Unsupervised Learning of Depth and Ego-Motion from Monocular Video in 3D

Reza Mahjourian[1], Martin Wicke and Anelia Angelova[2]

***Abstract*— We present a novel approach for unsupervised learning of depth and ego-motion from monocular video. The idea is to explicitly consider the inferred 3D geometry of the scene, enforcing consistency of the estimated 3D point clouds and ego-motion across consecutive frames. This is a challenging task and is solved by a novel (approximate) backpropagation algorithm for aligning 3D structures. Because we only require a simple video, learning depth and ego-motion on large and varied datasets becomes possible. We demonstrate this by training on the low quality uncalibrated video dataset and evaluating on KITTI, ranking among top performing prior methods which are trained on KITTI itself. A detailed version of this abstract will appear in [5]. See also http://sites.google.com/view/vid2depth.**

Introduction. This paper considers unsupervised learning of depth and ego-motion from monocular RGB-only videos [6]. The only form of supervision that we use comes from assumptions about consistency and temporal coherence between consecutive frames in a monocular video. While, most supervised methods for learning depth and ego-motion require carefully calibrated setups [4], here the only required input is video and the camera focal length.

Method. In order to learn depth in a completely unsupervised fashion, we rely on existence of ego-motion in the video. Given two consecutive frames from the video, a neural network produces single-view depth estimates from each frame, and an ego-motion estimate from the frame pair. Requiring that the depth and ego-motion estimates from adjacent frames are consistent serves as supervision for training the model. We propose a loss which directly penalizes inconsistencies in the estimated depth by directly comparing 3D point clouds in a common reference frame. Intuitively, assuming there is no significant object motion in the scene, one can transform the estimated point cloud for each frame into the predicted point cloud for the other frame by applying ego-motion or its inverse (Fig. 1). Unlike 2D losses that enforce local photometric consistency, the 3D loss considers the entire scene and its geometry. We show how to efficiently backpropagate through this loss [5].

Problem setup. At training time, the goal is to learn depth and ego-motion from a single monocular video stream. This problem can be formalized as follows: Given a pair of consecutive frames X_{t-1} and X_t, estimate depth D_{t-1} at time $t-1$, depth D_t at time t, and the ego-motion T_t representing the camera's movement (position and orientation) from time $t-1$ to t. Once a depth estimate D_t is available, it can be projected into a point cloud Q_t (using the camera intrinsic matrix K).

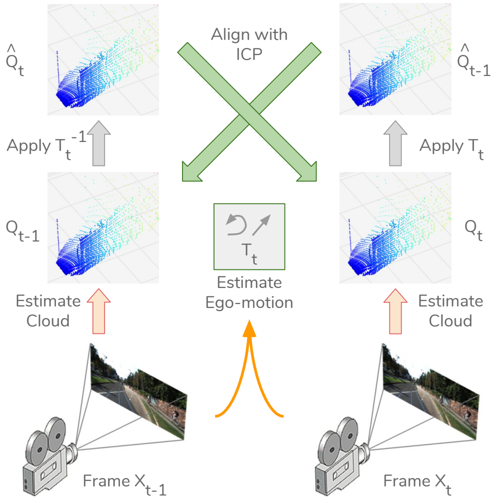

Fig. 1. The 3D loss: ICP is applied symmetrically in forward and backward directions to bring the depth and ego-motion estimates from two consecutive frames into agreement. The products of ICP generate gradients which are used to improve the depth and ego-motion estimates.

Given an estimate for T_t, the camera's movement from $t-1$ to t, Q_t can be transformed to get an estimate for the previous frame's point cloud: $\hat{Q}_{t-1} = T_t Q_t$. Note that the transformation applied to the point cloud is the inverse of the camera movement from t to $t-1$. $\hat{Q}_{t-1}$ can then be projected onto the camera at frame $t-1$ as $K\hat{Q}_{t-1}$. Combining this transformation and the projection onto the image, establishes a mapping from image coordinates at time t to image coordinates at time $t-1$. This mapping allows us to reconstruct frame $\hat{X}_t$ by warping X_{t-1} based on D_t, T_t. Comparing the reconstructed images $\hat{X}_t$, $\hat{X}_{t-1}$ to the input frames X_t, X_{t-1} respectively produces a differentiable image reconstruction loss that is based on photometric consistency per pixel i, j [6], [2] $L_{\text{rec}} = \sum_{ij} \|(X_t^{ij} - \hat{X}_t^{ij})\|$.

3D Point Cloud Alignment Loss. Instead of using the projections of $\hat{Q}_{t-1}$ or $\hat{Q}_t$ just to establish a mapping between coordinates of adjacent frames, we construct a loss function that directly compares point clouds $\hat{Q}_{t-1}$ to Q_{t-1}, or $\hat{Q}_t$ to Q_t. This 3D loss uses a well-known rigid registration method, Iterative Closest Point (ICP) [1], which computes a transformation that minimizes point-to-point distances between corresponding points in the two point clouds.

Because of the combinatorial nature of the correspondence computation, ICP is not differentiable. We can, however, approximate its gradients using the products of the alignment, allowing us to backpropagate errors for both the ego-motion

*This work was done while at Google Brain

[1] Reza Mahjourian is also with the University of Texas at Austin

[2] Contact author is `anelia@google.com`

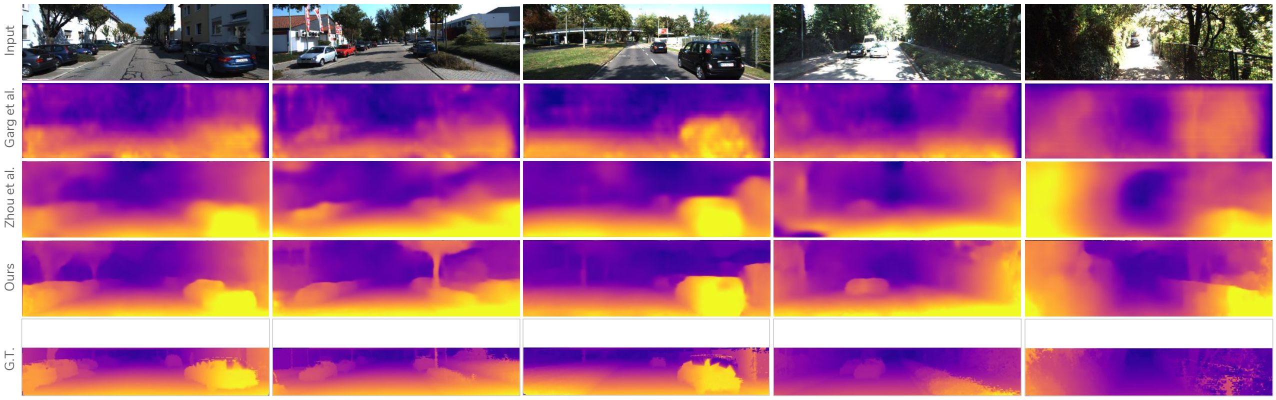

Fig. 2. Sample depth estimates from the KITTI Eigen test set, generated by our approach (4th row), compared to prior work [2], [6], and ground truth

Method	Superv.	Data	Abs Rel	Sq Rel	RMSE
Eigen et al.	Depth	K	0.203	1.548	6.307
Liu et al.	Depth	K	0.201	1.584	6.471
Ours (Bike)	-	Bike	0.211	1.771	7.741
Zhou et al.[6]	-	K	0.208	1.768	6.856
Zhou et al.[6]	-	CS+K	0.198	1.836	6.565
Ours	-	K	0.163	1.240	6.220
Ours	-	CS+K	**0.159**	**1.231**	**5.912**

TABLE I

DEPTH EVALUATION METRICS. KITTI TEST SET.

and depth estimates. ICP takes as input two point clouds A and B (*e. g.* $\hat{Q}_{t-1}$ and Q_{t-1}). Its main output is a best-fit transformation T' which minimizes the distance between the transformed points in A and their corresponding points in B: $\arg\min_{T'} \frac{1}{2} \sum_{ij} \|T' \cdot A^{ij} - B^{c(ij)}\|^2$, where $c(\cdot)$ denotes the point to point correspondence found by ICP. The secondary output of ICP is the residual $r^{ij} = A^{ij} - T'^{-1} \cdot B^{c(ij)}$, which reflects the residual distances between corresponding points after ICP's distance minimizing transform has been applied. The negative residuals are used as (approximate) gradients to improve depth per step, and T' as potential improvement on ego-motion estimation.

Overall, the full training loss is defined as:

$$L = \sum_s \alpha L^s_{\text{rec}} + \beta L^s_{\text{3D}} + \gamma L^s_{\text{sm}} + \omega L^s_{\text{SSIM}} \qquad (1)$$

where L^s_{rec} is the standard photometric reconstruction loss [6], L^s_{3D} is the ICP-based loss, L^s_{sm} and L^s_{SSIM} are the smoothness and the structural similarity (SSIM) losses, respectively, combined with hyper-parameters. The loss is applied at four scales s.

Experiments. We first test our algorithm on the popular KITTI dataset [3]. Our proposed approach consistently improves depth estimates, and outperforms the state-of-the-art for both depth (Table I, Fig. 2) and ego-motion. Our model lowers the mean absolute relative depth prediction error (in meters) from 0.208 [6] to 0.163, which is a significant improvement. Our method obtains ego-motion trajectory error of 0.013 ± 0.010 and significantly outperforms the method of [6] 0.021 ± 0.017, as well as other state-of-the-art methods.

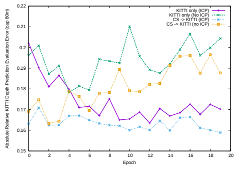

Fig. 3. Depth eval. error over time.

Evaluation of the 3D Loss. Fig. 3 plots the evaluation error from each model over time as training progresses. The points show the depth error at different training epochs on the evaluation set. Using the 3D loss improves performance across all stages of training and has a regularizing effect, by reducing overfitting. In contrast, just pre-training on the larger Cityscapes dataset is not sufficient to reduce overfitting or improve depth quality.

Learning from an uncalibrated video stream. We demonstrate that our proposed approach can consume and learn from any monocular video source with camera motion. We record a new dataset, the Bike Video Dataset, containing monocular video captured using a hand-held commercial phone camera while riding a bicycle. We train our depth and ego-motion model only on these videos, then evaluate the quality of its predictions by testing the trained model on the KITTI dataset. The results in Table I show that our model trained on the Bike videos is close in quality to the best unsupervised model of [6], which is trained on KITTI itself.

Acknowledgments We thank Tinghui Zhou, Clément Godard, Parvin Taheri and Oscar Ejdeha.

REFERENCES

[1] P. J. Besl and N.D. McKay, A Method for Registration of 3-D Shapes, IEEE Trans. on Pattern Analysis and Machine Intelligence, 1992

[2] R. Garg, G. Carneiro and I. Reid, Unsupervised CNN for single view depth estimation: Geometry to the rescue, ECCV 2016

[3] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, Vision meets robotics: The KITTI dataset, The International Journal of Robotics Research, 2013

[4] C. Godard, O. Aodha and G. Brostow, Unsupervised Monocular Depth Estimation with Left-Right Consistency, CVPR 2017

[5] R. Mahjourian, M. Wicke and A. Angelova, Unsupervised, Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints, CVPR, 2018

[6] T. Zhou, M. Brown, N. Snavely and D. Lowe, Unsupervised Learning of Depth and Ego-Motion from Video, CVPR, 2017