

# Self-supervised Object Motion and Depth Estimation from Video

Qi Dai<sup>1,3</sup> Vaishakh Patil<sup>1</sup> Simon Hecker<sup>1</sup> Dengxin Dai<sup>1</sup> Luc Van Gool<sup>1,2</sup> Konrad Schindler<sup>1,3</sup>

<sup>1</sup>Computer Vision Lab, ETH Zurich <sup>2</sup>VISICS, ESAT/PSI, KU Leuven

<sup>3</sup>Institute of Geodesy and Photogrammetry, ETH Zurich

daiq@ethz.ch {patil, heckers, dai, vangool}@vision.ee.ethz.ch schindler@geod.baug.ethz.ch

## Abstract

We present a self-supervised learning framework to estimate the individual object motion and monocular depth from video. We model the object motion as a 6 degree-of-freedom rigid-body transformation. The instance segmentation mask is leveraged to introduce the information of object. Compared with methods which predict dense optical flow map to model the motion, our approach significantly reduces the number of values to be estimated. Our system eliminates the scale ambiguity of motion prediction through imposing a novel geometric constraint loss term. Experiments on KITTI driving dataset demonstrate our system is capable to capture the object motion without external annotation. Our system outperforms previous self-supervised approaches in terms of 3D scene flow prediction, and contribute to the disparity prediction in dynamic area.

## 1. INTRODUCTION

Imagining a driving scenario in real world. The driver may encounter many dynamic objects (*e.g.* moving vehicles). The knowledge of their movements is of vital importance for the driving safety. We aim to solve the motion of individual object from video in the context of autonomous driving (*i.e.* the video is taken by a camera installed on a moving car). However, due to the entanglement of object movement and camera ego-motion, it is difficult to estimate the individual object motion from video.

This difficulty can be tackled by introducing the information of surrounding structure, *i.e.* a per-pixel depth map. Depth estimation from image is a fundamental problem in computer vision. Recently the view-synthesis based approach provides a self-supervised learning framework for depth estimation, without supervision of depth annotation. Strong baselines of depth prediction have been established in [10, 20, 35], most of which jointly train a depth and a pose network (for predicting camera ego-motion).

The depth and camera ego-motion can only explain the

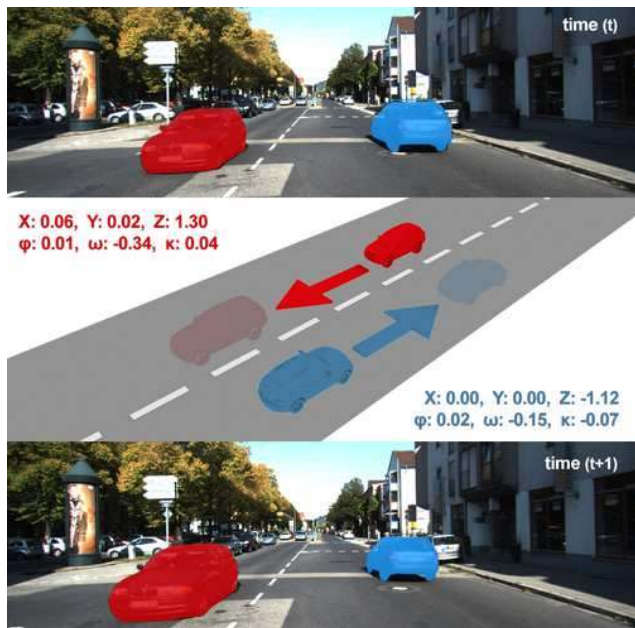


Figure 1. Our system predicts individual object motion by leveraging the instance-level segmentation mask. For each segmented object, three translation ( $X, Y, Z$ ) and three rotation elements ( $\varphi, \omega, \kappa$ ) are predicted. The prediction describes the object movement during the capture of two consecutive frames ( $I_t$  and  $I_{t+1}$ ), within the camera coordinate system of  $I_t$ . The unit of translation and rotation elements are meter and degree respectively.

pixel displacement in static background. To explain the motion of dynamic object, 2D optical flow ([33]) and 3D scene flow map ([3]) have been used to model the object motion. For example, Luo *et al.* [20] proposed to jointly estimate depth, camera ego-motion and optical flow map.

In this paper, we propose a self-supervised learning framework for estimating the individual object motion and the monocular depth from video. The object motion is modelled in the form of a 6 degree-of-freedom (*dof*) rigid-body transform. We further eliminate the scale ambiguity of motion prediction by imposing a novel geometric constraint loss term. Previous approaches use dense flow map

to model the motion, meaning a pixel-wise flow map is predicted. By contrast, our approach predicts a 6 *dof* rotation for the motion of individual object. The number of values to be estimated is significantly reduced from a pixel-wise prediction to 6 scalars per instance.

We perform evaluations of our framework on KITTI dataset. The result manifests the effectiveness of our system to predict individual object motion. Our system outperforms other self-supervised approaches in scene flow prediction, and improve the disparity prediction in dynamic area of the image.

## 2. Related Work

Our system is developed to solve the individual object motion from video, and provide monocular depth estimation. In this section we firstly present works related to depth estimation from image. Then some methods which address the object motion are introduced.

**Supervised Depth Estimation** The depth estimation is formulated as a regression problem in most supervised approaches, where the difference between the predicted depth and its ground truth is minimized. The manually defined feature is used in early work. Saxena *et al.* [28] propose to estimate the single-view depth by training Markov random field(MRF) with hand-crafted features. Liu *et al.* [19] integrate semantic labels with MRF learning. Ladicky *et al.* [15] improve the performance by combining the semantic labeling with the depth estimation.

Deep convolutional neural network (CNN) is good at extracting features and inspires many other methods. Eigen *et al.* [7] propose a CNN architecture to produce dense depth map. Based on this architecture, many variants have been proposed to improve the performance. Li *et al.* [18] improve the estimation accuracy by combining the CNNs with the conditional random field(CRF), while Laina *et al.* [16] use the more robust *Huber loss* as the loss function. Patil *et al.* [26] produce a more accurate depth estimation by exploiting spatio-temporal structures of depth across frames.

**Self-supervised Depth Estimation** The depth map can be learned from unlabeled video under a view-synthesis based framework [35]. This framework is primarily supervised by the image reconstruction loss, which is a function of depth prediction. Zhou *et al.* [35] proposed to jointly train two networks for estimating dense depth and camera ego-motion, respectively. The image is synthesized from the network outputs, following the traditional *Structure-from-motion* procedure. Extra constraint and additional information have been introduced to improve the performance, like the temporal depth consistency [22], the stereo matching [23] and the semantic information [34]. Godard *et al.* [11] achieved a significant improvement by compensating for image occlusion.

Besides estimating depth from the monocular video, [11, 20] have proposed to synthesize stereo image pairs for depth estimation. Here the stereo image pairs have been calibrated in advance, the pose network is thus no longer necessary. Depth prediction from this set-up is free of scale ambiguity issue, since the scale information is introduced from the calibrated stereo image pairs.

**Compensation for Object Motion** Most self-supervised monocular depth estimation approaches are subject to rigid scene assumption: scenes captured by video are assumed to be rigid. This assumption is not valid in most autonomous driving scenario, where many moving objects are presented.

The object motion can be solved by introducing the optical flow map. Yin *et al.* [33] proposed to estimate the residual flow on top of the rigid flow, which is computed from the predicted depth and camera ego-motion. This residual flow can only correct for small error but generally fail for big pixel displacement, *e.g.* when the object is moving fast. Lee *et al.* [17] proposed to estimate the residual flow from stereo video. Luo *et al.* [20] proposed to jointly train networks for depth, camera ego-motion, optical flow and motion segmentation, with enforcing the consistency between each prediction. In [27] a similar architecture is adopted, while the system is trained in a competitive collaboration manner. Both [27] and [20] produced State-of-the-art (SoTA) performance of optical flow prediction on KITTI dataset.

Beyond the scope of self-supervised learning, the estimation of optical flow has been addressed through end-to-end deep regression based methods [6, 13]. PWC-Net [29] further improves the efficiency by integrating the pyramid processing and cost volume into their system. Besides optical flow, scene flow [30] has been introduced to solve the object motion. Scene flow vector describes the 3D motion of a point. [31, 32, 24] estimated the scene flow by fitting a piece-wise rigid representations of motion. They decompose the scene into small rigidly moving plane and solve their motion by enforcing some constraints, like appearance or constant velocity consistency in [31]. Batraway *et al.* [2] introduced sparse LiDAR to estimate scene flow together with stereo images. DRISF [21] formulates the scene flow estimation as energy minimization in a deep structured model, which can be solved efficiently and outperforms all other approaches.

In this work we estimate the object motion by modelling it as a rigid-body transform. The scale ambiguity of motion prediction is solved by imposing a geometric constraint loss term. Our network output describes the object movement in 3D space. This is fundamentally different with the work of Casser *et al.* [4], where only an up-to-scale prediction is predicted. This means the magnitude information of motion is missing in their prediction. Neither did they provide the evaluation of the object motion prediction.

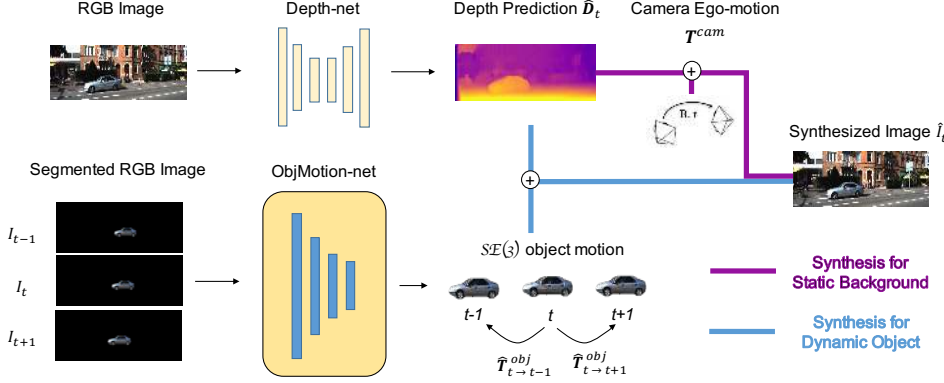


Figure 2. Framework overview. For view synthesis, each pixel is distinguished as either dynamic or static pixel. Dynamic pixel is synthesized from the individual object motion and depth prediction, while static pixel is reconstructed from the depth and camera ego-motion. The camera ego-motion is pre-computed from the visual odometry library [9]. The distinguish of static/dynamic pixel is based on the segmentation mask, provided by Mask R-CNN [12].

### 3. Method

We propose a framework for jointly training an object motion network (ObjMotion-net) and a depth network (Depth-net). We firstly explain the view synthesis for dynamic objects, and then provide an overview of our framework. Our networks are supervised by four losses, which are detailed in Sec. 3.3.

#### 3.1. Theory of View Synthesis

The target frame  $I_{tgt}$  is synthesized from the source frame  $I_{src}$ . For each pixel  $p_{tgt}$  in  $I_{tgt}$ , its correspondence  $p_{src}$  in  $I_{src}$  is required. The photometric consistency between the synthesized view  $\hat{I}_{tgt}$  and its reference  $I_{tgt}$  serves as the primary supervision in our system.

**Synthesis for Static Area** Suppose two consecutive frames from a video are given: the target frame  $I_{tgt}$  captured at time  $t$ , and the source frame  $I_{src}$  captured at time  $t+1$ . For pixel  $p_{tgt}$  in the static area of  $I_{tgt}$ , its correspondence  $p_{src}$  in  $I_{src}$  is computed from Eq. 1:

$$\begin{aligned} h(p_{src}) &\sim K T_{t \rightarrow s} X^t(p_{tgt}) \\ X^t(p_{tgt}) &= \hat{D}(p_{tgt}) K^{-1} h(p_{tgt}) \quad p_{tgt} \in S_0(I_{tgt}) \end{aligned} \quad (1)$$

where  $h(p)$  denotes the homogeneous pixel coordinates,  $K$  is the camera intrinsics,  $T_{t \rightarrow s}$  is the camera ego-motion for the reference system  $C_{tgt}$  and  $C_{src}$ ,  $X^t(p_{tgt})$  is the projected 3D point of  $p_{tgt}$  in the reference system  $C_{tgt}$ ,  $\hat{D}(p_{tgt})$  denotes the depth prediction scalar at  $p_{tgt}$ ,  $S_0(I_{tgt})$  refers to the static area of  $I_{tgt}$ .

**Synthesis for Dynamic Area** Pixel correspondence for dynamic object is computed from Eq. 2. Here the 3D point  $X^t(p_{tgt})$  is further transformed by a rigid-body transform  $\hat{T}_i^{obj} \in SE(3)$  (6 *dof*, 3 translations and 3 Euler angles). This process is illustrated in Fig. 3.

$$h(p_{src}) \sim K T_{t \rightarrow s} \hat{T}_i^{obj} X^t(p_{tgt}) \quad p_{tgt} \in S_i(I_{tgt}) \quad (2)$$

Here  $S_i(I_{tgt})$  refers to pixels in the dynamic area of  $I_{tgt}$ , whose 3D motion is described as  $\hat{T}_i^{obj}$ . Suppose there

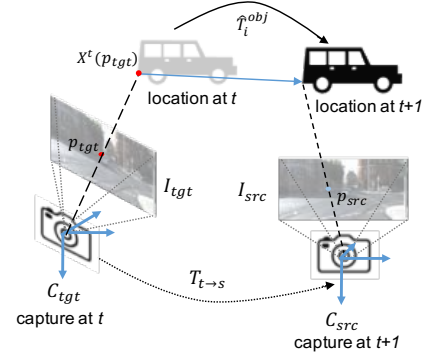


Figure 3. View synthesis for dynamic object. Firstly  $p_{tgt}$  is projected into the target camera reference system  $C_{tgt}$ , denoted as a 3D point  $X^t(p_{tgt})$ . This point is then transformed by  $\hat{T}_i^{obj}$  (for object motion) and  $T_{t \rightarrow s}$  (for camera ego-motion), and is finally projected onto  $I_{src}$  as the correspondence  $p_{src}$ .

are  $n$  moving objects in the scene, we estimate  $\hat{T}_i^{obj}$  ( $i = 1, \dots, n$ ) for each individual object. Then the target frame  $I_{tgt}$  is synthesized separately for static pixels according to Eq. 1, and for dynamic pixels according to Eq. 2.

Note we only focus on objects whose movement can be described by a rigid-body transform. These include cars, buses and trucks. Objects like pedestrians are not considered since their movement is too complicated to be described by a 6 *dof* rigid-body transform.

#### 3.2. Framework Overview

Fig. 2 provides an overview of our framework. It illustrates how the image is synthesized from the network output: depth and object motion for all instances in the scene. We distinguish between the static and dynamic area based on image segmentation mask. The segmentation masks are obtained from the pre-trained Mask R-CNN [12] model. They highlight instances which move rigidly in the scene.

The segmentation mask also distinguishes between different instances in the scene. We align the instance mask across time, and segment the temporal image sequence by the instance-aligned mask. One masked sequence example

is shown as  $I_{t-1}$ ,  $I_t$  and  $I_{t+1}$  in Fig. 2. This serves as the network input to predict the motion  $\hat{T}_{t \rightarrow t-1}^{obj}$  and  $\hat{T}_{t \rightarrow t+1}^{obj}$  for this specific object.

In implementation, the actual network prediction is the product of the camera ego-motion and the object motion (i.e.  $T_{t \rightarrow s} \times \hat{T}_i^{obj}$  in Eq. 2). We combine the camera ego-motion and object motion into one single transformation. This transformation is equivalent to a *pesudo* object motion where we assume the camera is static. The *actual* object motion can be decomposed based on the pre-computed camera ego-motion. Combining the camera and object motion together facilitates the employment of geometric constraint loss term (defined in Eq. 5), which encodes the magnitude information of object motion.

It is noteworthy that motionless objects are also highlighted by Mask R-CNN. They are treated equally as dynamic objects in our system. It is unnecessary to distinguish between static and dynamic objects, since the input of ObjMotion-net is a segmented image sequence, where only one individual instance is presented. The motion predictions of static objects are equal to the camera ego-motion.

**Object Motion Network** The ObjMotion-net is designed to predict individual object movement. It takes the masked image sequence (shown in Fig. 2) as input. All information irrelevant with the target object is excluded.

The idea of ObjMotion-net is inspired by the Pose-net. Both networks take image sequence as input, and output 6 motion parameters. As shown in [35], Pose-net is capable to infer the camera ego-motion. This indicates Pose-net can conduct feature extraction and matching which are indispensable procedures for motion inference. We suppose ObjMotion-net, which adopts a similar architecture, also has the capability to extract and match features, and can infer the individual object motion based on these information.

### 3.3. Loss Function

Our framework employs four loss terms: photometric loss  $L_p$ , left-right photometric loss  $L_{lrp}$ , disparity smoothness loss  $L_{disp}$  and geometric constraint loss  $L_{gc}$ .

**Photometric Loss**  $L_p$  penalizes the photometric inconsistency between the synthesized view  $\hat{I}$  and its reference view  $I$ .  $\hat{I}$  is synthesized based on the prediction from Depth-net and ObjMotion-net, thus  $L_p$  provides gradient on both networks. We adopt a robust image similarity measurement SSIM for  $L_p$  as formulated in Eq. 3, with  $\alpha = 0.85$ . The depth is predicted and supervised at multi-scale level to overcome the gradient locality [35].

$$L_p = \alpha \frac{1 - SSIM(I, \hat{I})}{2} + (1 - \alpha) \|I - \hat{I}\|_1 \quad (3)$$

Note we distinguish the static and dynamic area for the synthesized view  $\hat{I}$  when we compute its photometric loss. Instead of averaging the per-pixel photometric difference

over the whole image, we average the difference in static and dynamic area separately, and formulate the  $L_p$  by summing them. According to [3], the separation of photometric loss can compensate the unbalance between the static and dynamic image area, thus provide more supervision signal and contribute to the training of ObjMotion-net.

**Left-right Photometric Loss**  $L_{lrp}$  is imposed to solve the scale ambiguity of the monocular depth prediction. The direct output of our Depth-net is actually the disparity. It can be used to synthesize the left image from its right counterpart, and vice versa.  $L_{lrp}$  penalizes the photometric difference of the synthesized stereo images. This provides supervision to solve the scale ambiguity of disparity predictions.

**Disparity Smoothness Loss**  $L_{disp}$  is enforced to penalize a fluctuated disparity prediction. An edge-aware smoothness term is imposed as formulated in Eq. 4. Here the disparity smoothness ( $\partial_x d$  and  $\partial_y d$ ) is weighted by the exponential image gradient ( $e^{\|\partial_x I\|}$  and  $e^{\|\partial_y I\|}$ ).  $x$  and  $y$  refers to the gradient along the horizontal or vertical direction.

$$L_{disp} = |\partial_x d| e^{\|\partial_x I\|} + |\partial_y d| e^{\|\partial_y I\|} \quad (4)$$

**Geometric Constraint Loss** During experiments we found the translation of object motion tends to be predicted as small values. Similar phenomenon was also observed in [4]. We fix this issue by imposing a geometric constraint on the object translation prediction. This constraint provides the magnitude information of the object movement. The geometric constraint  $F_i^{t \rightarrow t+1}$  for the  $i$ -th object between time  $t$  to  $t + 1$  is computed as Eq. 5:

$$\begin{aligned} F_i^{t \rightarrow t+1} &= \bar{X}_i^{t+1} - \bar{X}_i^t \\ \bar{X}_i^m &= \left| \sum X_i^m(p) \right| \quad p \in S_i(I_m), m \in \{t, t+1\} \\ X_i^m(p) &= \hat{D}^m(p) K^{-1} h(p) \quad m \in \{t, t+1\} \end{aligned} \quad (5)$$

$F_i^{t \rightarrow t+1}$  is actually the vector from the 3D object center  $\bar{X}_i^t$  to  $\bar{X}_i^{t+1}$ . Here  $|\cdot|$  refers to the mean operator.  $X_i^m(p)$  is the projected 3D point of pixel  $p$  in the reference system  $C_m$ , while  $S_i(I_m)$  is the  $i$ -th object area of image  $I_m$ , with  $m \in \{t, t+1\}$  denoting the image capture time.

As mentioned in 3.2, our system predicts a *pesudo* object motion where we assume the camera is static. Ideally the predicted (pesudo) object translation are supposed to be equivalent with the geometric constraint. We impose the geometric constraint loss term  $L_{gc}$ , which is the L1-norm of the difference between the predicted translation of object motion  $\hat{t}_i$  and the geometric constraint  $F_i$ .

$$L_{gc} = \sum_i^n \|\hat{t}_i - F_i\|_1 \quad \hat{t}_i = [\hat{x}_i, \hat{y}_i, \hat{z}_i] \quad (6)$$

Here  $n$  is the number of instances appeared in the input image sequence. With introducing  $L_{gc}$ , the issue of the small



translation prediction can be fixed. Ablation studies are provided in Sec. 4.3.

Our final objective is a sum of all loss terms stated above, weighted by their corresponding weight:

$$L_{final} = \lambda_p \cdot L_p + \lambda_{lrp} \cdot L_{lrp} + \lambda_{disp} \cdot L_{disp} + \lambda_{gc} \cdot L_{gc} \quad (7)$$

## 4. Experiments

In this section, we firstly describe the implementation details, and demonstrate evaluation results on individual object motion, disparity, and scene flow prediction. Experiments are conducted on KITTI [8], a dataset provides driving scenes in real-world scenario.

The ObjMotion-net and Depth-net are trained jointly, since the view synthesis for dynamic scene requires both the object motion and depth prediction. However, these two networks can be run independently during test time inference. Their network inputs are irrelevant with each other.

### 4.1. Implementation Details

**Dataset and Preprocessing** KITTI raw dataset provides videos which cover various scenes. We resize all images into a fixed size  $192 \times 640$ , and format a temporal image sequence by concatenating  $I_{t-1}$ ,  $I_t$  and  $I_{t+1}$  horizontally.

The evaluation is performed on the training split of KITTI flow 2015 dataset, where the ground truth for disparity, optical flow and scene flow are available. Scenes covered by this training split are excluded during training. 40820 samples and 2070 samples are formatted for training and validation, respectively. Besides the raw dataset, we format another training set from the test split of the multi-view extension of KITTI flow 2015 dataset. Scenes in this split contain more moving vehicles. This contributes to the training of ObjMotion-net. There are 6512 training samples and 364 validation samples in this training set.

The segmentation mask for image is generated from the pre-trained Mask R-CNN model [12]. We segment objects which move rigidly in the scene. The instance is aligned according to the Intersection over Union (IoU) of the temporal mask sequence. For example,  $M_t^i$  is the mask of instance  $i$  at time  $t$ . Its aligned mask  $M_{t-1}^i$  and  $M_{t+1}^i$  are obtained by finding the instance mask with the maximum IoU at time  $t-1$  and  $t+1$ . For partially occluded or fast moving objects whose IoU is small, we further check the moving direction of the mask center. We assume the object moving direction between  $t-1$  to  $t$  and  $t$  to  $t+1$  (*i.e.* the 2D vector which connects mask centers) are similar. Aligned masks with significantly different moving direction are discarded. We also ignore very small instances (*i.e.* the number of mask pixels is less than 400 in one  $1392 \times 512$  image).

The camera ego-motion is required for view synthesis. Instead of training a pose network, we use the Libviso2 [9]

to estimate the camera ego-motion.

**Network Architecture** Our system contains two sub-networks, the ObjMotion-net and the Depth-net. The ObjMotion-net is designed based on the pose network in [33]. We adopt ReLU [25] activation for all convolutional layers. Batch normalization (BN) [14] is excluded in ObjMotion-net, since experiments demonstrate BN does not contribute to the performance.

For the Depth-net, we adopt the architecture in [33] as backbone. This structure consists of the encoder and the decoder part. The basic structure of ResNet50 is adopted for the encoder. While in decoder the combination of convolution and upsampling is used for upscaling the feature map. Skip connections between the encoder and the decoder are added to integrate global and local information. ReLU and BN are adopted for all layers of Depth-net except for the prediction layer, where the Sigmoid activation is used and BN is excluded.

### 4.2. Training Details

Our system is implemented using TensorFlow framework [1]. Color augmentation is performed on the fly. The network is optimized using Adam optimizer, with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  respectively. Our system is trained on a single TitanXP GPU. A stage-wise training strategy is adopted, with training Depth-net alone at the beginning, and then jointly training Depth-net and ObjMotion-net.

**Training Depth-net** We firstly train the Depth-net, since accurate depth prediction is necessary to compute the geometric constraint for the ObjMotion-net. According to Godard *et al.* [10], pre-training network on Cityscapes dataset [5] contributes to the performance of Depth-net. We trained the Depth-net on the Cityscapes dataset for 600K steps. The training was supervised by the left-right photometric consistency  $L_{lrp}$  ( $\lambda_{lrp} = 1.0$ ) and disparity smoothness  $L_{disp}$  ( $\lambda_{disp} = 0.5$ ). The learning rate and the batch size are 0.0001 and 2, respectively. The photometric loss  $L_p$  is excluded in this stage, since we do not model object motion. Employment of  $L_p$  will lead to inaccurate depth prediction in dynamic area. This issue is also observed by Luo *et al.* [20].

After the pre-training on Cityscapes, we continued to train the Depth-net on the formatted KITTI dataset for 500K steps. All hyper-parameters were kept same except for the  $\lambda_{disp}$ , which we changed to 25.0. We found in experiments that a higher smoothness penalization was indispensable, otherwise the disparity prediction became unreasonably fluctuated.

**Training ObjMotion-net with Depth-net** We then jointly optimize the Depth-net and ObjMotion-net. Besides the  $L_{lrp}$  and the  $L_{disp}$ , the photometric loss  $L_p$  and geometric constraint loss  $L_{gc}$  are imposed. The loss weights are set to



Figure 4. Visualization of Bird’s View Box and 3D Bounding Box. Our results (top row) and results from GeoNet [33] (bottom row) are presented. The ground truth is in red while the prediction is in blue. Our predictions have a larger overlapping with the ground truth box.

Method	Bird View	3D Box
CC [27]	43.10%	43.60%
GeoNet [33]	57.54%	56.00%
Ours (no $L_{gc}$ )	45.02%	43.67 %
Ours	<b>72.31%</b>	<b>70.61%</b>

Table 1. Average IoU of bird’s view box and 3D bounding box.

be  $\lambda_p = \lambda_{lrp} = 1.0$ ,  $\lambda_{disp} = 25.0$ , and  $\lambda_{gc} = 1.0$ . The learning rates for Depth-net and ObjMotion-net are 0.0001 and 0.0002, respectively. And the batch size is 2.

After training on the KITTI raw dataset for 200K iteration, we fine-tune the ObjMotion-net on the test split of KITTI Flow dataset 2015, with fixing the parameters of Depth-net. All hyper-parameters remain the same. The ObjMotion-net is trained for 100K iterations in this stage.

### 4.3. Individual Object Motion Evaluation

Our system predicts individual object motion in 3D space. To demonstrate the effectiveness of our system, we present the IoU of the bird’s view box and 3D bounding box. Take the example of 3D bounding box: the 3D bounding box for object  $i$  at time  $t$ , denoted as  $B_t^i$ , is transformed by its object motion prediction  $\hat{T}_{t+1}^i$ . Then the predicted location of box at time  $t + 1$ ,  $\hat{B}_{t+1}^i$  is obtained. We compute the IoU between  $\hat{B}_{t+1}^i$  and its ground truth  $B_{t+1}^i$ . The average IoU indicates the performance of our ObjMotion-net.

We evaluate on 80 temporal image pairs which are contained in both the training split of KITTI tracking (provides ground truth bounding box) and flow 2015 dataset. Objects are segmented by the pre-trained Mask R-CNN model [12]. We do not use the ground truth segmentation in flow 2015, since it only provides the segmentation for  $I_t$ , while the segmentation for  $I_{t-1}$  and  $I_{t+1}$  are necessary for motion prediction. In total 204 objects are selected from these 80 image pairs.

We compare our results with GeoNet [33] and CC [27]. Both approaches predict dense optical flow and depth map, from which we can compute the pixel-wise scene flow vector in 3D space. The individual object motion can then be

Method	Bad Pixel Percentage		
	bg	fg	all
CC [27]	35.03%	42.74%	36.20%
Monodepth2 [11]	<b>18.60%</b>	44.47 %	<b>22.50%</b>
EPC++ [20]	22.76%	26.63%	23.84 %
Ours (1st stage)	29.49%	19.62%	28.00%
Ours	28.08%	<b>16.65%</b>	26.36%

Table 2. Bad pixel percentage of disparity prediction

inferred by averaging over the object segmentation mask on the scene flow map. Their up-to-scale depth prediction are scaled by the median ground truth depth.

In Fig. 4, we compare our qualitative results with GeoNet [33]. It can be seen that our predicted bounding boxes have a higher overlapping with the ground truth. Quantitative results in Table 1 also show our system has a higher average IoU, with 72.31 % for bird’s view box and 70.61% for 3D bounding box. This demonstrates the effectiveness of our system to predict the individual object motion. We report the evaluation result without geometric constraint loss  $L_{gc}$ . It can be seen that without  $L_{gc}$  the performance is significantly worse. The lower IoU of CC [27] prediction results from its inaccurate depth prediction.

Besides IoU of bounding box, Cao *et al.* [3] proposed to evaluate the motion speed and direction. However, they did not publish their models and evaluation details. Thus it is not possible to compare with them.

### 4.4. Monocular Disparity Evaluation

To demonstrate the contribution of ObjMotion-net towards disparity estimation (in particular for dynamic area), we evaluate on the training split of flow 2015 dataset and report the average bad pixel percentage (BPP) of disparity prediction. A pixel is considered as bad if its prediction error  $\geq 3px$  or  $\geq 5\%$ . Besides, the ground truth segmentation masks for moving objects are provided. This makes it possible to evaluate within dynamic region (*fg* in Table 2), which in our case is the Region-of-Interest. The BPP for static background (*bg*) and overall area (*all*) are also presented.

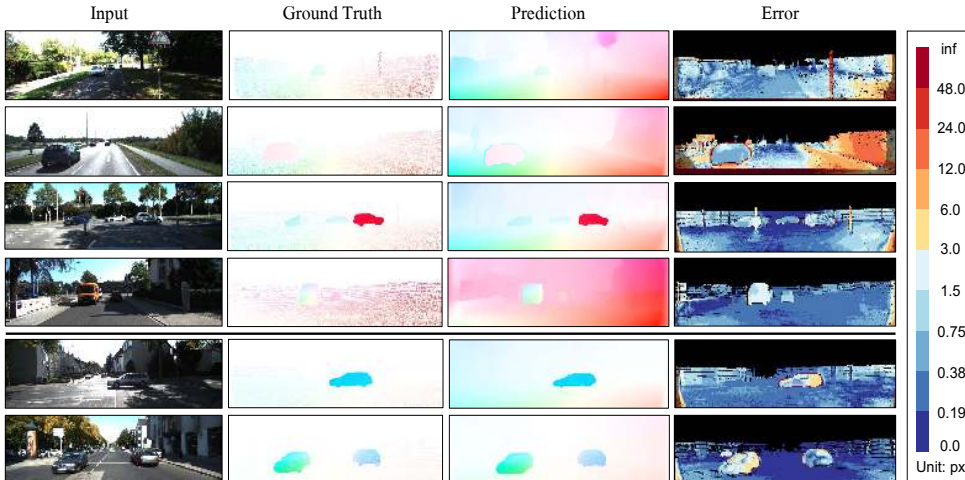


Figure 5. Visualization of optical flow prediction for KITTI flow 2015 training split. The error magnitude is encoded into different colors according to the legend at the right-hand side. Basically good pixel is in blue while bad pixel is in orange/red. The top four rows show some succeeded samples, where the error of most pixels in dynamic region are below 3px. The bottom two rows show two imperfect examples. The bias of bad pixel is slightly over the threshold due to the imperfection of view synthesis.

Method	bg	fg	all
GeoNet [33]	66.8%	90.4%	70.7%
Mono + Geo	39.4%	70.9%	44.7%
EPC++ [20]	> 22.8%	> 70.4%	> 60.3%
CC [27]	50.2%	<b>60.0%</b>	51.8%
Ours	38.2%	65.9%	<b>42.8%</b>

Table 3. Bad pixel percentage of scene flow prediction. For the evaluation of *Mono + Geo*, the disparity from Monodepth2 [11] and the flow prediction from GeoNet [33] are used.

Our result achieves SoTA performance in terms of BPP in fg. We produce the lowest value of 16.65% in foreground, which is better than BPP of Monodepth2 (44.47%) and EPC++ (26.63%). It is important to note that the published models of other works (CC, Monodepth2 and EPC++) have been trained on the test images, since they adopted another training split (Eigen split [7]). Our system did not witness the test images, but still produce a better performance in foreground and comparable result in overall area. This demonstrates the disparity prediction for dynamic objects can be improved through modelling the object motion explicitly.

#### 4.5. Scene Flow Evaluation

We present the evaluation results on 3D scene flow in Table 3. The evaluation conducts on the predicted disparity for two consecutive frames:  $\hat{D}(I_t)$  and  $\hat{D}(I_{t+1})$ , and the 2D optical flow map  $\hat{F}_{t \rightarrow t+1}$ . In our system, we do not have a component to explicitly predict the pixel-wise optical flow. The optical flow prediction is obtained through view synthesis with taking the object motion into account. We present the visualization of our optical flow results in Fig. 5.

Our result achieves the best BPP in overall area (42.8%), compared with 44.7% from Mono + Geo, and 51.8% from CC. In foreground, our BPP (65.9%) is worse than the results of CC (60.0%). This is because our flow results are

synthesized from depth and object motion prediction. Any subtle bias in view synthesis (like camera intrinsics, depth, object motion) may result in an error larger than the bad pixel threshold (3px). We present two imperfect flow prediction examples in the last two rows of Fig. 5. Although some bad pixels are presented in the dynamic area, the magnitude of their bias are barely over the threshold due to the imperfect view synthesis.

Nevertheless, our system is capable to capture the holistic object motion. Our system produces the lowest BPP in overall area, and it can be seen from Fig. 5 that the bias of most dynamic pixels are below the bad pixel threshold.

## 5. Conclusion

We have presented a self-supervised learning framework for individual object motion and depth estimation. The object motion is modelled as a rigid-body transformation. Our system is able to learn the object motion from unlabelled video. This contributes to scene flow prediction, and improve the depth estimation in dynamic area of the scene.

It would be interesting to explore the following questions in future: 1) Integrate scale information from other sources. In our system the scale information of object motion is extracted from the depth prediction (the *control signal*). It would be difficult to apply our system in case where the depth prediction is not reliable. In future we can try to integrate scale information from other sources, like stereo image pairs, or sparse depth *ground truth* from Lidar. 2) Estimate the motion of pedestrians. Currently we focus on the object motion which can be described by a rigid-body transformation. While non-rigid motion, like the movement of pedestrians, is also common in driving scenario. We may dissect pedestrians into smaller parts which is moving rigidly, or try to model its motion in a different way.



## References

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. [5](#)
- [2] Ramy Battrawy, René Schuster, Oliver Wasenmüller, Qing Rao, and Didier Stricker. Lidar-flow: Dense scene flow estimation from sparse lidar and stereo images. *IROS*, 2019. [2](#)
- [3] Zhe Cao, Abhishek Kar, Christian Häne, and Jitendra Malik. Learning independent object motion from unlabelled stereoscopic videos. In *CVPR*. 2019. [1](#), [4](#), [6](#)
- [4] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. *AAAI*, 2019. [2](#), [4](#)
- [5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [5](#)
- [6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2015. [2](#)
- [7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, pages 2366–2374, 2014. [2](#), [7](#)
- [8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012. [5](#)
- [9] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Intelligent Vehicles Symposium (IV)*, 2011. [3](#), [5](#)
- [10] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. [1](#), [5](#)
- [11] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *ICCV*, 2018. [2](#), [6](#), [7](#)
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. [3](#), [5](#), [6](#)
- [13] 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*, pages 2462–2470, 2017. [2](#)
- [14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. [5](#)
- [15] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014. [2](#)
- [16] 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*, pages 239–248. IEEE, 2016. [2](#)
- [17] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning residual flow as dynamic motion from stereo videos. *IROS*, 2019. [2](#)
- [18] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015. [2](#)
- [19] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260. IEEE, 2010. [2](#)
- [20] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *arXiv preprint arXiv:1810.06125*, 2018. [1](#), [2](#), [5](#), [6](#), [7](#)
- [21] Wei-Chiu Ma, Shenlong Wang, Rui Hu, Yuwen Xiong, and Raquel Urtasun. Deep rigid instance scene flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3614–3622, 2019. [2](#)
- [22] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018. [2](#)
- [23] Diogo Martins, Kevin Van Hecke, and Guido De Croon. Fusion of stereo and still monocular depth estimates in a self-supervised learning context. In *IEEE Robotics and Automation Letters (ICRA)*, pages 849–856, 2018. [2](#)
- [24] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. [2](#)
- [25] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. [5](#)
- [26] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. *arXiv preprint arXiv:2001.02613*, 2020. [2](#)
- [27] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019. [2](#), [6](#), [7](#)



- [28] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006. [2](#)
- [29] 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*, pages 8934–8943, 2018. [2](#)
- [30] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 722–729. IEEE, 1999. [2](#)
- [31] Christoph Vogel, Konrad Schindler, and Stefan Roth. Piecewise rigid scene flow. In *ICCV*, pages 1377–1384, 2013. [2](#)
- [32] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *ECCV*, pages 756–771. Springer, 2014. [2](#)
- [33] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018. [1](#), [2](#), [5](#), [6](#), [7](#)
- [34] Junming Zhang, Katherine A Skinner, Ram Vasudevan, and Matthew Johnson-Roberson. Dispsegnet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery. *IEEE Robotics and Automation Letters (ICRA)*, 4(2):1162–1169, 2019. [2](#)
- [35] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. [1](#), [2](#), [4](#)