DEEP MONOCULAR VIDEO DEPTH ESTIMATION USING TEMPORAL ATTENTION

Haoyu Ren, Mostafa El-khamy, Jungwon Lee

Samsung Semiconductor Inc., San Diego, CA, USA

ABSTRACT

Monocular video depth estimation (MVDE) plays a crucial role in 3D computer vision. In this paper, we propose an end-to-end monocular video depth estimation network based on temporal attention. Our network starts by a motion compensation module where the spatial temporal transformer network (STN) is utilized to warp the input frames using the estimated optical flow. Next, a temporal attention module is used to combine features from the warped frames, while emphasizing the temporal consistency. A monocular depth estimation network is used to estimate the depth from the temporally combined features. Experimental results demonstrate that our proposed framework achieves better performance compared to the state-of-the-art single image depth estimation (SIDE) networks, as well as existing MVDE methods.

***Index Terms*—** Depth estimation, temporal attention, spatial temporal transformer, optical flow

1. INTRODUCTION

Depth prediction from monocular images [1][2][3][4] and videos [5][6][7][8] using CNNs has a surge of interest in recent years. In particular, the depth map can be used to infer the 3D structure, which is the basic element of many topics in 3D vision, such as image reconstruction, image rendering, and shallow depth of the field. Most of current researchers focus on extracting depth from single image only (SIDE). Recently, some researchers start using temporal information to assist the depth estimation, which is helpful to understand the context of the moving objects. These methods focus on learning the camera pose [7][6] or intrinsic matrix [8] to obtain the 3D geometry. The 3D geometry is utilized to back-project 2D frames to 3D with estimated depth. These frames are aligned in 3D space and perspective projected to 2D again. The photometric consistency is minimized to optimize the whole framework.

There are two problems of the above framework. Firstly, these methods require additional network to estimate the 3D geometry such as the camera pose or the intrinsic matrix. This makes the whole framework slow. Secondly, most of the existing MVDE methods are unsupervised since the photometric error doesn't consider the ground-truth depth. The accuracy will be limited compared to supervised methods such as applying supervised SIDE method frame by frame.

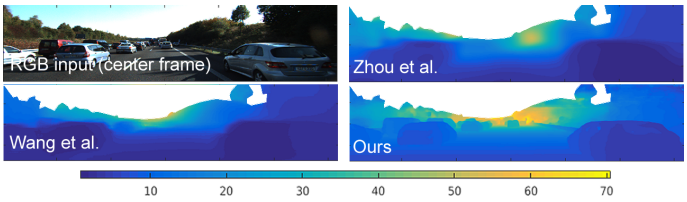

Fig. 1. Our results compared to existing video depth estimation methods from Zhou et al. [6] and Wang et al. [7]. The sky region is masked by white since sky doesn't have meaningful depth.

In this paper, we propose an end-to-end monocular video depth estimation method based on temporal attention, which can achieve high accuracy and considerable efficiency concurrently. Our method consists of three modules, the motion compensation, the temporal attention, and the depth estimation. In the motion compensation module, an efficient spatial temporal transformer network (STN) is designed to obtain the optical flow to align the multi-frame input with a reference frame. The photometric error is minimized when optimizing STN, so that the optical flow ground-truth is not required during the training. In the temporal attention module, a temporal attention layer is adopted to emphasize the temporal consistency among warped frames. Then a depth estimation network is applied to obtain the final predicted depth. Our framework is trained end-to-end, while optimizing the motion compensation module and the depth estimation module at the same time. Experimental results demonstrate that our video depth estimation framework can achieve state-of-the-art performance on the commonly-used KITTI dataset, with better efficiency compared to existing SIDE and MVDE methods. Figure 1 illustrates the outputs of our method on images from KITTI dataset, as well as the outputs of existing MVDE methods [6][7]. It can be observed that our estimated depth is clearly better than other approaches.

2. MONOCULAR VIDEO DEPTH ESTIMATION

As illustrated in Figure 2, our proposed video depth estimation network consists of three modules, motion compensation, temporal attention, and depth estimation. It takes three con-

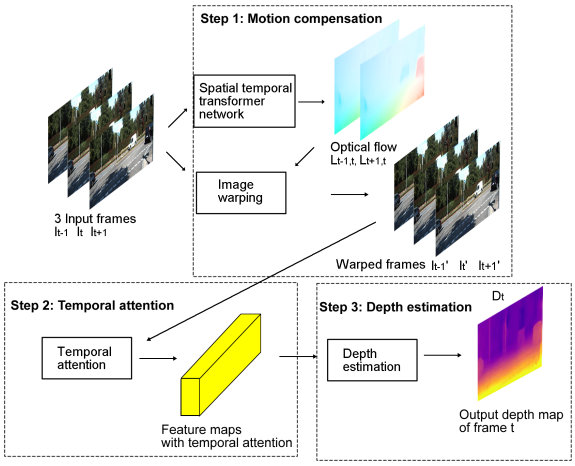

Fig. 2. Our proposed video depth estimation framework.

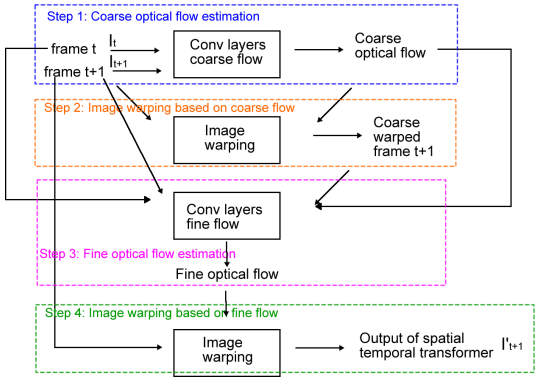

Fig. 3. Our coarse-to-fine spatial temporal transformer network (STN) for motion compensation.

secutive RGB frames I_{t-1}, I_t, I_{t+1} as input, and outputs the depth for the center frame D_t.

2.1. Motion compensation

The motion compensation module takes three consecutive frames I_{t-1}, I_t, I_{t+1} as input and outputs the motion compensated frames $I^{'}_{t-1}, I^{'}_t, I^{'}_{t+1}$ by image warping with optical flow. There have been many optical flow algorithms proposed in these years, such as FlowNet2 [9] or PWCNet [10]. These methods can obtain dense accurate optical flow, but with the cost of large network size. It will increase the overall video depth estimation network size and the computational cost significantly. In this paper, we propose the use of an efficient spatial temporal transformer network (STN) to compensate the motion between frames fed to the video depth estimation network. We will compensate three consecutive frames in our implementation, but for simplicity we first introduce the motion compensation between two frames.

As given in Fig. 3, we adopt a coarse-to-fine design in the spatial temporal transformer to obtain the optical flow. First, a coarse estimate of the flow is obtained by fusing the two input frames and feeding into 5 convolutional layers with 32 3×3 filters. The estimated flow is applied to warp the target frame producing coarse warped frame t+1 (see orange dashed box). This warped image is then processed together with the coarse flow and the original images through a fine flow estimation module (see pink dashed box), which is implemented by another 5 convolutional layers with 32 3×3 filters to obtain a fine optical flow. This fine flow is utilized for final image warping, which produces the output of STN (see green dashed box). Similar operations are applied between frame t-1 and frame t.

To train the spatial temporal transformer, we optimize its parameters to minimize the photometric loss between the warped frames and the reference frame, as given in Eq. 1

$$Loss_{flow} = ||I^{'}_{t-1} - I_t||^2 + ||I^{'}_{t+1} - I_t||^2. \quad (1)$$

In the ablation study, we show that our video depth estimation network with STN motion compensation can achieve similar accuracy to the one based on state-of-the-art optical flow estimator such as PWCNet [10].

2.2. Temporal attention

Sometimes due to the inaccurate optical flow estimation or occlusion, the warped frames will have inconsistency between moving objects. Directly using these warped frames may confuse the following depth estimation network. To solve this problem, we propose the temporal attention module to emphasize the consistent information among motion compensated frames.

The architecture of our temporal attention module is illustrated in Figure 4. For better generalization, we apply the temporal attention module on the feature maps $F(I^{'}_{t-1}), F(I^{'}_t), F(I^{'}_{t+1})$ generated from the warped frames $I^{'}_{t-1}, I^{'}_t, I^{'}_{t+1}$. All of these three feature maps have same size $R^{C \times W \times H}$. We start by combining these feature maps to obtain a feature map $M \in R^{3C \times W \times H}$. Next, M is reshaped to 2D as $M_r \in R^{3C \times (WH)}$, and we perform a matrix multiplication between M_r and the transpose of M_r. A softmax layer is applied to obtain the temporal attention map $A \in R^{3C \times 3C}$

$$A_{ij} = \frac{e^{M^j_r \cdot M^i_r}}{\sum_{j=1}^{3C} e^{M^j_r \cdot M^i_r}}, \quad (2)$$

where M^i_r and M^j_r are one dimensional vectors in $R^{(WH)}$, $\cdot$ is the dot product between two vectors. A_{ij} measures the impact of the frame generating the j^{th} channel on the frame generating the i^{th} channel. If i and j come from the same frame, it measures a kind of self-attention. After softmax, we adopt a matrix multiplication between the temporal attention map A and M_r, and reshape their result to $R^{3C \times H \times W}$. The result is

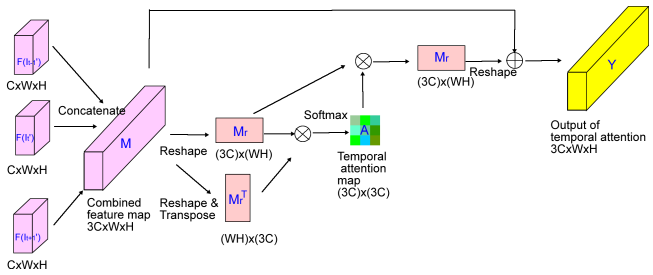

Fig. 4. Our proposed temporal attention module. The inputs are the feature maps $F(I^{'}_{t-1}), F(I^{'}_{t}), F(I^{'}_{t+1})$ generated from the motion compensated frames.

multiplied by a scale parameter η and perform element-wise sum with M to obtain the output feature maps Y, as

$$Y^i = \eta \sum_{j=1}^{3C} (A_{ij} M^j) + M^i. \quad (3)$$

where $Y^i \in R^{W \times H}$ is a single channel feature map. η is learnt during the training, and initialized as 0.

In the experiments, we show that our temporal attention module can improve the accuracy of video depth estimation.

2.3. Depth estimation

After obtaining the feature maps Y from the temporal attention module, we feed these feature maps into the depth estimation network to obtain the final estimated depth map for the center frame t D_t. We use an encoding-decoding architecture as in [11]. We uniformly quantize the depth-range into K bins in the log scale. Then the depth estimation network is actually doing a kind of classification task. We use a soft classification loss, where the expected depth is the weighted sum of the quantized depth probabilities and the quantized depth values, and is the predicted depth for each pixel location. Let p_i^j represents the probability of pixel i being quantized depth j, and B is the maximum quantized depth, the expected quantized depth D_i of pixel i is calculated as $D_i = \sum_{j=1}^{B} j \times p_i^j$. A Huber loss is utilized to measure the difference between the predicted quantized depth $D_{i,t}$ of frame t and the ground truth quantized depth $D^*_{i,t}$ of frame t, $Loss_{depth} = Huber(D_{i,t}, D^*_{i,t})$.

Our network is trained end-to-end. The motion compensation module and the depth estimation module are optimized at the same time. The final loss function is a weighted sum between the motion compensation loss in Eq. 1 and the depth estimation loss, as given in Eq. 4. We set $w_{flow} = 0.5$, and $w_{depth} = 1$ empirically.

$$Loss_{all} = w_{flow} Loss_{flow} + w_{depth} Loss_{depth}. \quad (4)$$

When training our network, the learning rate is set to 0.0001 at the first 10 epochs, and then decreases to 0.00001 for the following epochs. We use Adam to optimize the network with $\beta_1 = 0.9, \beta_2 = 0.999$.

3. EXPERIMENTS

3.1. Datasets

We use KITTI dataset [15] to evaluate our approach. We use the same train/test split (Eigen's split) as [8][6][7], meaning that we use 28 sequences for training and 28 sequences for testing in the categories of 'city', 'residential' and 'road'. Since each of the sequences are captured with a stereo pair, we have 2 monocular sequences giving 56 monocular sequences for training. Consecutive 3-frame video clips are cropped as the training example. Since the size of KITTI images varies, we zero pad them to a uniform size 1248×384 for training. We quantize the ground-truth depth from $[0m, 80m]$ to $K = 200$ bins. The depth higher than $80m$ is truncated to $80m$. During the testing, we evaluate the accuracy of all labeled pixels without any depth cap. The commonly-used evaluation metric, including mean relative absolute error (REL), mean relative squared error (sqREL), root mean square error (RMSE), and δ threshold are utilized to evaluate the performance.

3.2. Experimental results on the KITTI dataset

In Table 1, we show the accuracy of our proposed video depth estimation network on the KITTI dataset, compare to existing video depth estimation methods [8][6][7], as well as applying the state-of-the-art single image depth estimation methods [12][13][14][11] frame by frame. We tested their official models of other approaches on a single Tesla-V100 GPU and reported the accuracy and speed. It can be seen that our method achieves significantly better accuracy compared to other video depth estimation methods (rows with 'frames' as 3), where our relative error is only 1/3. The reason is that these methods are all trained by minimizing the photometric error of the original input video, without using ground-truth depth information. Such unsupervised method is clearly worse compared to our supervised method. Compared to the state-of-the-art SIDE methods (rows with 'frames' as 1) including the SIDE backbone [11] used in our framework, our method benefits from the temporal information, so that we can achieve lower RMSE and REL. We also give the accuracy of our network without the temporal attention module. The RMSE increases from 4.771m to 5.182m. These results verify the effectiveness of our proposed spatial temporal transformer and temporal attention module. Due to our efficient design of STN and temporal attention module, the efficiency is still similar to [11]. In the supplementary video [1], we show that our MVDE method also gives better depth details compared to applying SIDE method frame by frame.

[1] https://youtu.be/3B4fgfjD8J0

Table 1. Depth estimation accuracy and efficiency of different networks on the KITTI dataset. Rows with 'frames' as 1 denote that this method is single-frame based. The accuracy of other papers are generated by their official models.

Method	frames	REL (%)	sqREL	RMSE (in meter)	δ_1	δ_2	δ_3	speed (in second)
Eigen et al. [12]	1	20.34	1.548	6.307	0.702	0.789	0.905	0.864
Godard et al. [13]	1	14.86	1.344	5.927	0.803	0.922	0.964	0.578
Ren et al. [11]	1	6.64	0.271	5.156	0.935	0.979	0.992	0.348
Fu et al. [14]	1	7.88	0.294	4.991	0.931	0.976	0.992	0.748
Zhou et al. [6]	3	20.81	1.768	6.856	0.678	0.885	0.957	1.367
Wang et al. [7]	3	14.82	1.187	5.496	0.812	0.938	0.975	1.478
Ariel [8]	3	12.80	0.959	5.230	N/A	N/A	N/A	N/A
Ours w/o temporal attention	3	4.33	0.241	5.182	0.939	0.982	0.993	0.362
Ours	3	**3.78**	**0.209**	**4.771**	**0.952**	**0.993**	**0.998**	0.370

Table 2. Depth estimation accuracy of our framework with different optical flow estimation modules on KITTI dataset.

Method	optical flow	REL (%)	RMSE	speed
Ours	STN	3.78	4.771	0.370
Ours	PWCNet	3.74	4.754	0.688

Table 3. Depth estimation accuracy of our networks with different number of input frames on KITTI dataset.

Method	attention	frames	REL (%)	RMSE
Ours	N/A	3	4.33	5.182
Ours	Yes	3	3.78	4.771
Ours	N/A	5	4.25	5.147
Ours	Yes	5	**3.76**	**4.641**

3.3. Ablation study

In this section, we give ablation study of the key modules of our video depth estimation framework. We first compare our spatial temporal transformer to the state-of-the-art optical flow estimation methods such as PWCNet [10]. In Table 2, it can be seen that the depth estimation accuracy of our network based on STN is only slightly worse than the one based on PWCNet. The reason is that our video depth estimation network is trained end-to-end with STN. Although the accuracy of the learnt optical flow is lower, it can be compensated by the following temporal attention module and the depth estimation module, with a supervised depth estimation loss. In addition, our STN only consists of 10 convolutional layers, its network size is much smaller than PWCNet. If we replace STN with PWCNet, the network size and computational cost will significantly increase.

Next, we evaluate the performance of our framework with more time stamps. Intuitively, using more time stamps will always be helpful for video-based vision applications. In Table 3, we can find that using 5 frames is consistently better than using three frames with same temporal attention module. This result is consistent to our intuition: the more temporal information we use, the better accuracy we can get.

In Fig. 5, we visualize the difference map between the input and output of the temporal attention module. For easier visualization, we apply the temporal attention module on warped RGB frames, so there will be three difference maps for R/G/B channels respectively. The larger value, the more difference (emphasized or suppressed) between the input and

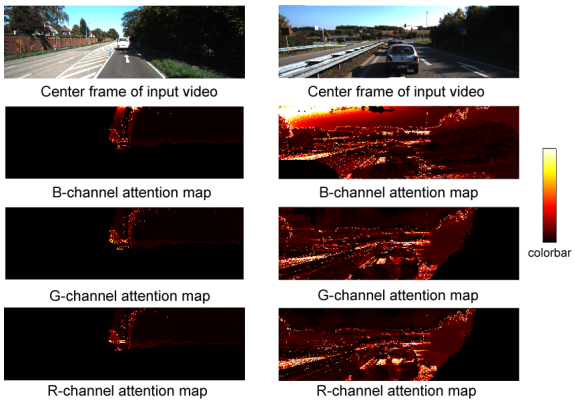

Fig. 5. Differences between the input and output of temporal attention module.

output of the temporal attention module. For the left image of Fig. 5, we observe that the attention focuses on the car since the car is the most important moving object. For the right image, the attention focus on all major regions with motion. Compared to the left image, the illumination of the right image is more complicated (see the shadow), and the objects are closer to the camera. So the temporal consistency is more complicated. These results make sense since optical flow based warping can handle the case of 2D motion (w/o occlusion), but is relatively difficult to deal with the motion in the depth direction. In this scenario, our proposed temporal attention can give us more insights.

4. CONCLUSION

In this paper, we proposed an end-to-end video depth estimation network. A motion compensation module based on spatial temporal transformer network is first applied to estimate the optical flow without ground-truth optical flow supervision. The input frames are warped based on the estimated flow, and further combined by the temporal attention module to enhance the temporal consistency. An encoding-decoding network is further utilized to estimate the depth. Compared to state-of-the-art methods, our MVDE with temporal attention achieve better performance on the benchmark KITTI dataset.

5. REFERENCES

[1] David Eigen and Rob 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*, 2015, pp. 2650–2658.

[2] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 239–248.

[3] Kevin Karsch, Ce Liu, and Sing Bing Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2144–2158, 2014.

[4] Raul Dıaz and Amit Marathe, "Soft labels for ordinal regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[6] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[7] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey, "Learning depth from monocular videos using direct methods," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2022–2030.

[8] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *arXiv 1904.04998*, 2019.

[9] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[10] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[11] Haoyu Ren, Mostafa El-khamy, and Jungwon Lee, "Deep robust single image depth estimation neural network using scene understanding," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Worlshop*, 2019.

[12] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.