Instance-wise Depth and Motion Learning from Monocular Videos

Seokju Lee1  Sunghoon Im2  Stephen Lin3  In So Kweon1
KAIST1  DGIST2  Microsoft Research Asia3
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 \( T_{t \rightarrow s} \). 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 \( D_t \) as shown in Fig. 1(a).
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 $T_{s\rightarrow t}$. 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][17] 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.

1This 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.
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.

<table>
<thead>
<tr>
<th></th>
<th>Inverse warping</th>
<th>Forward warping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inputs</td>
<td>$I_1, D_1, P^i_{1 \rightarrow 2}$ (inverse ego-motion)</td>
<td>$I_1, D_1, P^i_{1 \rightarrow 2}$ (forward ego-motion)</td>
</tr>
<tr>
<td>Pros.</td>
<td>Dense registration by grid sampling (suitable for static regions)</td>
<td>Geometry corresponds to reference (suitable for dynamic regions)</td>
</tr>
<tr>
<td>Cons.</td>
<td>Errors induced on moving objects.</td>
<td>Holes are generated.</td>
</tr>
</tbody>
</table>

**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, $\mathcal{F}_{fw}(I_1, D_1, P^i_{1 \rightarrow 2}, K) \rightarrow \tilde{I}^w_{1 \rightarrow 2}$ as follows:

$$p_2 \sim KP^{-1}_{1 \rightarrow 2}D^i_1(p_1)(K^+)^{-1}p_1.$$  

The forward warping cannot be interpolated by the grid sampling (inverse of grid sampling) since it is a rasterization procedure. 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 (outside 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 $D^i_1(p_1)$ of the reference frame by a factor of $\alpha$. The camera intrinsic matrix, $K^+$, is also upsampled by multiplying $\alpha$ 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, $\mathcal{L}_{rec} = ||I_2 - \tilde{I}^i_{1 \rightarrow 2}||_1$. 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 ($\tilde{D}^i_{1 \rightarrow 2} = \tilde{D}^i_{1 \rightarrow 2}$) and transform the target depth map to the reference frame ($\hat{D}_{2 \rightarrow 1}^{i,c} = \hat{D}_{2 \rightarrow 1}^{i,c}$). The depth inconsistency map is designed as the difference between $\tilde{D}^i_{1 \rightarrow 2} = \tilde{D}^i_{1 \rightarrow 2}$ and $\hat{D}_{2 \rightarrow 1}^{i,c} = \hat{D}_{2 \rightarrow 1}^{i,c}$. We optimize this term for boosting the geometric consistency between nearby frames, in addition to the photometric consistency.
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. We validate the disparity comparison to recent monocular training methods.

**4 Experiments**

**Implementation details** For DepthNet, we use DispResNet [24] and ResNet 18-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 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 partition, best results are written in boldface.

**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