Self-Supervised Monocular Scene Decomposition and Depth Estimation

Sadra Safadoust, Fatma Güney

KUIS AI Center, Koç University

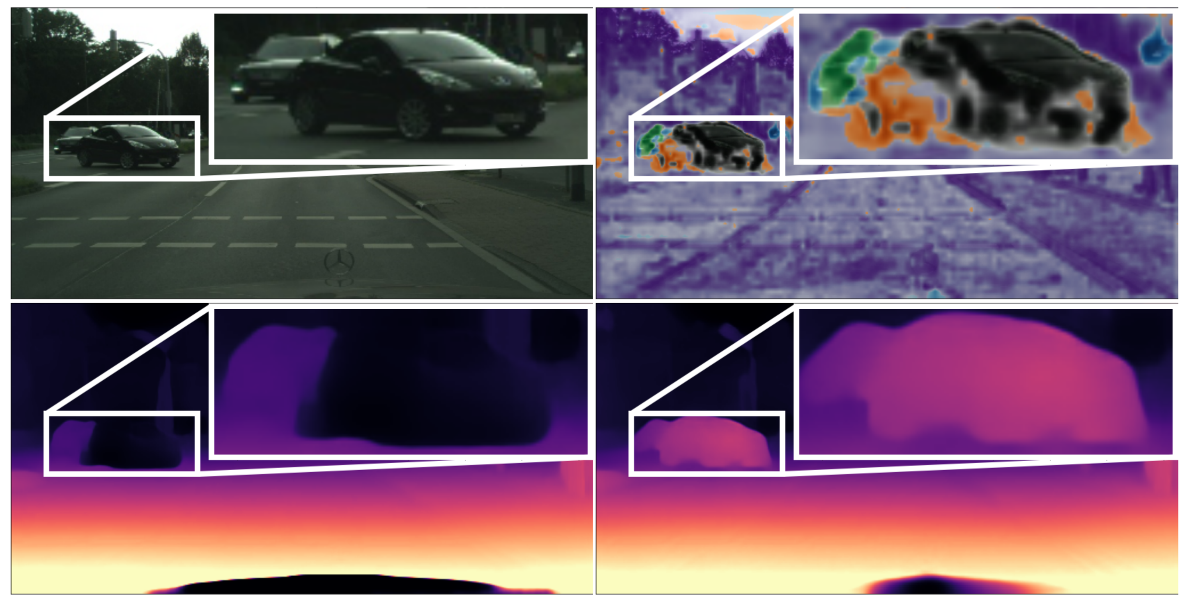
Self-Supervised Monocular Depth

Current monocular depth estimation methods

- either assume a static scene and fail in foreground regions with independently moving objects
- or require a separate segmentation step to identify the dynamic objects in the foreground.

MonoDepthSeg

We introduce **MonoDepthSeg** to jointly estimate depth and segment moving objects from monocular video without using any ground-truth labels.



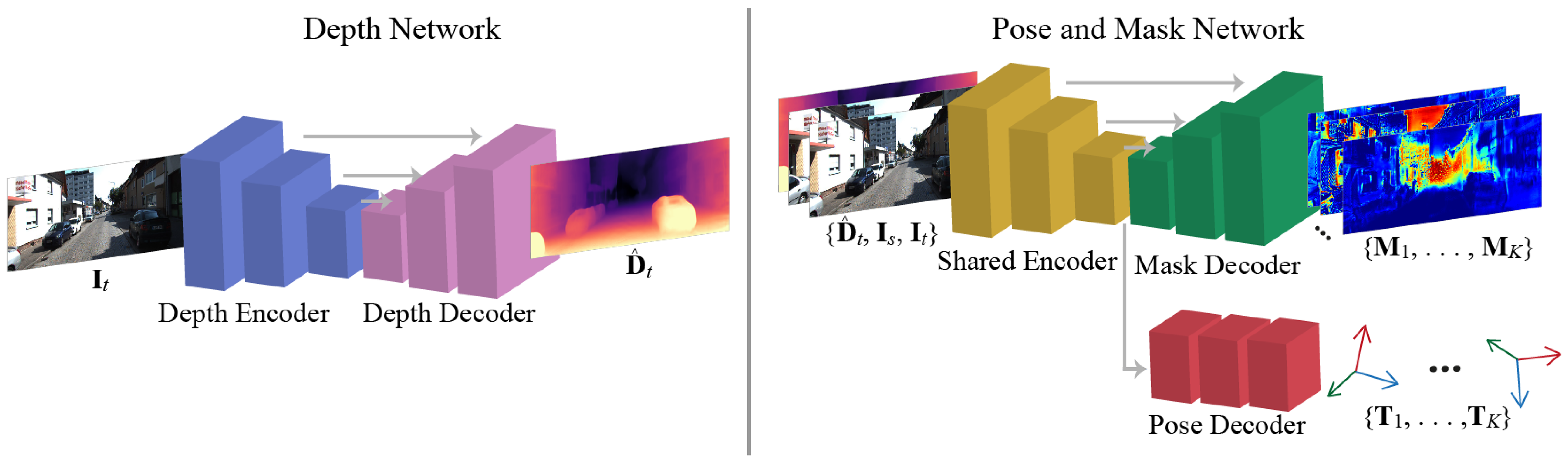
We decompose the scene into a fixed number of components where each component corresponds to a region on the image with its own transformation matrix representing its motion. This improves the results in regions with moving objects (bottom-right) compared to the current approaches [1] (bottom-left), while simultaneously recovering a decomposition of the scene, mostly corresponding to moving regions (top-right).

Our framework consists of

- depth network to estimate per-pixel depth values
- pose and mask network to divide the image into components and estimate a separate pose for each component.

The two networks are trained jointly and end-to-end.

Model Architecture



- Given a single target image $\mathbf{I}_t$, the depth network outputs the depth estimate $\hat{\mathbf{D}}_t$.
- The shared encoder maps the two consecutive input frames $\mathbf{I}_s$ and $\mathbf{I}_t$ and the depth estimate $\hat{\mathbf{D}}_t$ to a common representation.
- The mask decoder produces the same resolution K masks $\{\mathbf{M}_1, \ldots, \mathbf{M}_K\}$.
- The pose decoder maps the same encoded representation into rigid transformations $\{\mathbf{T}_1, \ldots, \mathbf{T}_K\}$ corresponding to the masks.

Methodology

- We represent the motion of each component using a 3D rigid transformation $\mathbf{T} = [\mathbf{R}, \mathbf{t}] \in \mathbf{SE}(3)$
- We encourage masks to be layered according to a pre-defined depth order d_i. This helps to account for occlusions:

$$\mathbf{M}_i(\mathbf{p}) = \frac{e^{d_i \mathbf{M}'_i(\mathbf{p})}}{\sum_{j=1}^{K} e^{d_j \mathbf{M}'_j(\mathbf{p})}}$$

- For every pixel $\mathbf{p}$ on the target image $\mathbf{I}_t$, we compute the corresponding 3D point $\mathbf{x}$ using its depth value $\hat{\mathbf{D}}_t$ and the intrinsic camera matrix $\mathbf{K}$:

$$\mathbf{x} = \hat{\mathbf{D}}_t(\mathbf{p})\ \mathbf{K}^{-1}\ \mathbf{p}$$

- We transform the 3D point $\mathbf{x}$ using the masks and rigid transformations to obtain $\mathbf{x}'$:

$$\mathbf{x}' = \sum_{i=1}^{K} \mathbf{M}_i(\mathbf{p})\ \mathbf{T}_i\ \mathbf{x} = \sum_{i=1}^{K} \mathbf{M}_i(\mathbf{p})\ (\mathbf{R}_i\ \mathbf{x} + \mathbf{t}_i)$$

- We project the transformed point $\mathbf{x}'$ to find the corresponding point $\mathbf{p}'$ on the source image $\mathbf{I}_s$:

$$\mathbf{p}' = \mathbf{K}\ \mathbf{x}'$$

- We reconstruct the target image $\mathbf{I}_t$ by sampling pixels from the source image $\mathbf{I}_s$ and obtain the warped image $\hat{\mathbf{I}}_s$, such that $\hat{\mathbf{I}}_s(\mathbf{p}) = \mathbf{I}_s(\mathbf{p}')$.
- We use an edge-aware smoothness loss $\mathcal{L}_{smooth}$ over the mean-normalized inverse depth values and define the photometric loss as follows:

$$\mathcal{L}_{\text{photo}}(\mathbf{p}) = \min_{s} \Big[(1-\alpha)\, |\mathbf{I}_t(\mathbf{p}) - \hat{\mathbf{I}}_s(\mathbf{p})| + \frac{\alpha}{2} \Big(1 - \text{SSIM}(\mathbf{I}_t, \hat{\mathbf{I}}_s)(\mathbf{p})\Big) \Big]$$

- Our final loss is defined as:

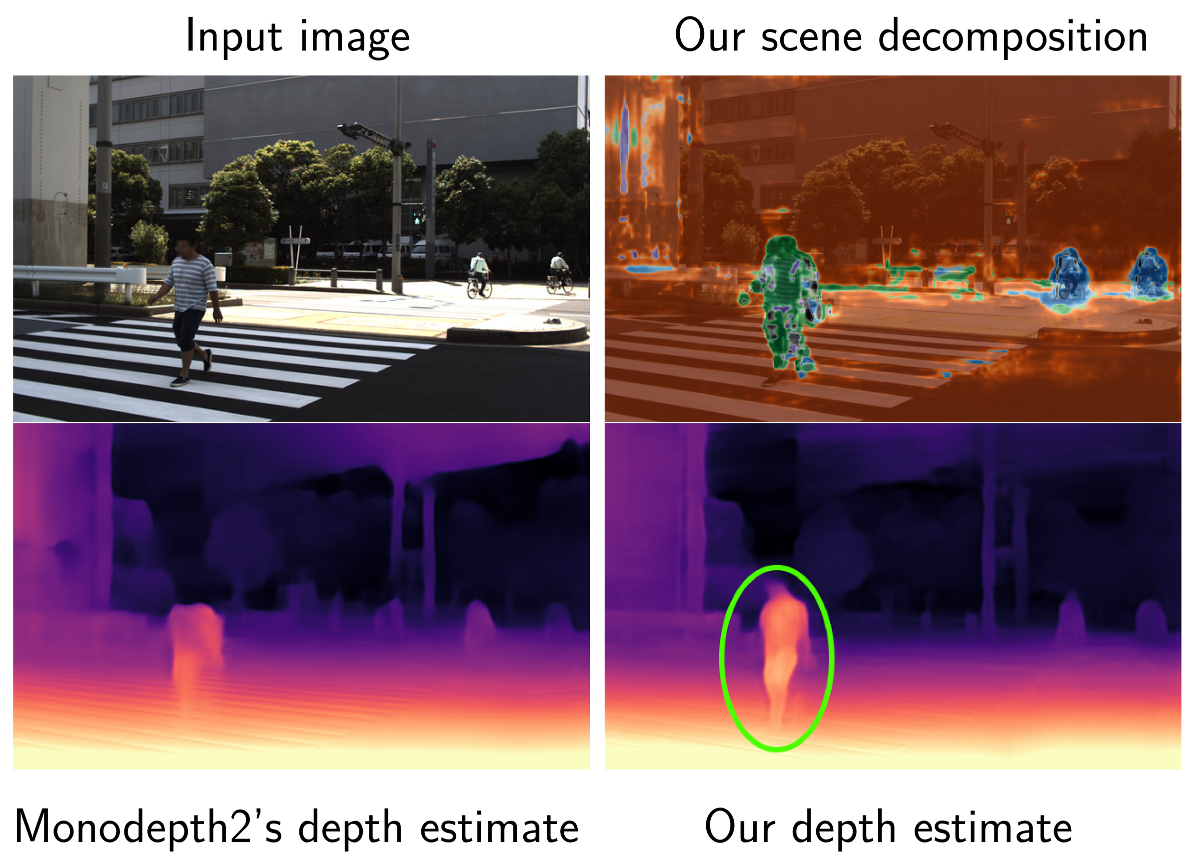
$$\mathcal{L} = \frac{1}{N} \sum_{\mathbf{p}} \mathcal{L}_{photo}(\mathbf{p}) + \lambda\ \mathcal{L}_{smooth}(\mathbf{p})$$

Quantitative Results

	Method	Abs Rel	RMSE	Car	Person
DDAD	PackNet [2]	0.23	17.92	0.38	0.20
	Monodepth2 [1]	0.22	17.63	0.25	0.21
	Ours	**0.19**	**16.61**	**0.24**	**0.17**

	Method	Abs Rel		RMSE	
		Moving	All	Moving	All
CityScapes	Monodepth2 [1]	0.158	0.170	8.043	8.155
	Ours	**0.143**	**0.142**	**7.649**	**7.361**
KITTI	Monodepth2 [1]	0.143	**0.110**	5.949	**4.642**
	Ours	**0.138**	**0.110**	**5.796**	4.700

Qualitative Results



References

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

[2] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," 2020.

Contact Information

Web: https://kuis-ai.github.io/monodepthseg

Email: {ssafadoust20, fguney}@ku.edu.tr