6] Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems* (pp. 2366-2374). 2, 3, 4, 5, 6, 7, 8

[7] Engel, J., Schöps, T., & Cremers, D. (2014, September). LSD-SLAM: Large-scale direct monocular SLAM. In *Proc. European Conference on Computer Vision* (pp. 834-849). Springer. 2

[8] Finn, C., Goodfellow, I., & Levine, S. (2016). Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems* (pp. 64-72). 4

[9] Furukawa, Y., Curless, B., Seitz, S. M., & Szeliski, R. (2010, June). Towards internet-scale multi-view stereo. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1434-1441). 1, 2

[10] Garg, R., Carneiro, G., & Reid, I. (2016, October). Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Proc. European Conference on Computer Vision* (pp. 740-756). Springer International Publishing. 4, 7, 8

[11] Geiger, A., Lenz, P., & Urtasun, R. (2012, June). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3354-3361). 1, 3, 4, 6, 7

[12] Godard, C., Mac Aodha, O., & Brostow, G. J. (2016). Unsupervised monocular depth estimation with left-right consistency. arXiv preprint arXiv:1609.03677. 2, 3, 4, 7, 8

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778). 5, 8

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, Vol. 9(8) (pp. 1735-1780). 4

[15] Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. International Conference on Machine Learning* (pp. 448-456). 6

[16] Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 6

[17] Kuznietsov, Y., Stückler, J., & Leibe, B. (2017, February). Semi-supervised deep learning for monocular depth map prediction. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6647-6655). 3, 4, 7, 8

[18] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016, October). Deeper depth prediction with fully convolutional residual networks. In IEEE 3D Vision (3DV), 2016 Fourth International Conference on (pp. 239-248). 3

[19] Liu, F., Shen, C., Lin, G., & Reid, I. (2016). Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38(10) (pp. 2024-2039). 2, 4, 7, 8

[20] Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104. 4

[21] Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4040-4048). 3

[22] Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, Vol. 31(5) (pp. 1147-1163). 1, 2

[23] Newcombe, R. A., Lovegrove, S. J., & Davison, A. J. (2011, November). DTAM: Dense tracking and mapping in real-time. In *Proc. IEEE International Conference on Computer Vision* (pp. 2320-2327). 2

[24] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241). Springer, Cham. 5

[25] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252. 8

[26] Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015, June). Unsupervised learning of video representations using LSTMs. In *Proc. International Conference on Machine Learning* (pp. 843-852). 4

[27] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112). 2, 5

[28] Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., & Fragkiadaki, K. (2017). SfM-Net: Learning of Structure and Motion from Video. arXiv preprint arXiv:1704.07804. 2, 3

[29] Williams, B., Klein, G., & Reid, I. (2007, October). Real-time SLAM relocalisation. In *Proc. IEEE International Conference on Computer Vision* (pp. 1-8). 2

[30] Wu, C. (2011). VisualSFM: A visual structure from motion system. http://ccwu.me/vsfm/ 2

[31] Xingjian, S. H. I., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems* (pp. 802-810). 2, 4, 5

[32] Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. arXiv preprint arXiv:1704.07813. 2, 3, 4, 5, 6, 7, 8