Supplementary Material for
“Multi-Object Discovery by Low-Dimensional Object Motion”

Sadra Safadoust Fatma Güney
KUIS AI Center and Department of Computer Engineering, Koç University
ssafadoust20@ku.edu.tr fguney@ku.edu.tr

In this supplementary document, we first provide the derivations of the basis for the space of possible optical flows in Section A. Then in Section B, we provide the details of the projection of the input flow into the space spanned by the bases. In Section C, we show additional qualitative results and in Section D, we provide an evaluation of our depth estimations for the foreground objects on MOVi datasets. Finally, in Section E, we show depth evaluation results for our model assuming known camera intrinsics.

A. Derivation of Basis

Assume that the world coordinate system coincides with the camera coordinate system and let $\mathbf{X} = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ denote the coordinates of a 3D point in the world. Assume that the scene is static and the camera is moving rigidly with angular velocity $\omega \in \mathbb{R}^3$ and linear velocity $v \in \mathbb{R}^3$, corresponding to the rotational and translational part of the motion. Then, following [4, 7], $\mathbf{X}'$, the instantaneous velocity of the point $\mathbf{X}$, can be calculated as follows:

$$\mathbf{X}' = \begin{bmatrix} \mathbf{x}' \\ \mathbf{y}' \\ \mathbf{z}' \end{bmatrix} = -(\omega \times \mathbf{X} + v) = \begin{bmatrix} \omega_3 \mathbf{y} - \omega_2 \mathbf{z} - v_1 \\ \omega_1 \mathbf{z} - \omega_3 \mathbf{x} - v_2 \\ \omega_2 \mathbf{x} - \omega_1 \mathbf{y} - v_3 \end{bmatrix} \tag{1}$$

Let f_x, f_y be the focal lengths and (c_x, c_y) denote the principal point of the camera. The pixel $\mathbf{p} = [u, v]^T$ corresponding to the 3D point $\mathbf{X}$ can be calculated as:

$$\mathbf{p} = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \mathbf{x} f_x / \mathbf{z} + c_x \\ \mathbf{y} f_y / \mathbf{z} + c_y \end{bmatrix} \tag{2}$$

Therefore, we can write:

$$\begin{aligned} \frac{\mathbf{x}}{\mathbf{z}} &= \frac{(u - c_x)}{f_x} = f_x^{-1} \bar{u} \\ \frac{\mathbf{y}}{\mathbf{z}} &= \frac{(v - c_y)}{f_y} = f_y^{-1} \bar{v} \end{aligned} \tag{3}$$

where we have defined $\bar{u} = u - c_x$ and $\bar{v} = v - c_y$. The instantaneous flow of a pixel $\mathbf{p}$ can be computed by taking derivatives of Eq. (2) with respect to time as follows:

$$\mathbf{p}' = \begin{bmatrix} u' \\ v' \end{bmatrix} = \frac{1}{\mathbf{z}^2} \begin{bmatrix} f_x (\mathbf{z}\mathbf{x}' - \mathbf{x}\mathbf{z}') \\ f_y (\mathbf{z}\mathbf{y}' - \mathbf{y}\mathbf{z}') \end{bmatrix} \tag{4}$$

By substituting the values from Eq. (1) into Eq. (4) we can write:

$$\begin{aligned} \begin{bmatrix} u' \\ v' \end{bmatrix} &= \frac{1}{\mathbf{z}^2} \begin{bmatrix} f_x \left(\mathbf{z} \left(\omega_3 \mathbf{y} - \omega_2 \mathbf{z} - v_1 \right) - \mathbf{x} \left(\omega_2 \mathbf{x} - \omega_1 \mathbf{y} - v_3 \right) \right) \\ f_y \left(\mathbf{z} \left(\omega_1 \mathbf{z} - \omega_3 \mathbf{x} - v_2 \right) - \mathbf{y} \left(\omega_2 \mathbf{x} - \omega_1 \mathbf{y} - v_3 \right) \right) \end{bmatrix} \\ &= \frac{1}{\mathbf{z}^2} \begin{bmatrix} f_x (-\mathbf{z} v_1 + \mathbf{x} v_3 + \mathbf{x}\mathbf{y}\omega_1 - (\mathbf{x}^2 + \mathbf{z}^2)\omega_2 + \mathbf{y}\mathbf{z}\omega_3) \\ f_y (-\mathbf{z} v_2 + \mathbf{y} v_3 + (\mathbf{y}^2 + \mathbf{z}^2)\omega_1 - \mathbf{x}\mathbf{y}\omega_2 - \mathbf{x}\mathbf{z}\omega_3) \end{bmatrix} \end{aligned} \tag{5}$$

By plugging the values from Eq. (3), and using disparity $d = 1/\mathbf{z}$, we can re-write Eq. (5) as:

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} -f_x\, d & 0 \\ 0 & -f_y\, d \\ \bar{u}\, d & \bar{v}\, d \\ {f_y}^{-1}\, \bar{u}\, \bar{v} & f_y + {f_y}^{-1}\, \bar{v}^2 \\ -(f_x + {f_x}^{-1}\, \bar{u}^2) & -{f_x}^{-1}\, \bar{u}\, \bar{v} \\ f_x\, {f_y}^{-1}\, \bar{v} & -f_y\, {f_x}^{-1}\, \bar{u} \end{bmatrix}^T \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} \tag{6}$$

Therefore, we can define a basis for the space of possible instantaneous optical flows for a given frame as

$$\mathcal{B}_0 = \{\mathbf{b}_{\mathbf{T}x}, \mathbf{b}_{\mathbf{T}y}, \mathbf{b}_{\mathbf{T}z}, \mathbf{b}_{\mathbf{R}x}, \mathbf{b}_{\mathbf{R}y}, \mathbf{b}_{\mathbf{R}z}\} \tag{7}$$

where we define:

$$\begin{aligned} \mathbf{b}_{\mathbf{T}x} &= \begin{bmatrix} f_x\, d \\ 0 \end{bmatrix}, & \mathbf{b}_{\mathbf{R}x} &= \begin{bmatrix} {f_y}^{-1}\, \bar{u}\, \bar{v} \\ f_y + {f_y}^{-1}\, \bar{v}^2 \end{bmatrix} \\ \mathbf{b}_{\mathbf{T}y} &= \begin{bmatrix} 0 \\ f_y\, d \end{bmatrix}, & \mathbf{b}_{\mathbf{R}y} &= \begin{bmatrix} f_x + {f_x}^{-1}\, \bar{u}^2 \\ {f_x}^{-1}\, \bar{u}\, \bar{v} \end{bmatrix} \\ \mathbf{b}_{\mathbf{T}z} &= \begin{bmatrix} -\bar{u}\, d \\ -\bar{v}\, d \end{bmatrix}, & \mathbf{b}_{\mathbf{R}z} &= \begin{bmatrix} f_x\, {f_y}^{-1}\, \bar{v} \\ -f_y\, {f_x}^{-1}\, \bar{u} \end{bmatrix} \end{aligned} \tag{8}$$

Our goal is to have basis vectors that do not depend on the values of focal lengths. Since basis vectors can be scaled arbitrarily, we can scale $\mathbf{b}_{\mathbf{T}x}$ and $\mathbf{b}_{\mathbf{T}y}$ by $1/f_x$ and $1/f_y$, respectively, to make them independent of f_x and f_y. By assuming $f_x = f_y$, $\mathbf{b}_{\mathbf{R}z}$ becomes $[\bar{v}, -\bar{u}]^T$ which is also free of focal lengths. We can write $\mathbf{b}_{\mathbf{R}x}$ and $\mathbf{b}_{\mathbf{R}y}$ as:

$$\begin{aligned} \mathbf{b}_{\mathbf{R}x} &= f_y \begin{bmatrix} 0 \\ 1 \end{bmatrix} + {f_y}^{-1} \begin{bmatrix} \bar{u}\,\bar{v} \\ \bar{v}^2 \end{bmatrix} \\ \mathbf{b}_{\mathbf{R}y} &= f_x \begin{bmatrix} 1 \\ 0 \end{bmatrix} + {f_x}^{-1} \begin{bmatrix} \bar{u}^2 \\ \bar{u}\,\bar{v} \end{bmatrix} \end{aligned} \tag{9}$$

Therefore, if we define:

$$\begin{aligned} \mathbf{b}_{\mathbf{R}^1x} &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \mathbf{b}_{\mathbf{R}^2x} &= \begin{bmatrix} \bar{u}\,\bar{v} \\ \bar{v}^2 \end{bmatrix} \\ \mathbf{b}_{\mathbf{R}^1y} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \mathbf{b}_{\mathbf{R}^2y} &= \begin{bmatrix} \bar{u}^2 \\ \bar{u}\,\bar{v} \end{bmatrix} \end{aligned} \tag{10}$$

we can replace $\mathbf{b}_{\mathbf{R}x}$ with the pair $\mathbf{b}_{\mathbf{R}^1x}$ and $\mathbf{b}_{\mathbf{R}^2x}$. Similarly, we replace $\mathbf{b}_{\mathbf{R}y}$ with the pair $\mathbf{b}_{\mathbf{R}^1y}$ and $\mathbf{b}_{\mathbf{R}^2y}$ [1]. Therefore, we can use the set of eight basis vectors

$$\mathcal{B}_0 = \{\mathbf{b}_{\mathbf{T}x}, \mathbf{b}_{\mathbf{T}y}, \mathbf{b}_{\mathbf{T}z}, \mathbf{b}_{\mathbf{R}^1x}, \mathbf{b}_{\mathbf{R}^2x}, \mathbf{b}_{\mathbf{R}^1y}, \mathbf{b}_{\mathbf{R}^2y}, \mathbf{b}_{\mathbf{R}z}\} \tag{11}$$

as a basis for the space of possible flows. Note that the space covered by this basis is actually slightly bigger because we cannot enforce the $f_x = f_y$ constraint in the decomposition of the rotational flows [1].

We normalize $\mathbf{b}_{\mathbf{T}x}, \mathbf{b}_{\mathbf{T}y}$, and $\mathbf{b}_{\mathbf{T}z}$ so that each vector has norm 2 before multiplication by d, and normalize $\mathbf{b}_{\mathbf{R}^1x}, \mathbf{b}_{\mathbf{R}^2x}, \mathbf{b}_{\mathbf{R}^1y}, \mathbf{b}_{\mathbf{R}^2y}$, and $\mathbf{b}_{\mathbf{R}^1z}$ to have norm 1.

B. Projection of Flow

We project input flow $\mathbf{F}$ into span$(\{\mathcal{B}_1 \cup \mathcal{B}_2 \cup \ldots \cup \mathcal{B}_K\})$ where each $\mathcal{B}_i$ is a set of 8 vectors defined as:

$$\mathcal{B}_i = \{\mathbf{m}_i \mathbf{b} \mid \mathbf{b} \in \mathcal{B}_0\}. \tag{12}$$

Consider an aribtrary ordering on the elements of $\mathcal{B}_i$ and define $\mathbf{v}_i{}^j$ as the j'th element in $\mathcal{B}_i$, reshaped into a $2HW$ vector. We define the matrix $\mathbf{S}_i \in \mathbb{R}^{2HW \times 8}$ as:

$$\mathbf{S}_i = \left[\mathbf{v}_i^1 \mid \mathbf{v}_i^2 \mid \ldots \mid \mathbf{v}_i^8\right] \tag{13}$$

Then, we define $\mathbf{S} \in \mathbb{R}^{2HW \times 8K}$ as follows:

$$\mathbf{S} = \left[\mathbf{S}_1 \mid \mathbf{S}_2 \mid \ldots \mid \mathbf{S}_K\right] \tag{14}$$

We calculate the singular value decomposition of $\mathbf{S}$:

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{15}$$

The columns of $\mathbf{U}$ corresponding to non-zero singular values of $\mathbf{S}$ span the column space of $\mathbf{S}$, i.e. span$(\{\mathcal{B}_1 \cup \mathcal{B}_2 \cup \ldots \cup \mathcal{B}_K\})$. Since the columns of $\mathbf{U}$ form an orthonormal set, we can project $\mathbf{F}$ into the column space of $\mathbf{S}$ as follows:

$$\hat{\mathbf{F}} = \mathbf{U}'\mathbf{U}'^T\mathbf{F} \tag{16}$$

where $\mathbf{U}'$ is the matrix whose columns are the columns of $\mathbf{U}$ corresponding to non-zero singular values of $\mathbf{S}$. In practice, we select columns of $\mathbf{U}$ that correspond to singular values larger than 10^{-5}.

C. Additional Qualitative Results

Qualitative results for CLEVR and ClevrTex datasets are provided in Fig. 1. We show additional visualizations, including the post-processing results for MOVi datasets in Fig. 2. We can see that PPMP [5] suffers from over-segmentation, with or without post-processing, especially in the MOVi datasets, whereas our method achieves much better results, as reflected in the quantitative performance. Our results for the KITTI dataset are visualized in Fig. 3. It can be seen that we can segment objects such as cars and pedestrians successfully. We also visualize the depth estimations of our model.

D. Depth Evaluation on MOVi

In this section, we evaluate the performance of our depth model on the foreground objects in each of the MOVi datasets. We evaluate the performance for both the Full model and the translation-only (Only-T) model. Note that as explained in the main paper, with the rotation-only model, the depth network is not trained. The results are presented in Table 1. We only evaluate the foreground objects. We use the median scaling approach [12] to convert the predicted depths into the metric scale and cap the depths to 10 meters in all datasets.

The depth network of the Only-T models achieves better results on the MOVi{C, D, E} datasets. This is expected because the depth and segmentation networks are trained jointly. As a result, in the Full model, the errors in the estimation of rotation affect the depth estimations negatively, whereas, in the Only-T model, the depth estimations are not affected by erroneous rotation estimations. However, In the simpler MOVi-A dataset, as explained in the main paper, we found that the depth network in the Only-T model cannot predict the depth correctly. Therefore, we did not include the results of this version of the model for MOVi-A.

E. KITTI with Intrinsics

In our formulation, we produce basis vectors that do not depend on the values of focal lengths f_x and f_y, which results in a set of 8 basis vectors as in Eq. (11), instead of 6 as in Eq. (7), for each of K regions. This means that our method can work without knowing the values of f_x and f_y. On the other hand, monocular depth estimation methods assume a known intrinsics matrix, i.e. f_x, f_y, c_x, and c_y are provided in the dataset. In order to make a fair comparison with monocular depth estimation methods, we train a version of our model on KITTI, where we also assume a

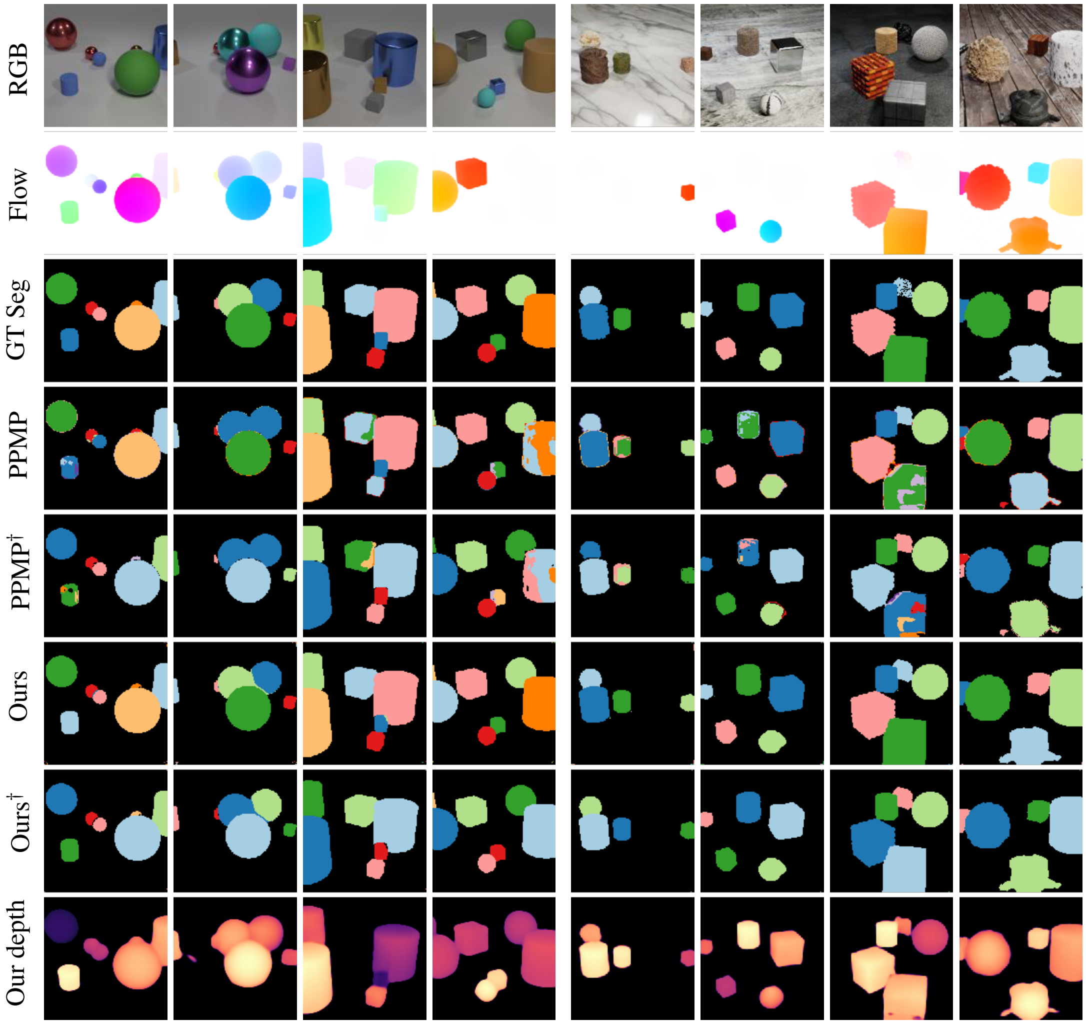

Figure 1: **Visualization of Depth and Segmentation Results on CLEVR and CLEVRTEX datasets**. The first four columns are from CLEVR, and the last four columns are from CLEVRTEX. † indicates post-processing.

Dataset	Abs Rel ↓	Sq Rel ↓	RSME ↓	RMSE log ↓	log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
MOVi-A (Full)	0.113	0.348	1.483	0.226	0.061	0.813	0.912	0.949
MOVi-A (Only-T)	-	-	-	-	-	-	-	-
MOVi-C (Full)	0.225	0.604	1.845	0.299	0.100	0.609	0.856	0.946
MOVi-C (Only-T)	0.166	0.446	1.437	0.217	0.068	0.779	0.932	0.978
MOVi-D (Full)	0.544	2.863	3.744	1.381	0.415	0.348	0.547	0.657
MOVi-D (Only-T)	0.357	1.598	2.603	0.730	0.225	0.540	0.757	0.847
MOVi-E (Full)	0.274	1.198	2.965	0.582	0.162	0.565	0.747	0.829
MOVi-E (Only-T)	0.244	0.989	2.596	0.559	0.159	0.596	0.761	0.842

Table 1: **Depth Evaluation on Foreground Objects on MOVi Datasets**. Only-T refers to the version of our model where we only use the basis vectors corresponding to translation.

known intrinsics matrix and create 6 basis vectors according to Eq. (7) for each region using the values of known focal lengths. We report the depth estimation results with improved ground truth on the Eigen split of the KITTI dataset

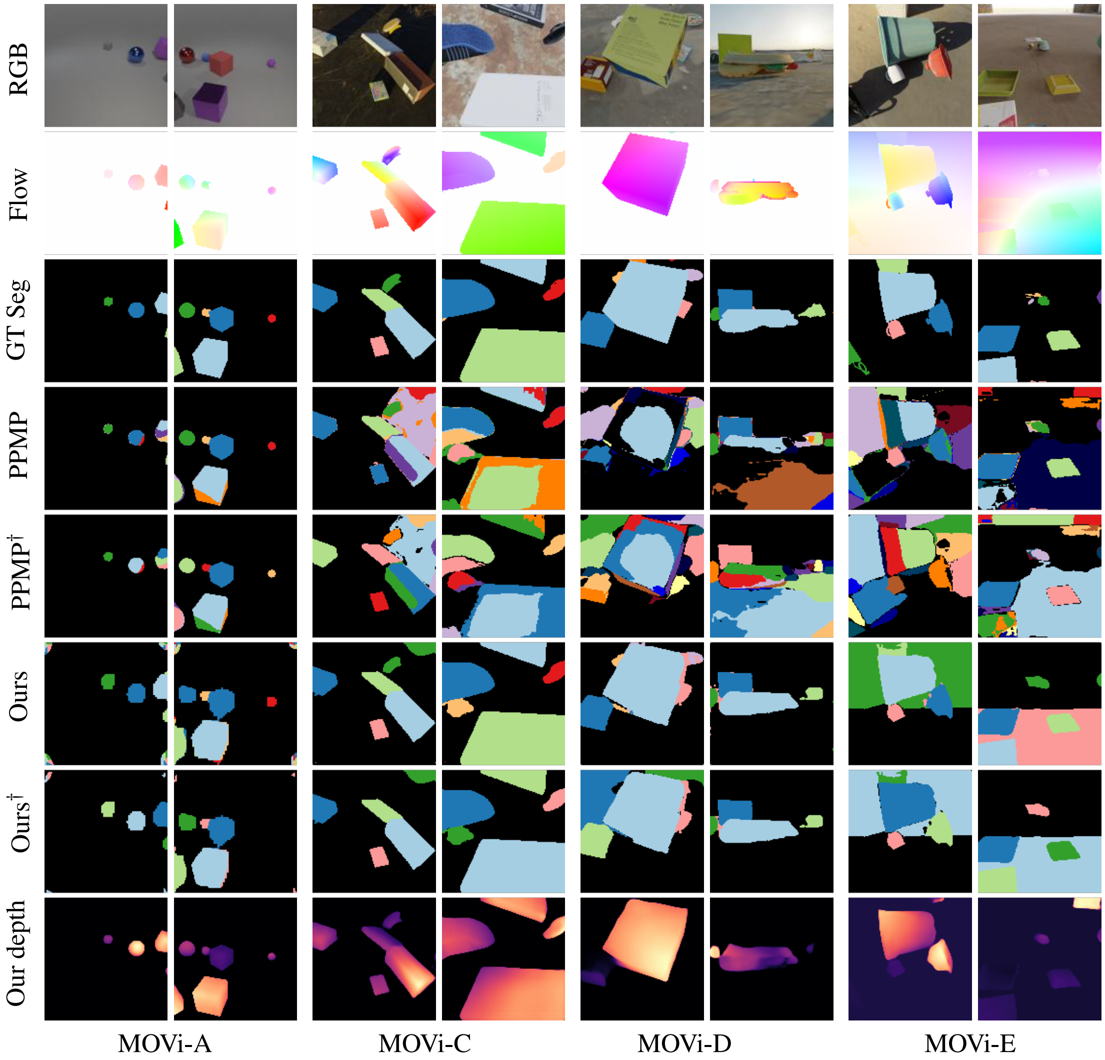

Figure 2: **Visualization of Depth and Segmentation Results on MOVi datasets**. † indicates post-processing.

in Table 2. We can see that when we use a known camera intrinsic matrix (Ours-intrinsics), the performance is improved compared to our original model (Ours), and we can achieve better results that are comparable to the state-of-the-art in all metrics.

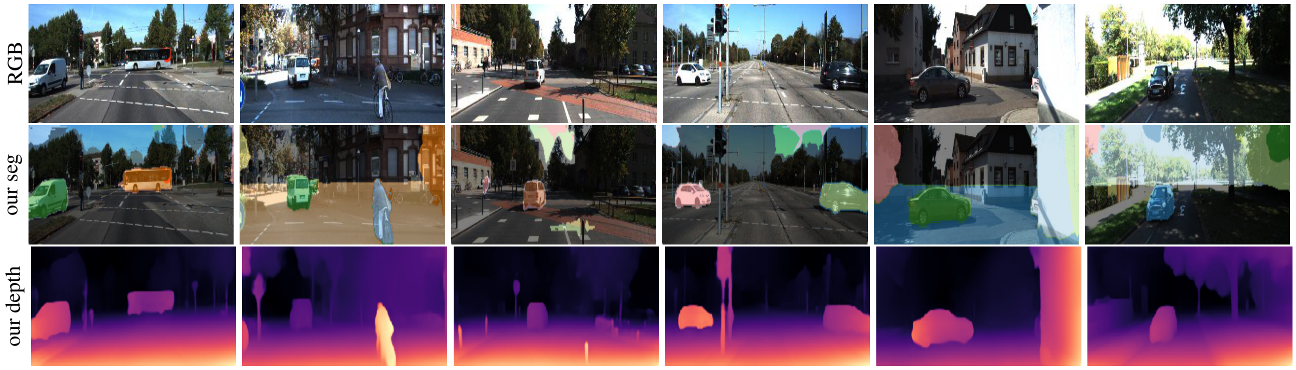

Figure 3: **Visualization of Segmentation and Depth Results on KITTI**.

	Abs Rel ↓	Sq Rel ↓	RSME ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Zhou et al. [12]	0.176	1.532	6.129	0.244	0.758	0.921	0.971
Mahjourian et al. [8]	0.134	0.983	5.501	0.203	0.827	0.944	0.981
GeoNet [11]	0.132	0.994	5.240	0.193	0.833	0.953	0.985
DDVO [10]	0.126	0.866	4.932	0.185	0.851	0.958	0.986
Ranjan et al. [9]	0.123	0.881	4.834	0.181	0.860	0.959	0.985
EPC++ [6]	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Ours	0.107	1.539	4.027	0.149	0.911	0.971	0.989
Monodepth2 [2]	0.090	0.545	3.942	0.137	0.914	0.983	**0.998**
Ours-intrinsics	0.084	0.509	**3.450**	0.132	0.931	0.980	0.993
PackNet-SfM [3]	**0.078**	**0.420**	3.485	**0.121**	**0.934**	**0.986**	0.996

Table 2: **Evaluation of Depth Estimation on KITTI.** We use the Eigen split of KITTI using improved ground truth. Note that all methods, except Ours, use the camera intrinsics matrix. Ours-intrinsics uses the intrinsics matrix and achieves comparable performance to state-of-the-art methods.

References

[1] Richard Strong Bowen, Richard Tucker, Ramin Zabih, and Noah Snavely. Dimensions of motion: Monocular prediction through flow subspaces. In *3DV*, 2022. 2

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 5

[3] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 5

[4] David J. Heeger and Allan D. Jepson. Subspace methods for recovering rigid motion I: Algorithm and implementation. *IJCV*, 1992. 1

[5] Laurynas Karazija, Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised Multi-object Segmentation by Predicting Probable Motion Patterns. In *NeurIPS*, 2022. 2

[6] 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. *PAMI*, 42(10):2624–2641, 2019. 5

[7] Yi Ma, Stefano Soatto, Jana Kosecka, and S. Shankar Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. SpringerVerlag, 2003. 1

[8] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018. 5

[9] 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 *CVPR*, 2019. 5

[10] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 5

[11] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 5

[12] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2, 5