

3D Object Aided Self-Supervised Monocular Depth Estimation

Songlin Wei, Guodong Chen*, Wenzheng Chi*, Zhenhua Wang and Lining Sun

Abstract—Monocular depth estimation has been actively studied in fields such as robot vision, autonomous driving, and 3D scene understanding. Given a sequence of color images, unsupervised learning methods based on the framework of Structure-From-Motion (SfM) simultaneously predict depth and camera relative pose. However, dynamically moving objects in the scene violate the static world assumption, resulting in inaccurate depths of dynamic objects. In this work, we propose a new method to address such dynamic object movements through monocular 3D object detection. Specifically, we first detect 3D objects in the images and build the per-pixel correspondence of the dynamic pixels with the detected object pose while leaving the static pixels corresponding to the rigid background to be modeled with camera motion. In this way, the depth of every pixel can be learned via a meaningful geometry model. Besides, objects are detected as cuboids with absolute scale, which is used to eliminate the scale ambiguity problem inherent in monocular vision. Experiments on the KITTI depth dataset show that our method achieves State-of-The-Art performance for depth estimation. Furthermore, joint training of depth, camera motion and object pose also improves monocular 3D object detection performance. To the best of our knowledge, this is the first work that allows a monocular 3D object detection network to be fine-tuned in a self-supervised manner.

I. INTRODUCTION

Monocular depth estimation is a challenging task for 3D scene understanding. Research on monocular depth estimation has grown rapidly due to its relatively low cost, ease of deployment, and freedom from complex calibration. Among them, unsupervised learning based methods reformulate the problem of depth estimation as an image reconstruction problem by using the photometric consistency between consecutive image frames as a supervisory signal. Current self-supervised learning based methods for depth estimation can even achieve comparable accuracy to supervised learning methods, which makes unsupervised depth estimation very attractive because it avoids the laborious collection of large amounts of data in real-world environments. The unsupervised neural network typically consists of two sub-networks for learning the depth map and the relative pose between consecutive frames, respectively. By using the estimated depth of the target image and the motion of the camera to transform nearby frames, a reconstructed view is obtained to compare with the target view. Then the photometric error between the reconstructed view and the target view is used to train the whole network. However, when there

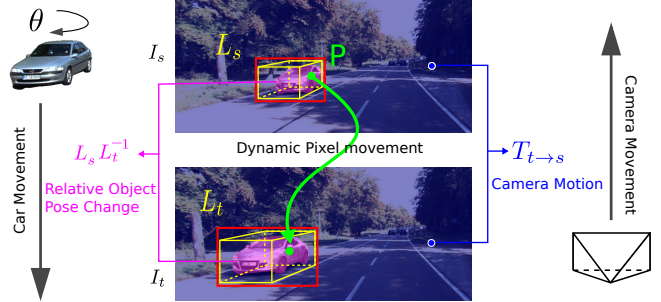


Fig. 1. The dynamic pixel movement between target view I_t and source view I_s is explicitly modeled with the detected 3D object poses L_t and L_s , while the static pixel is modeled with $T_{t \rightarrow s}$. In this way, the depth of every pixel can be learned via a meaningful geometric model. Detail description is given in Section III-A.

are dynamically moving objects in the scene that violate the static world assumption, the accuracy of predicted depth is affected.

In order to address this problem, some researchers such as [1] [2] try to eliminate the impact of dynamic objects by masking out the pixels which are unexplainable by the camera motion. However, the pixels belonging to the dynamic objects are not modeled. Another work [3] employs FlowNet [4] to model the movement of dynamic pixels directly and revise the loss function accordingly. However, the accuracy of depth estimation is limited with the performance of the unsupervised FlowNet. Different from these methods, in this work, we propose to exploit 3D object detection to model the dynamic moving objects explicitly, as illustrated in Fig. 1. By modeling the movement of pixels on the dynamic object, the depth of the dynamic pixels will no longer be coupled with the camera motion T to formulate the appearance loss but rather the object pose L . Then the depth of the dynamic pixels can be learned via a meaningful appearance loss.

Moreover, scale ambiguity is inherent in monocular vision, i.e., the depth and translation of the relative camera pose between two frames can only be recovered up to an unknown scale. Guizilini et al. [5] proposed to use the velocity of the camera to constrain the scale. The velocity can be obtained through the IMU unit. However, velocity information has accumulated drift over time, for example, when the camera is moving at a constant speed. In this work, we propose to exploit the absolute scale of detected objects to constrain the translation of camera movement so as to recover the true scale.

After introducing 3D object detection into the monocular depth estimation problem, the photometric loss between the warped and target images can be back-propagated to the

*Corresponding author

Research was partially supported by the National Key R&D Program of China grant #2019YFB1310201.

The authors are with the Robotics and Microsystems Center, School of Mechanical and Electric Engineering, Soochow University, Suzhou 215021, China {slwei, wzchi, chengudong, wangzhenhua, lining.sun}@suda.edu.cn

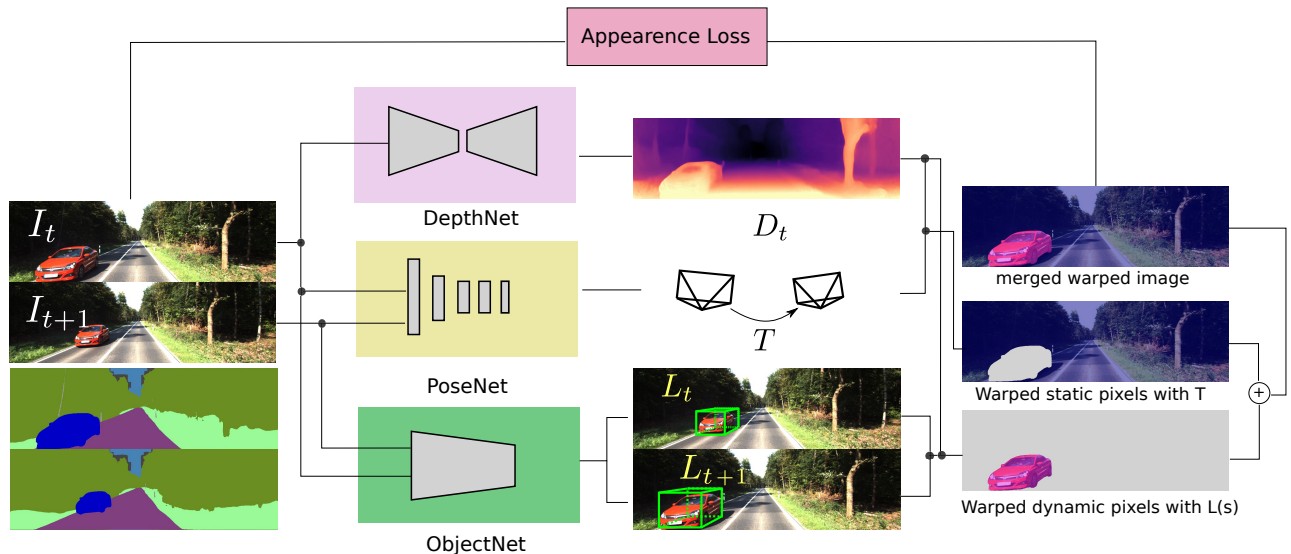


Fig. 2. Overview of the network. Image I_t is passed into DepthNet to predict depth map D_t . Both target view I_t and source views I_s are passed into PoseNet to predict the camera relative pose T , and the two views are also passed into ObjectNet to detect objects in each view. The reconstructed image is warped with D_t , T and $L(s)$ for static background and dynamic objects, respectively. Finally, the appearance loss of the reconstructed image and target image is calculated. The whole network can then be trained by minimizing the appearance loss. Especially, the ObjectNet can be trained along with DepthNet and PoseNet in an unsupervised manner.

3D object detection network. Therefore, 3D object detection networks can be fine-tuned in a self-supervised manner. This has been shown in experiments to benefit both object detection and depth estimation.

Our main contributions are summarized as follows:

- We novelly introduce monocular 3D object detection into the self-supervised depth estimation problem to address the dynamic object movements. After we detect 3D objects in each view, the per-pixel correspondence can be built either by relative object pose change or camera motion. In this way, the depth of every pixel can be learned via a meaningful geometric model.
- A novel scale ambiguity elimination method is proposed with the absolute scale of the detected 3D objects.
- The 3D object detection network is fine-tuned in a self-supervised manner along with depth and pose estimation networks.

II. RELATED WORK

Simultaneously estimating geometric structure and camera relative pose has long been well studied in Structure-from-Motion. Whilst Deep learning is recently adopted to deal with more challenging scenarios where traditional feature extraction and matching pipeline fails. The most appealing aspect of deep learning methods is that it learns depth knowledge as a prior and can estimate the dense map given a single image without predicting the camera pose during test time.

1) *Self-supervised Monocular Depth Estimation*: Instead of directly regressing depth values using convolutional neural networks as supervised methods do, self-supervised methods redefine depth prediction as an image reconstruction problem. Garg et al. [6] proposed an autoencoder structured

network to predict the depth map of an image. Then, the fixed displacement between the two views of the binocular camera is used to warp one of the views, and the photometric error of the reconstructed view and with the opposite view is used to train the network. However, this approach produces "texture-copy" artifacts and depth discontinuities. To address this issue, Godard et al. [7] proposed to predict both left-to-right and right-to-left disparities and further enforce the spatial coherence between them, thereby improving performance. Zhou et al. [1] extend the paradigm to monocular vision. The authors propose to use temporal image sequences to simultaneously predict single-view depth and multi-view camera pose via separate DepthNet and PoseNet, which quickly became the de facto standard for monocular depth estimation. Mahjourian et al. [2] and Liu et al. [8] employ geometric models from Structure-from-motion (SfM) to estimate camera ego-motion instead of PoseNet. Guizilini et al. [5] proposes a novel 3D packing and unpacking layer to better preserve the structural details of the input color image during feature shrinkage and expansion, however at the cost of higher computation. Our work adopts canonical PoseNet and DepthNet structures in [1], and introduces additional object detection networks.

2) *Dynamic Objects Handling*: The unsupervised learning based depth estimation method works well in static scenes. However, the presence of dynamic moving objects brings extra challenges, as it breaks the static world assumption. Liu et al. [8] proposed a unified framework for jointly learning optical flow and depth to segment rigid regions. Then the photometric and geometric loss are built on rigid region to avoid the influence of dynamic regions. Klingner et al. [9] applied cross-task training for both semantic segmentation and depth estimation. The feature map learned from semantic

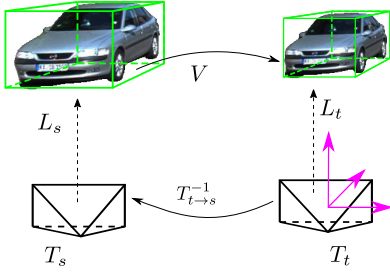


Fig. 3. Coordinate System: The target frame T_t is set as the reference frame. The solid curves illustrate the camera and object movements with arrows indicating the transformation direction. The dotted lines with arrows illustrate the single view object detections. The relationship of the transformation matrices of different poses is given in Equation (2).

segmentation is proved to be beneficial for depth estimation. Besides, the pixels which are labeled with dynamic class objects are also masked out. Other works try to model the dynamic objects explicitly. Yang et al. [3] propose to use an optical flow network to build correspondences of pixels between frames including dynamic pixels. The authors model the dynamic objects with 3D scene flow. Casser et al. [10] propose to take advantage of semantic labels to mask out the objects, which are then fed into a separate similar network with PoseNet to predict the object motion separately. With the predicted motion of the objects, the geometric reprojection model can again be applied to dynamic objects.

Different from these works, we use a monocular 3D object detector [11] to detect objects in each frame and model the dynamic pixels directly with the detected object poses, as illustrated in Fig. 1. We also take advantage of the absolute scale of the detected 3D objects to propose a novel scale loss to constrain the PoseNet to predict the right scale.

III. METHOD

Depth estimation is the problem of estimating the dense depth map D_t given a target view image I_t at time t , as illustrated in Fig. 2, where D_t has the same resolution as I_t with each pixel representing the point distance from the camera. Self-supervised depth estimation tackle this problem with a novel view-synthesis method. The temporally preceding and succeeding frames denoted as $\{I_{t-1}, I_{t+1}\}$ are feed together with I_t into the network for predicting their relative transformation matrices with frame I_t , denoted as $T_{t \rightarrow s} \in SE(3)$, $s \in \{-1, 1\}$. Suppose the camera intrinsic matrix is $K \in \mathbb{R}^{3 \times 3}$, and the homogeneous coordinate of a pixel in the target view is denoted as \bar{p}_t . Following the work of [1] [12], we can obtain the projected homogeneous pixel coordinates in the source image \bar{p}_s as:

$$\bar{p}_s = K T_{t \rightarrow s} D(p_t) K^{-1} \bar{p}_t \quad (1)$$

However, this model only works for the static points in the world. Next, to address this problem, we describe our method of modeling the dynamic points with the detected 3D object poses.

A. Modeling Dynamic Moving Points

To ease illustrating the geometric relationships of camera frames and object frames, the world coordinate system is built at the target image frame, as illustrated in Fig. 3. The camera pose at time s is denoted as T_s . Suppose an object O is observed both in the source and target frame with poses L_s and L_t respectively. Further suppose it moves from time s to t with transformation matrix V . Then we have

$$L_t = T_{t \rightarrow s}^{-1} L_s V. \quad (2)$$

For a physical point $P = [x, y, z, 1]^T$ in the object frame, the position of P relative to the object does not change assuming the object is rigid. Then the 3D coordinate of point P at different times can be obtained by

$$P_t = L_t P \quad (3)$$

$$P_s = L_s V^{-1} P \quad (4)$$

The coordinates are different when the object is moving, that is, V is not the identity matrix. Then the homogeneous coordinate of the corresponding pixel in view I_s can be acquired through projection:

$$\bar{p}_s = K T_{t \rightarrow s} P_s, \quad (5)$$

Notice that there is an implicit conversion from homogeneous coordinates to non-homogeneous coordinates before multiplying to K . After substituting V from (2) and P from (3) into (4) and further substituting P_s from (4) into (5), we have

$$\bar{p}_s = K L_s L_t^{-1} D(p_t) K^{-1} \bar{p}_t \quad (6)$$

where $D(p_t) K^{-1} \bar{p}_t = P_t$ is the back projected 3D coordinates for pixel p_t . Comparing (6) with (1), we can find that the projected pixel of dynamic object points in source frames dependent *only* on object poses from 3D object detector. It does not dependent on camera ego-motion T nor object motion V . Indeed, when the object is not moving, then $T_{t \rightarrow s}$ and $L_s L_t^{-1}$ are equal according to (2). Therefore, the per-pixel correspondence between \bar{p}_s and \bar{p}_t which corresponding to dynamic objects is established through object pose change $L_s L_t^{-1}$.

B. Self-Supervised Learning for Depth Estimation

After the rigid background pixels and the dynamic object pixels are projected into the source frames, the synthesized target view $\tilde{I}_{s \rightarrow t}$ can be acquired through sampling from each source view I_s :

$$\tilde{I}_{s \rightarrow t} = M^{(0)} I_s \langle p_s^{(0)} \rangle + \sum_{i=1} M^{(i)} I_s \langle p_s^{(i)} \rangle, \quad (7)$$

where the superscript (0) denotes the pixel from the rigid background and (i) denotes the pixel from the object i , and $M^{(0)}$ is the binary mask of the rigid background, $M^{(i)}$ is the binary mask of the i -th object. All the masks are acquired in the source image, therefore not overlapping with each other. $\langle \rangle$ is a differentiable bilinear sampling operator. p_s can be computed with (1) if the pixel is static, and (6) if it is

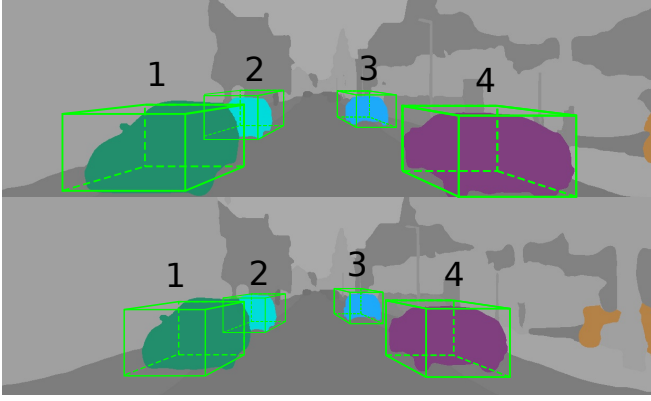


Fig. 4. An example of data association for objects detected in both target view (Up) and source view (Bottom). Each detected object is shown as a cuboid indexed with numbers. Objects with the same numbers are associated.

dynamic. Now the photometric error between the composite view and the target view is formulated as

$$\mathcal{L}_{ap} = \min_s pe(I_t, \tilde{I}_{s \rightarrow t}) \quad (8)$$

Herein, pe is the appearance similarity function and the minimum reprojection loss of all source views for each point is applied. The function pe is defined as:

$$pe(I_t, \tilde{I}_{s \rightarrow t}) = \frac{\alpha}{2}(1 - SSIM(I_t, \tilde{I}_{s \rightarrow t})) + (1 - \alpha)\|I_t - \tilde{I}_{s \rightarrow t}\|_1, \quad (9)$$

which considers both $SSIM$ [13] and $L1$ differences. α is a balancing constant. In projection geometry, the same color image corresponds to numerous different real scenes in the physical world. To make the depth estimation problem solvable, an edge-aware smoothness regularization term [7] is adopted here:

$$\mathcal{L}_{sm} = |\partial_x d_t^*| \exp(-\partial_x I_t) + |\partial_y d_t^*| \exp(-\partial_y I_t) \quad (10)$$

where d_t^* is mean value normalized depth of D_t . We follow Godard et al. [14] to predict the depth map at multiple image scales. Finally, the combined loss function is formulated as:

$$\mathcal{L}_{photo} = \lambda \mathcal{L}_{sm} + \sum_l \mathcal{L}_{ap}^l, \quad (11)$$

where λ is a weighting constant. The superscript l denotes the appearance loss is calculated at image scale l . Then, each subnetwork can be trained by minimizing the final loss \mathcal{L}_{photo} by gradient descent algorithms.

C. Monocular 3D Object Detection

Now we discuss how to acquire the object pose in the source and target views. We use a 3D region proposal network similar to [11] and follow the convention of autonomous driving literature to use 3D cuboids to represent objects in monocular images. Each cuboid C is comprised of 2D bounding box $B_{2d} = [x, y, w, h]_{2d}$, 3D bounding box $B_{3d} = [w, h, l]_{3d}$, center location $t_c = [x, y, z]_c$ and rotation

angle θ with respect to the camera frame, as illustrated in Fig. 1. The object pose L relative to the camera frame is

$$L = \begin{bmatrix} \cos \theta & 0 & \sin \theta & x_c \\ 0 & 1 & 0 & y_c \\ \sin \theta & 0 & \cos \theta & z_c \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (12)$$

Before the detected object pose L can be used to model the dynamic point, the objects detected in every single source and target view must be associated. The problem is also known as data association. Following Wei et al. [15], we consider both the object cuboid shape, location, and 2D box for the association. The score s for each potential association is calculated as

$$s(C^i, C^j) = IoU_{2D}(C^i, C^j) + D(C^i, C^j) + \alpha S(C^i, C^j), \quad (13)$$

where $\alpha = 0.5$ for balancing the similarity for object distance and shapes. IoU_{2D} is the Intersection over Union (IoU) of the two 2D boxes, defined as:

$$D(C^i, C^j) = \exp(-|t_c^i - t_c^j|), \quad (14)$$

which considers the object centroid distance. While S is defined as

$$S(C^i, C^j) = \sum_{s \in w, h, l} \exp(-|B_{3d}^i[s] - B_{3d}^j[s]|), \quad (15)$$

which considers the shape differences of each dimension. As shown in Fig. 4, each object that appeared in the target view is associated with the object with the highest score in the source view. Then, the projection model proposed in (6) can be applied to all the associated objects.

To further improve the precision of the projection model, we take advantage of panoptic segmentation to distinguish whether every pixel in the 2D box is on or off the object. As shown in Fig. 1, some pixels in the 2D bounding box are actually from the background. To remedy this problem, we use the off-the-shelf tool Panoptic-DeepLab [16] to perform panoptic segmentation for each view. All the training images are preprocessed and are fed together with the color images into our network.

D. Scale Recovery with Object Poses

The detected object poses and sizes have true scale because the object detection network is trained with the real size and distance of objects in similar scenes. As a result, the estimated depths of dynamic points that minimize the reprojection appearance loss are consistent with the true scale. For the static points, we design a scale loss to constrain the translation of the camera ego-motion instead of using the additional camera velocity information as proposed in [5]. Specifically, we check the reprojection error of objects with camera ego-motion $T_{s,t}$ against the relative object pose change $L_s L_t^{-1}$. If the two errors are close enough, the object is then determined as static since the reprojection error can also be explained with the camera ego-motion. Then we use the translation of static objects to constrain the translation of the camera. This can be justified by Equation (2), as $L_s L_t^{-1}$

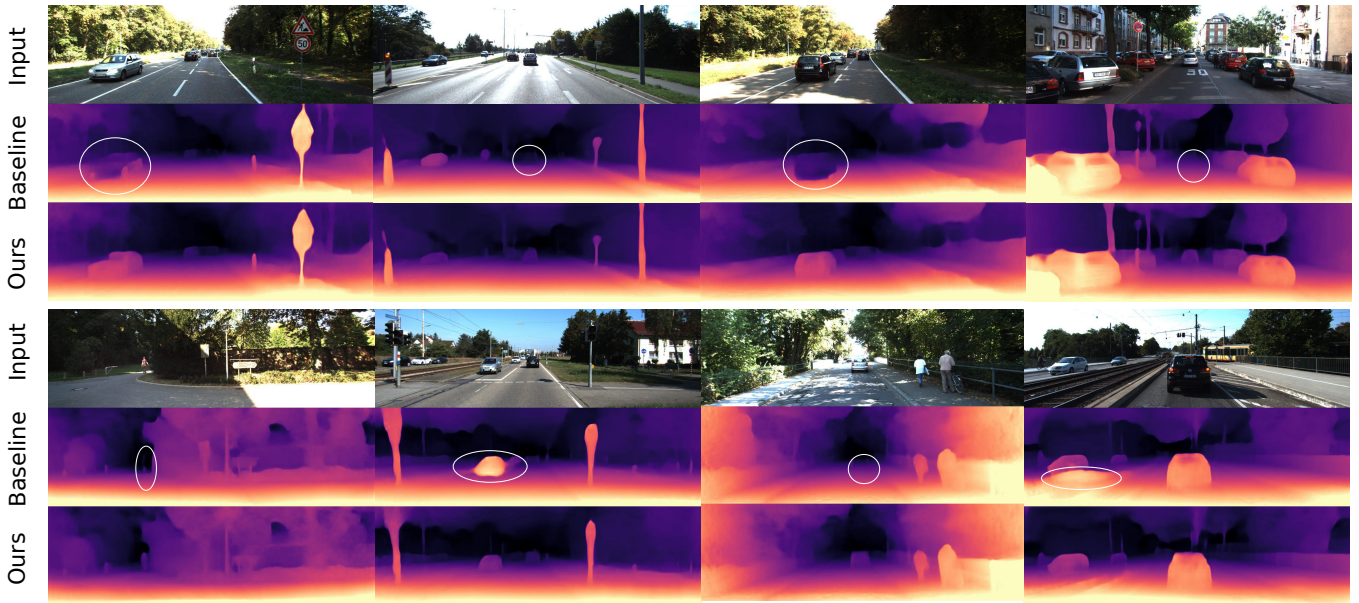


Fig. 5. Qualitative comparison of our method with Godard et al. [14] as the baseline. Every 3 rows show the input image, baseline depth, and ours. Despite the fact that the baseline method can handle moving objects e.g. cars most of the time. There are still some frames observed failing to predict the depth of moving objects while our method works reliably as marked in the circles.

equals to $T_{t \rightarrow s}$ for static objects with V set to identity matrix. Therefore, we propose the scale loss as follows:

$$\mathcal{L}_{scale} = |\text{tran}(T_{t \rightarrow s}) - \frac{1}{N} \sum_i \text{tran}(L_s^{(i)} L_t^{(i-1)})| \quad (16)$$

where N is the number of detected static objects and tran is the function to extract the translation part from a transformation matrix. The translation part of object pose changes is averaged over all static objects. Finally, the total loss function is

$$\mathcal{L}_{final} = \mathcal{L}_{photo} + \beta \mathcal{L}_{scale} \quad (17)$$

where \mathcal{L}_{photo} is from Equation (11) and β is a weighting constant set to 0.05. The scale loss part is effective to recover the true scale of the predictions without decreasing the network performance.

IV. EXPERIMENTS

A. Network Structure and Implementation Details

Our network is comprised of three subnetworks, DepthNet, PoseNet, and ObjectNet, as illustrated in Fig. 2. The three subnetworks are connected by the appearance loss between the reconstructed view and the target view. Specifically, the reconstructed view is segmented into two parts: static background and dynamic objects. The dynamic part is obtained by warping the source with D_t together with the detected object pose L by ObjectNet. While the static part is obtained by warping the source view D_t together with T predicted by PoseNet. The DepthNet and PoseNet are similar with [1]. The DepthNet is an auto-encoder like structure for single view depth estimation. It accepts a single color image and outputs inverse depth map $Disp_t$ with the same resolution. The inverse depth is converted to depth map D_t by $1/(a * Disp_t + b)$, with specifically chosen parameters

a and b to constrain the depth to be in the valid range $[0.1, 80]$ for the KITTI dataset [17]. The ResNet18 [18] with pre-trained weights on ImageNet [19] is used as the encoder part for feature extraction. The decoder part has 4 layers of building blocks each consisting of convolution and upsampling. The PoseNet accepts two temporally connected image sequence and outputs the 6 DoF relative camera pose $T_{t \rightarrow s}$. The 3D object detection network is similar to [11] which adopts DenseNet121 [20] as backbone feature extractor and followed by a Region Proposal Network (RPN) layers that use 3D anchor boxes to perform monocular 3D object detection.

Our proposed network is implemented with PyTorch [21] framework. The panoptic segmentation of color images are precomputed with [16]. The ObjectNet is also pretrained on KITTI *object* dataset [17]. Regarding data augmentation for DepthNet and PoseNet, we use common random horizontal flips and color augmentation with random brightness, contrast, saturation, and hue dithering, ranging from ± 0.2 , ± 0.2 , ± 0.2 and [14] proposed ± 0.1 . ObjectNet has random flipping disabled, but detected object poses are flipped when necessary. The network is trained for 15 epochs with a learning rate of 0.0001 and drops to 0.00001 for the last 5 epochs. Training takes about 48 hours on an Nvidia RTX 3090 24G graphic card. ObjectNet is frozen for the first 15 epochs to make training more stable. In the last 5 epochs, ObjectNet is fine-tuned together with DepthNet and PoseNet, excluding feature extraction layers to reduce computational cost.

B. Depth Evaluation on KITTI Dataset

There are total 61 scenes in KITTI raw dataset [17] consisting of city, residential and road. Following Eigen

et al. [22] and Zhou et al. [1], a total number of 39810 training images and 4424 validation images from both the left and right camera with static frames removed are used for training. The test set contains 697 images from different scenes from the training set. We compare our work with other networks using similar depth encoder layers. A qualitative comparison of depth estimation with baseline Monodepth2 [14] is given in Fig. 5. We can find that our network can better predict the depth of dynamic objects as marked in the figure. In some highly dynamic scenes, our network can predict reliable depth for moving objects while the baseline network fails in some frames. We also quantitatively evaluate the depth estimation results with standard metrics as defined in [22]. A detailed comparison with other state-of-the-art work is shown in Table I. Our method achieves improved performance without scaling the depth with median value, which is commonly adopted by other works. Following the convention on KITTI depth evaluation, only certain area $A = \{(x,y)|x \in (0.03594771, 0.96405229), y \in (0.40810811, 0.99189189)\}$ of meaningful points in the depth map is used when calculating all the metrics, where the pixel coordinates are normalized to $[0, 1]$ according to the depth map width and height.

C. Scale Recovery and Pose Evaluation

To verify the proposed network predicting the right scale in the depth map, we record the scale at each training step. The scale is defined as $scale = med(D_t^*) / med(D_t)$, where D_t^* and D_t are the ground truth depth map and the predicted depth map, respectively and med is the median value operator. The med is taken over the entire batch of inputs with batch size as 6. To ease the numerical convergence of the network, we scale the detected objects with 0.02, which constrains the right scale to be 50. As shown in Fig. 6, the predicted scale of our network is stabilized at 50 while the baseline network converges to the wrong scale.

We also performed an ablation study of how scale loss affects the depth estimation as shown in Table. II. We first train the network without scale loss and manually set the object scale to a fixed number. Not surprisingly, the results are even a little worse than baseline because the scales of the camera motion and the detected objects are inconsistent. The scale of camera motion (translation) is from PoseNet and the scale of object pose is from a separate ObjectNet. Without scale loss, the depths of rigid background and dynamic objects which are indirectly determined by the scales of camera motion and objects respectively are with different scales, hence leading to a worse result. With scale loss which enforces the predicted scales are consistent, our model is performed as expected.

D. Fine-Tuning for 3D Object Detection

The 3D object detection network is fine-tuned in a self-supervised manner with depth estimation. After 5 epochs of fine-tuning together with DepthNet and PoseNet, we observed improvements in object orientation angles, as shown

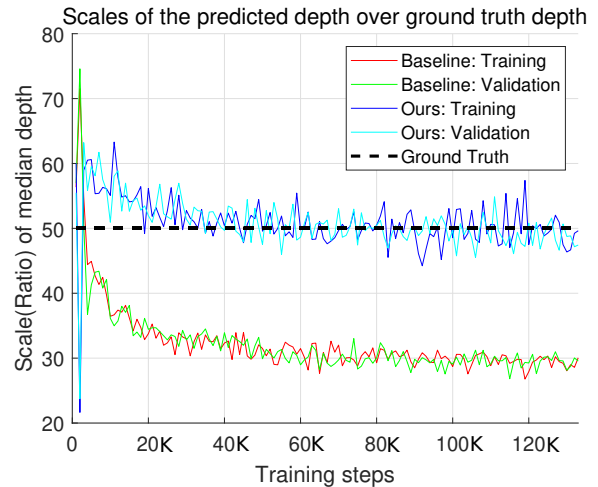


Fig. 6. The *scale* of the ground truth depth map and the predicted depth map. The predicted scale of our network converges to the true scale 50 marked with red dotted line after about 3 epochs while the baseline network converges around 30.



Fig. 7. We observed orientation angles refinements over some detected objects. Left column shows the cuboid detected before fine-tuning. Right columns show the refinements after fine-tuning.

in Fig. 7. Some objects which are partially occluded are detected more consistently on orientations. We also reevaluate the fine-tuned ObjectNet on the original KITTI 3D object dataset. The evaluation is conducted on the val1 [11] data splits. We compared the Average Precision (AP) [27] on 3 core tasks including Bird Eye View (BEV), 3D Object detection and 2D detection under 3 different settings: Easy, Moderate and Hard [27][17]. The AP is calculated with $IoU \geq 0.7$. All the metrics are improved after fine-tuning as shown in Table III. This shows the effectiveness of our proposed self-supervised training of monocular 3D object detection.

TABLE I

QUANTITATIVE RESULTS FOR DEPTH EVALUATION ON KITTI USING EIGEN SPLIT. ALL THE WORKS TO BE COMPARED ARE MONOCULAR METHODS WHICH ARE TRAINED WITH TEMPORAL IMAGES ONLY. WE FOLLOW THE EVALUATION METHOD PROVIDED BY [1] AND THE DEPTH IS CAPPED AT 80 METERS. ALL THE RESULTS IN THE TABLE ARE USING THE SAME EVALUATION METHOD WITHOUT FURTHER ONLINE REFINEMENTS.

| Method | Error Metrics | | | | Accuracy Metrics | | |
|------------------------------------|---------------|--------|-------|----------|------------------|-------------------|-------------------|
| | Abs Rel | Sq Rel | RMSE | RMSE log | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Zhou et al. [1] (SfMLearner) | 0.183 | 1.595 | 6.709 | 0.270 | 0.734 | 0.902 | 0.959 |
| Yang et al. [23] | 0.182 | 1.481 | 6.501 | 0.267 | 0.725 | 0.906 | 0.963 |
| Mahjourian et al. [2] | 0.163 | 1.240 | 6.220 | 0.250 | 0.762 | 0.916 | 0.968 |
| Zhan et al. [24] | 0.135 | 1.132 | 5.585 | 0.229 | 0.820 | 0.933 | 0.971 |
| Casser et al. [10] | 0.141 | 1.025 | 5.28 | 0.215 | 0.816 | 0.945 | 0.979 |
| Godard et al. [14] (Monodepth2) | 0.115 | 0.903 | 4.863 | 0.193 | 0.877 | 0.959 | 0.981 |
| Klingner et al. [9] (SGDepth full) | 0.113 | 0.835 | 4.693 | 0.191 | 0.879 | 0.961 | 0.981 |
| Guizilini et al. [25] | 0.117 | 0.854 | 4.714 | 0.191 | 0.873 | 0.963 | 0.981 |
| Ours | 0.115 | 0.831 | 4.734 | 0.190 | 0.877 | 0.960 | 0.983 |

TABLE II

ABLATION STUDY OF OUR MODEL ON KITTI [26] DATASET USING EIGEN SPLIT [22]. THE BASELINE IS ADOPTED FROM [14]. ALL THE METRICS ARE OBTAINED THROUGH THE EVALUATION METHOD [1]. NOTICE THAT THE PREDICTED DEPTHS ARE CAPPED AT 80 METERS.

| Method | Object Detection | Scale Loss | Joint Training | Error Metrics | | | | Accuracy Metrics | | |
|------------------------------|------------------|------------|----------------|---------------|--------|-------|----------|------------------|-------------------|-------------------|
| | | | | Abs Rel | Sq Rel | RMSE | RMSE log | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Ours baseline | | | | 0.115 | 0.903 | 4.851 | 0.193 | 0.877 | 0.959 | 0.981 |
| Ours (w/o scale loss) | ✓ | | | 0.126 | 0.879 | 4.799 | 0.198 | 0.844 | 0.957 | 0.982 |
| Ours (w/o joint fine-tuning) | ✓ | ✓ | | 0.116 | 0.844 | 4.758 | 0.190 | 0.871 | 0.960 | 0.983 |
| Ours (full) | ✓ | ✓ | ✓ | 0.115 | 0.831 | 4.734 | 0.190 | 0.877 | 0.960 | 0.983 |

TABLE III

DIFFERENT METRICS ON VAL1 DATA SPLIT FOR CARS BEFORE AND AFTER FINE-TUNING.

| Tasks | AP_{BEV} (IoU ≥ 0.7) | | | AP_{3D} (IoU ≥ 0.7) | | | 2D Detection | | |
|--------|------------------------------|--------|--------|-----------------------------|--------|--------|--------------|--------|--------|
| | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard |
| Before | 0.2416 | 0.1937 | 0.1589 | 0.1808 | 0.1447 | 0.1235 | 0.8865 | 0.8260 | 0.6719 |
| After | 0.2422 | 0.1844 | 0.1642 | 0.1846 | 0.1511 | 0.1292 | 0.8971 | 0.8357 | 0.6758 |

E. Ablation study

A detailed ablation study of our model is given in Table II. We adopt Monodepth2 [14] (monocular variation) as the baseline and gradually apply object detection, scale loss, and joint training technique to the baseline. We observed improvements when object detection is combined together with scale loss as discussed in Sec. IV-C. Furthermore, the joint training of depth prediction and object detection improves the results even more as proposed.

V. CONCLUSION

We have presented a novel method for handling dynamically moving objects in monocular depth estimation. By modeling dynamic points on objects with detected object poses, every pixel in the scene can be explained by camera motion or object pose changes. The benefits of introducing object detection for depth estimation are threefold. First, explicit modeling of dynamic points is shown to have better depth estimation performance. Second, static object poses can be used to regulate the ego-motion of the camera. The result shows that our network not only achieves state-of-the-art monocular depth estimation accuracy but also produces

the correct scale. Finally, the object detection network is also fine-tuned in a self-supervised manner, which is proven to benefit the original object detection network.

REFERENCES

- [1] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.
- [2] R. Mahjourian, M. Wicke, and A. 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*, 2018, pp. 5667–5675.
- [3] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding," in *Proc. European Conference on Computer Vision*, 2018, pp. 691–709.
- [4] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.
- [5] V. Guizilini, R. Ambruş, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2482–2491.
- [6] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Proc. European Conference on Computer Vision*, Amsterdam, Netherlands, Oct. 2016, pp. 740–756.

- [7] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.
- [8] L. Liu, G. Zhai, W. Ye, and Y. Liu, "Unsupervised learning of scene flow estimation fusing with local rigidity," in *IJCAI*, 2019, pp. 876–882.
- [9] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *European Conference on Computer Vision*, 2020, pp. 582–600.
- [10] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Unsupervised monocular depth and ego-motion learning with structure and semantics," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2019, pp. 381–388.
- [11] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9287–9296.
- [12] V. M. Babu, K. Das, A. Majumdar, and S. Kumar, "Undemon: Unsupervised deep network for depth and ego-motion estimation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1082–1088.
- [13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [14] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.
- [15] S. Wei, G. Chen, W. Chi, Z. Wang, and L. Sun, "Object clustering with dirichlet process mixture model for data association in monocular slam," *IEEE Transactions on Industrial Electronics*, pp. 1–1, 2022.
- [16] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *CVPR*, 2020.
- [17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.
- [22] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2366–2374, 2014.
- [23] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia, "Unsupervised learning of geometry with edge-aware depth-normal consistency," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- [24] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 340–349.
- [25] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," in *International Conference on Learning Representations*, 2019.
- [26] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [27] A. Simonelli, S. R. Buló, L. Porzi, M. López-Antequera, and P. Kotschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.