Impact of Pseudo Depth on Open World Object Segmentation with Minimal User Guidance (Supplementary Material)

Robin Schön Katja Ludwig Rainer Lienhart
Chair for Machine Learning and Computer Vision, University of Augsburg
{robin.schoen, katja.ludwig, rainer.lienhart}@uni-a.de

1. Unsupervised Monocular Depth Estimation from Video Data

This section aims at acquainting the reader with the general idea of unsupervised monocular depth estimation. This task consists in the utilization of unannotated video data with the purpose of training a network for the task of monocular depth estimation. It should be explicitly mentioned, that despite the necessity of video data during training, the resulting depth estimation network will be trained to predict depth maps for single images. This renders the depth estimator useful for downstream tasks on images, which is especially useful for our purposes. Although we mention the existence of a considerable amount of literature (see [1–3, 5–9]), we will specifically use the MonodepthV2 framework as described in [2].

Despite certain specific differences, most of these training strategies are based on the same principle:

- In a video sequence, we assume frames which are temporally close to each other to display the same scene.
- We predict the depth map of one of the two frames, as well as their relative camera pose.
- We use these predictions and the intrinsic camera parameters to project one image onto the other.
- In order to train the depth prediction network and the pose prediction network, we compute a simple photometric loss between the warped and the target image.

This training strategy is visualized in Figure 1 and more closely explained in the following text.

Let t and t' be two close points in time in a video (e.g. at 3 frames distance). We assume that the corresponding frames I_t and $I_{t'}$ display the same partially static scene, from two slightly different points of view. In order to warp the image $I_{t'}$ onto I_t, we will need the relative pose between the images. Since we only have single camera at our disposal (instead of two cameras with a fixed known distance), we will have to guess the relative camera position in the form of a rotation and a translation. This subtask is carried out by the means of a relative pose estimation network which outputs the transformation

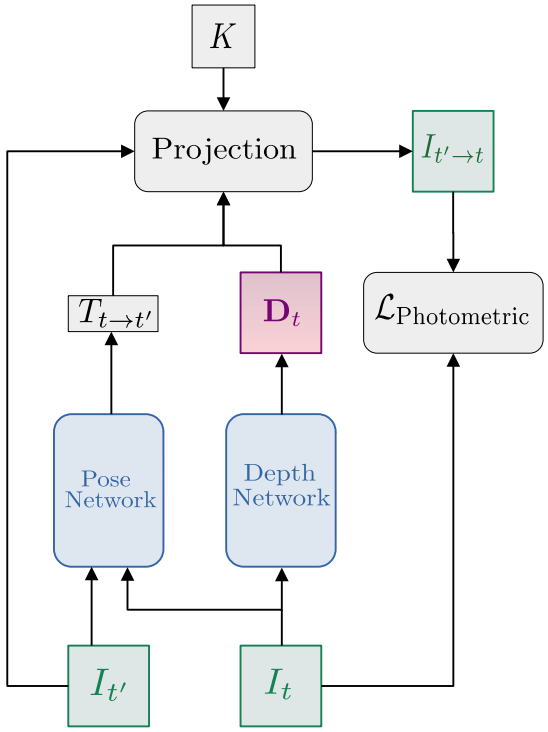

Figure 1. The pose network computes the relative pose $T_{t\to t'}$ between the images I_t and $I_{t'}$. The depth network computes the depth map $\mathbf{D}_t$. Together with the intrinsic camera parameters K, we can use these estimates to warp $I_{t'}$ onto I_t obtaining $I_{t\to t'}$. The two images I_t and $I_{t\to t'}$ are then compared by the means of a photometric loss (L1 loss and SSIM).

$$T_{t\to t'} = g(I_t, I_{t'}) \tag{1}$$

between the two frames. More specifically, $T_{t\to t'}$ denotes the rotation and subsequent translation which are necessary to get from a point in I_t to the corresponding point in $I_{t'}$.

The second piece of information necessary for warping task will be the depth map

$$\mathbf{D}_t = h(I_t) \tag{2}$$

which is predicted by the depth network h. As can be observed, the depth prediction only ever happens on single images, rendering the resulting trained network viable for single image input. Both networks, g and h, will be trained jointly. The third necessary ingredient are the intrinsic camera parameters K, which are assumed to be known beforehand.

We can then warp the image $I_{t'}$ onto I_t, obtaining

$$I_{t'\to t} = I_{t'} \left\langle \text{Projection}(\mathbf{D}_t, T_{t\to t'}, K) \right\rangle \tag{3}$$

where $\langle\cdot\rangle$ denotes a differentiable sampling operator and $I_{t'\to t}$ is the result of warping the image $I_{t'}$ onto I_t. In order for the sampling operation to be differentiable (see [4, 9]), the pixel values are a linear interpolation of the four closest pixels at integer positions in the image from which we sample. The projection operation effectively allows us to compute for each coordinate p_t in I_t the corresponding pixel position $p_{t\to t'}$ in $I_{t'}$. This transformation of coordinates can be formulated (see [9]) as

$$p_{t\to t'} \sim K T_{t\to t'} \mathbf{D}_t K^{-1} p_t. \tag{4}$$

We now compare the two images $I_{t\to t'}$ and I_t by the means of photometric loss, that expresses their difference. Minimizing this difference during training will imply the improvement of the two network-predicted components in this computation: the depth map and the relative pose. In our case of MonodepthV2 the image difference is expressed as

$$\begin{aligned}\mathcal{L}_{\text{Photometric}} = &\frac{\alpha}{2}(1 - \text{SSIM}(I_t, I_{t'\to t})) \\ &+ (1-\alpha)||I_t - I_{t'\to t}||_1,\end{aligned} \tag{5}$$

where SSIM denotes the structural similarity index measure and $||\cdot||_1$ is a simple L1 loss. The authors of [2] set the relative loss weight α to 0.85. MonodepthV2 specifically uses an additional gradient based smoothing loss, multiple neighbouring frames, and a masking scheme for pixel positions on which the feedback is deemed to be of insufficient quality. A detailed explanation of these auxiliary techniques would, however, go beyond the scope of conveying the general training idea.

After training, the depth network is used in isolation for the purpose of obtaining pseudo depth maps on new images.

References

[1] Vincent Casser, Sören Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI*, 2019. 1

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 1, 2

[3] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2

[5] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*, 2020. 1

[6] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[7] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 1

[8] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6871–6880, 2019. 1

[9] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 1, 2