Instance-wise Depth and Motion Learning from Monocular Videos

Seokju Lee[1] **Sunghoon Im**[2] **Stephen Lin**[3] **In So Kweon**[1]

KAIST[1] DGIST[2] Microsoft Research Asia[3]
seokju91@gmail.com

Abstract

We present an end-to-end joint training framework that explicitly models 6-DoF motion of multiple dynamic objects, ego-motion and depth in a monocular camera setup without supervision. Our technical contributions are two-fold. First, we highlight the fundamental difference between inverse and forward projection while modeling an individual motion of rigid object, and propose a geometrically correct projection pipeline using a differentiable forward projection module. Second, we design a unified instance-aware photometric and geometric consistency loss that holistically imposes self-supervisory signals for every background and object region. Through extensive experiments conducted on the KITTI and Cityscapes dataset, our framework is shown to outperform the state-of-the-art depth and motion estimation methods.

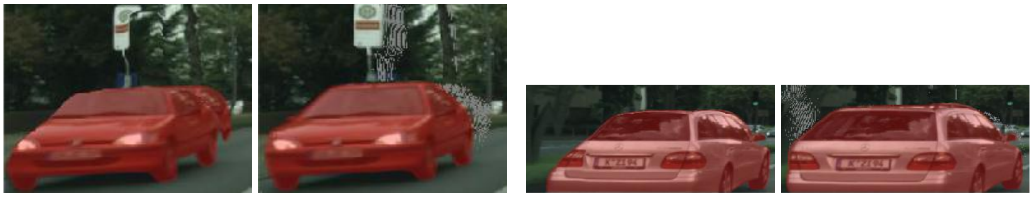
1 Introduction

Recent advances in deep neural networks (DNNs) have led to a surge of interest in depth prediction using monocular images [8, 9] and stereo images [22, 5]. These supervised methods require a large amount and broad variety of training data with ground-truth labels. Studies have shown significant progress in unsupervised learning of depth and ego-motion from unlabeled image sequences [29, 10, 25, 21, 24]. The joint optimization framework uses a network for predicting single-view depth and pose, and exploits view synthesis of images in the sequence as the supervisory signal. However, these works ignore or mask out regions of moving objects for pose and depth inference.

In this work, rather than consider moving object regions as a nuisance under the ***assumption of static structure***, we utilize them as an important clue for estimating 3D object motions. Factorizing camera and object motion in monocular sequences is a challenging problem, especially in complex urban environments that contain plenty of dynamic objects. To address this problem, we propose a novel framework that explicitly models 3D motions of dynamic objects and ego-motion together with scene depth in a monocular camera setting. Our unsupervised method relies solely on monocular video for training (without any ground-truth labels) and imposes a unified photometric and geometric consistency loss on synthesized frames from one time step to the next in a sequence. Given two consecutive frames from a video, the proposed neural network produces depth, 6-DoF motion of each moving object, and the ego-motion between adjacent frames. In this process, we leverage the instance mask of each dynamic object, obtained from off-the-shelf instance segmentation and optical flow modules. Our main contributions are the following:

Forward image projection Differentiable depth-based rendering was introduced in [29], where the target view I_t is reconstructed by sampling pixels from a source view I_s based on the target depth map D_t and the relative pose $T_{t\to s}$. The warping procedure is effective in static scene areas, but the regions of moving objects cause warping artifacts because the 3D structure of the source image I_s may become distorted after warping based on the target image's depth D_t [3] as shown in Fig. 1(a).

Machine Learning for Autonomous Driving Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a) Inverse projection [3] and forward projection. (b) Reversed warping [27] and forward projection.

Figure 1: Different rendering techniques on dynamic objects. Inverse projection and reversed inverse warping cause significant appearance distortion and ghosting effect, while our forward projection technique preserves the object appearance. This is a *video figure*, best viewed in *Adobe Reader*.

To build a geometrically plausible formulation, we introduce forward warping (or projection) which maps the source image I_s to the target viewpoint based on the source depth D_s and the relative pose $T_{s\rightarrow t}$. [1] There is a well-known remaining issue with forward warping that the output image may have holes. Thus, we propose the differentiable and hole-free forward warping module that works as a key component in our instance-wise depth and motion learning from monocular videos.

Unified photometric and geometric consistency Existing works [18, 2] have successfully estimated independent object motion with stereo cameras. On the other hand, estimation from monocular video captured in the dynamic real world, where both agents and objects are moving, suffers from *motion ambiguity*, as only temporal clues are available. To address this issue, we introduce instance-aware view synthesis and unified projection consistency into the training loss. While warping each component, we impose a geo-consistency loss as well as a photo-consistency for each instance that constrains the estimated geometry from all input frames to be consistent. The proposed learning framework shows the state-of-the-art performance on monocular depth and motion estimation.

2 Related Works

Unsupervised depth and ego-motion learning Several works [29, 25, 21, 24, 23, 13] have studied a joint self-supervised learning of depth and ego-motion from monocular sequences with a basic concept of *Structure-from-Motion (SfM)*. Along with the photo-consistency proposed by Zhou *et al*. [29], several works [21, 1, 7] impose geometric constraints between nearby frames with a static structural assumption. Semantic knowledge is also used to enhance the feature representation for monocular depth estimation [6, 14]. **Our novelty**–The aforementioned studies have a limitation on dealing with moving objects due to the rigidity assumption, which leads to performance degradation in estimating object depths. To handle this, stereo pairs are leveraged during the training process as an auxiliary as presented by Godard *et al*. [10] and Hur *et al*. [15]. Please note that the monocular-based approaches are differentiated from the methodology of learning through stereo videos.

Learning object motion Recently, the joint optimization of dynamic object motion along with depth and ego-motion has gained interest as a new research topic. Cao *et al*. [2] propose a self-supervised framework with a given 2D bounding box to learn scene structure and 3D object motion from *stereo* videos. The disparity from the paired images, which is *deterministic*, enables computing the 3D motion vector of each instance using simple mean filtering. Gordon *et al*. [12] propose a motion field network to estimate a pixel-wise transformation. It receives two consecutive rough images, which are, however, ambiguous and unclear inputs to explicitly disentangle the motion of camera and non-rigid objects. Hence, we suggest to design the network to determine the object motion by looking at the residual signal between two images caused by pure object motion. Casser *et al*. [3, 4] and Klingner *et al*. [17] present an unsupervised image-to-depth framework that models the motion of moving objects and cameras with given segmentation knowledge. **Our novelty**–The aforementioned studies use the inverse warping technique when rendering dynamic objects, which causes appearance distortion, presented in Fig. 1. Thus, we propose a *geometrically correct* projection method in dynamic situations, which is a fundamental problem in 3D geometry.

[1] This is different from the reversed optical flow leveraged in [19, 26, 20]. Since flow-based warping techniques do not consider a geometric structure, serious distortions will appear where multiple source pixels warped to the same target locations, *e.g*., object boundaries, as shown in Fig. 1(b). Our *forward* and *inverse warping* is not about temporal order, but rather which coordinate frame from which to conduct the geometric transformation when warping from the reference to the target view.

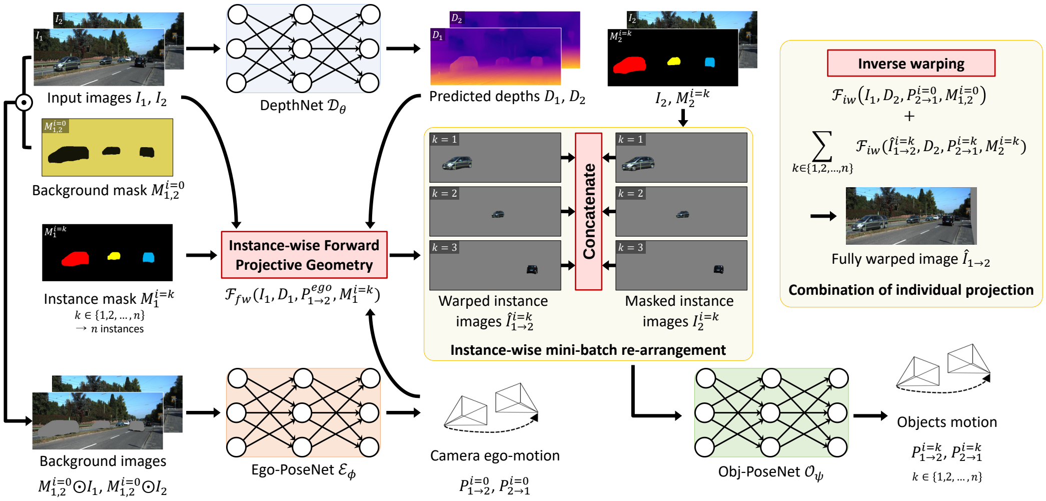

I_2
I_1
Input images I_1, I_2
$M_{1,2}^{i=0}$
Background mask $M_{1,2}^{i=0}$
$M_1^{i=k}$
Instance mask $M_1^{i=k}$
$k \in \{1,2,\dots,n\}$
$\rightarrow n$ instances
Background images
$M_{1,2}^{i=0} \odot I_1$, $M_{1,2}^{i=0} \odot I_2$
DepthNet $\mathcal{D}_\theta$
Instance-wise Forward Projective Geometry
$\mathcal{F}_{fw}(I_1, D_1, P_{1\to2}^{ego}, M_1^{i=k})$
Ego-PoseNet $\mathcal{E}_\phi$
D_2
D_1
Predicted depths D_1, D_2
Camera ego-motion
$P_{1\to2}^{i=0}$, $P_{2\to1}^{i=0}$
$M_2^{i=k}$
I_2, $M_2^{i=k}$
$k = 1$
$k = 2$
$k = 3$
Concatenate
Warped instance images $\hat{I}_{1\to2}^{i=k}$
Masked instance images $I_2^{i=k}$
Instance-wise mini-batch re-arrangement
Obj-PoseNet $\mathcal{O}_\psi$
Objects motion
$P_{1\to2}^{i=k}$, $P_{2\to1}^{i=k}$
$k \in \{1,2,\dots,n\}$
Inverse warping
$\mathcal{F}_{iw}(I_1, D_2, P_{2\to1}^{i=0}, M_{1,2}^{i=0})$
$+$
$\sum_{k\in\{1,2,\dots,n\}} \mathcal{F}_{iw}(\hat{I}_{1\to2}^{i=k}, D_2, P_{2\to1}^{i=k}, M_2^{i=k})$
Fully warped image $\hat{I}_{1\to2}$
Combination of individual projection

Figure 2: Overview of the proposed frameworks.

3 Methodology

We introduce an end-to-end joint training framework for instance-wise depth and motion learning from monocular videos without geometric annotation as illustrated in Fig. 2. Our main contribution lies in applying the inverse and forward warping in appropriate projection situations. In Table 1, we describe the difference between them. Following their characteristics, we propose a geometrically correct warping method in dynamic situations, which is a fundamental problem in 3D geometry.

Table 1: Comparisons between inverse and forward warping.

	Inverse warping	Forward warping
Inputs	I_1, D_2, $P_{2\to1}^{i=0}$ (inverse ego-motion)	I_1, D_1, $P_{1\to2}^{i=0}$ (forward ego-motion)
Pros.	Dense registration by grid sampling (suitable for static region).	Geometry corresponds to reference (suitable for dynamic region).
Cons.	Errors induced on moving objects.	Holes are generated.

Modeling object motion via forward projective geometry We model the geometry of the moving object with a two-stage warping procedure. In the first step, we define an *intermediate frame* which is transformed by camera motion with reference geometry. The *intermediate frame* is reconstructed by forward projective geometry, $\mathcal{F}_{fw}(I_1, D_1, P_{1\to2}^{i=0}, K) \rightarrow \hat{I}_{1\to2}^{fw}$ as follows: $p_2 \sim KP_{1\to2}^{i=0} D_1^{\uparrow}(p_1)(K^{\uparrow})^{-1} p_1$. The forward warping cannot be interpolated by the *grid sampling* [16] since it is a rasterization procedure (inverse of *grid sampling*). In order to make this operation differentiable, we use sparse tensor coding to index the homogeneous coordinates p_2 of a pixel in I_2. Invalid coordinates (exiting the view where $p_2 \notin \{(x,y) | 0 \leq x < W, 0 \leq y < H\}$) of the sparse tensor are masked out. We then convert this sparse tensor to be dense by taking the nearest neighbor value of the source pixel. However, irregular holes are generated during the sparse coding. Since we need to feed those forward projected images into the neural networks in the next step, the size of the holes should be minimized. To fill these holes as much as possible, we pre-upsample the depth map $D_1^{\uparrow}(p_1)$ of the reference frame by a factor of α. The camera intrinsic matrix, $K^{\uparrow}$, is also upsampled by multiplying α to the focal length and principal point. For the second step, the pure object motion is obtained by using this *intermediate frame* and pre-computed instance knowledge as inputs of Obj-PoseNet. The final warped view is reconstructed by inverse warping with each composite motion of an individual instance.

Unified instance-aware projection consistency Basically, the proposed DepthNet, Ego-PoseNet, and Obj-PoseNet are jointly optimized together with a self-supervisory signal of an image reconstruction loss, $\mathcal{L}_{rec} = ||I_2 - \hat{I}_{1\to2}||_1$. Along with this photo-consistency, we propose an instance-wise geometric consistency loss. With the predicted depth, composite motion, and instance prior, we warp the reference depth map of each object to the target frame ($\hat{D}_{1\to2}^{iw,i=k}$) and transform the target depth map to the reference frame ($\hat{D}_{2\to1}^{sc,i=k}$). The depth inconsistency map is designed as the difference between $\hat{D}_{1\to2}^{iw,i=k}$ and $\hat{D}_{2\to1}^{sc,i=k}$. We optimize this term for boosting the geometric consistency between nearby frames, in addition to the photometric consistency.

Table 2: Ablation study on KITTI Eigen split for both background (bg.) and object (obj.) areas.

Instance knowledge	Geometric consistency	Object warping		AbsRel		
		inverse	forward	all	bg.	obj.
✗	✗	✗	✗	0.156	0.142	0.396
✗	✓	✗	✗	0.137	0.124	0.309
✓	✗	✓	✗	0.151	0.138	0.377
✓	✓	✓	✗	0.146	0.131	0.362
✓	✗	✗	✓	0.143	0.133	0.285
✓	✓	✗	✓	**0.124**	**0.119**	**0.178**

Table 3: Evaluation on KITTI 2015 scene flow training set. We evaluate the disparity compared to recent monocular training methods.

Method	D1			D2		
	bg.	fg.	all	bg.	fg.	all
CC [24]	35.0	42.7	36.2	–	–	–
SC-SfM [1]	36.0	46.5	37.5	–	–	–
EPC++ (mono) [20]	30.7	34.4	32.7	**18.4**	84.6	65.6
Ours + GeoNet [28]	**26.8**	**30.4**	**27.4**	28.9	**32.3**	**29.4**

Table 4: Monocular depth estimation results on the KITTI (K) Eigen test split and Cityscapes (C) test set. Models pretrained on Cityscapes and fine-tuned on KITTI are denoted by ‘C+K’. Models trained with semantic knowledge are denoted by ’S’. For each partition, best results are written in **boldface**.

Method	Backbone	Training	Test	Error metric ↓				Accuracy metric ↑		
				AbsRel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
EPC++ [20]	DispNet	K	K	0.141	1.029	5.350	0.216	0.816	0.941	0.976
CC [24]	DispResNet	K	K	0.140	1.070	5.326	0.217	0.826	0.941	0.975
SC-SfM [1]	DispResNet	K	K	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Ours	DispResNet	K (S)	K	**0.124**	**0.886**	**5.061**	**0.206**	**0.844**	**0.948**	**0.979**
GLNet [7]	ResNet18	K	K	0.135	1.070	5.230	0.210	0.841	0.948	**0.980**
Monodepth2 [11]	ResNet18	K	K	0.132	1.044	5.142	0.210	0.845	0.948	0.977
Struct2Depth [3]	ResNet18	K (S)	K	0.141	1.026	5.290	0.215	0.816	0.945	0.979
Ariel *et al.* [12]	ResNet18	K (S)	K	0.128	0.959	5.230	0.212	0.845	0.947	0.976
Ours	ResNet18	K (S)	K	**0.112**	**0.855**	**4.690**	**0.192**	**0.874**	**0.958**	**0.980**
CC [24]	DispResNet	C+K	K	0.139	1.032	5.199	0.213	0.827	0.943	0.977
SC-SfM [1]	DispResNet	C+K	K	0.128	1.047	5.234	0.208	0.846	0.947	0.976
Ours	DispResNet	C+K (S)	K	**0.119**	**0.863**	**4.984**	**0.202**	**0.856**	**0.950**	**0.980**
Ariel *et al.* [12]	ResNet18	C+K (S)	K	0.124	0.930	5.120	0.206	0.851	0.950	0.978
Ours	ResNet18	C+K (S)	K	**0.109**	**0.812**	**4.623**	**0.191**	**0.875**	**0.958**	**0.979**
Struct2Depth [4]	ResNet18	C (S)	C	0.145	1.737	7.280	0.205	0.813	0.942	0.978
Ours	ResNet18	C (S)	C	**0.128**	**1.584**	**5.917**	**0.197**	**0.852**	**0.951**	**0.980**

Table 5: Absolute trajectory error (ATE) on KITTI-VO.

Method	Seq. 09	Seq. 10
SfM-Learner [29]	0.021 ± 0.017	0.020 ± 0.015
GeoNet [28]	0.012 ± 0.007	0.012 ± 0.009
CC [24]	0.012 ± 0.007	0.012 ± 0.008
Struct2Depth [3]	0.011 ± 0.006	0.011 ± 0.010
GLNet [7]	0.011 ± 0.006	0.011 ± 0.009
Ours (w/o inst.)	0.012 ± 0.008	0.011 ± 0.010
Ours (w/ inst.)	$\mathbf{0.010 \pm 0.013}$	$\mathbf{0.011 \pm 0.008}$

Table 6: Relative translation t_{err} (%) and rotation r_{err} ($^{\circ}/100m$) errors on KITTI-VO.

Method	Seq. 09		Seq. 10	
	t_{err}	r_{err}	t_{err}	r_{err}
GeoNet [28]	39.4	14.3	29.0	8.6
SC-SfM [1]	11.2	3.4	10.1	5.0
Ours (w/o inst.)	10.2	5.2	10.1	4.8
Ours (w/ inst.)	**8.6**	**2.9**	**9.2**	**4.5**

4 Experiments

Implementation details For DepthNet, we use DispResNet [24] and ResNet18-based encoder-decoder structure. The structures of Ego-PoseNet and Obj-PoseNet are the same, but the weights are not shared. They consist of seven convolutional layers and regress the relative pose as three Euler angles and three translation vectors. The image resolution is set to 832×256 and the video data is augmented with random scaling, cropping, and horizontal flipping.

Monocular depth and scene flow estimation First, we conduct an ablation study to validate the effect of our forward projective geometry and instance-wise geometric consistency term on monocular depth estimation as shown in Table 2. The inverse warping on the objects slightly improves the depth estimation; however, we observe that Obj-PoseNet does not converge well, while the forward warping on the objects improves the depth estimation on both background and object areas. This shows that well-optimized Obj-PoseNet helps to boost the performance of DepthNet and they complement each other. The significant performance improvement comes from the instance-wise geometric loss incorporated with forward projection while warping the object areas. Second, we validate the disparity results on the KITTI 2015 scene flow training set as in Table 3. The foreground (fg.) results show the superiority on handling dynamic regions. Third, Table 4 shows the Eigen split test results, where ours achieves state-of-the-art performance in the single-view depth prediction task with unsupervised monocular training. The advantage is evident from using instance masks and constraining the instance-wise photometric and geometric consistencies.

Visual odometry We evaluate the performance of our Ego-PoseNet on the KITTI visual odometry (-VO) dataset. Following the evaluation setup of SfM-Learner [29], we use sequences 00-08 for training, and sequences 09 and 10 for tests. We test the performance of visual odometry under two conditions: with and without instance masks. Table 5 shows the results of absolute trajectory error (ATE), and Table 6 shows the results of relative errors (t_{err}, r_{err}). Both experiments show state-of-the-art performance.

5 Conclusion

In this work, we propose a unified framework that predicts monocular depth, ego-motion, and 6-DoF motion of multiple objects by training monocular videos. Our main contributions are (1) differentiable forward warping, and (2) unified instance-wise projection consistency loss. We show that our method outperforms the existing unsupervised methods of monocular depth and motion estimation.

References

[1] J.-W. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NeurIPS*, 2019.

[2] Z. Cao, A. Kar, C. Hane, and J. Malik. Learning independent object motion from unlabelled stereoscopic videos. In *CVPR*, 2019.

[3] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI*, 2019.

[4] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *CVPRw*, 2019.

[5] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[6] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *CVPR*, 2019.

[7] Y. Chen, C. Schmid, and C. Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, 2019.

[8] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

[9] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[10] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[11] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019.

[12] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *ICCV*, 2019.

[13] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020.

[14] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *ICLR*, 2020.

[15] J. Hur and S. Roth. Self-supervised monocular scene flow estimation. In *CVPR*, 2020.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[17] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *ECCV*, 2020.

[18] P. Liu, I. King, M. R. Lyu, and J. Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *CVPR*, 2020.

[19] P. Liu, M. Lyu, I. King, and J. Xu. Selflow: Self-supervised learning of optical flow. In *CVPR*, 2019.

[20] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019.

[21] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018.

[22] 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 *CVPR*, 2016.

[23] S. Pillai, R. Ambruş, and A. Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *ICRA*, 2019.

[24] A. Ranjan, V. Jampani, K. Kim, D. Sun, J. Wulff, and M. J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019.

[25] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018.

[26] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *CVPR*, 2019.

[27] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu. Occlusion aware unsupervised learning of optical flow. In *CVPR*, 2018.

[28] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018.

[29] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.