

Multi-view Monocular Depth and Uncertainty Prediction with Deep SfM in Dynamic Environments^{*}

Christian Homeyer^{1,2}[0000–0002–0953–5162], Oliver Lange¹[0000–0001–7461–6869], and
Christoph Schnörr²[0000–0002–8999–2338]

¹ Robert Bosch GmbH, Hildesheim, Germany {christian.homeyer}@bosch.com

² IPA Group, Heidelberg University, Germany

Abstract. 3D reconstruction of depth and motion from monocular video in dynamic environments is a highly ill-posed problem due to scale ambiguities when projecting to the 2D image domain. In this work, we investigate the performance of the current State-of-the-Art (SotA) deep multi-view systems in such environments. We find that current supervised methods work surprisingly well despite not modelling individual object motions, but make systematic errors due to a lack of dense ground truth data. To detect such errors during usage, we extend the cost volume based Deep Video to Depth (DeepV2D) framework [70] with a learned uncertainty. Our Deep Video to certain Depth (DeepV2cD) model allows i) to perform en par or better with current SotA and ii) achieve a better uncertainty measure than the naive Shannon entropy. Our experiments show that a simple filter strategy based on the uncertainty can significantly reduce systematic errors. This results in cleaner reconstructions both on static and dynamic parts of the scene.

Keywords: Deep learning · Structure-from-Motion · Uncertainty prediction · Depth prediction · Visual odometry · Monocular video · Supervised learning.

1 Introduction

Reconstruction of scene geometry and camera motion is an important task for autonomous driving and related downstream tasks, e.g. collision avoidance and path planning. In particular, reconstruction from *monocular* videos has become an important research direction, due to the possibility of low-cost realizations in concrete autonomous systems.

Traditional non-learning algorithms [77] either fail to resolve the inherent ambiguities resulting from moving objects or do not scale well for dense predictions [3]. *Deep learning* based systems can resolve these shortcomings, but their accuracy is dependent on architecture and learning strategy. Depth reconstruction from a single image [17,48,44,42,18] is an even more challenging problem, as no motion cues can act as input. *Single-view* networks may not generalize well from one dataset to another [48,14] or need to be trained on massive datasets [59,58]. Using a true *multi-view* learning approach [72,69,70,75,23] turned out to be favorable in terms of performance. An

^{*} Supported by Robert Bosch GmbH.

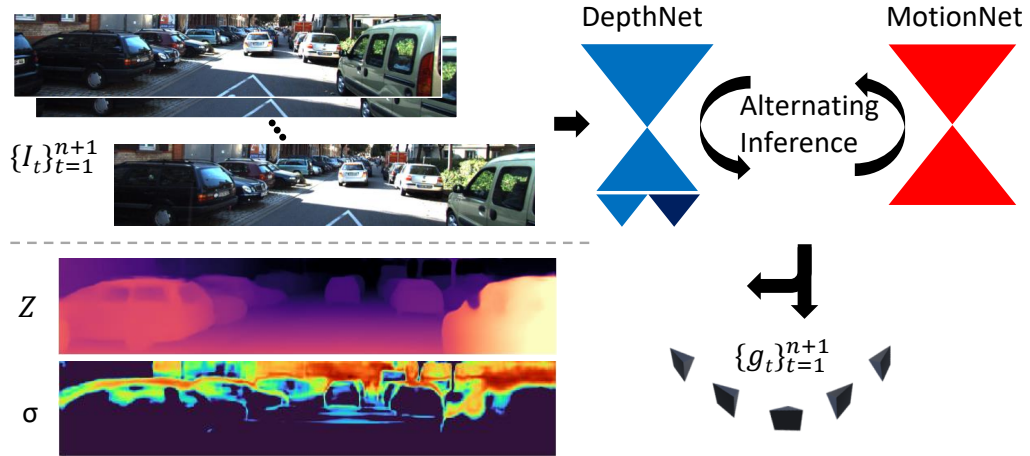


Fig. 1: Based on an image sequence, we predict camera pose graph, depth and uncertainty. Warm colors indicate high uncertainty. The uncertainty prediction helps to identify falsely reconstructed regions, which typically occur due to missing lidar supervision.

unsupervised learning strategy can leverage large-scale datasets without gathering expensive depth groundtruth. While unsupervised learning in this area has seen impressive progress [19,21,22,24,75,23], accurate depth on dynamic objects cannot be learned when objects move colinear to the camera, since the photometric loss cannot resolve the underlying ambiguity [84]. *Supervised* training can leverage a signal for these cases, which is why we base our work on a supervised multi-view framework.

Recent developments have shown, that modeling scene geometry explicitly inside the architecture [70,69,23] leads to better reconstruction results than loosely coupling neural networks with a common training loss. Inside this paradigm, learned features are aligned temporally and spatially based on scene geometry and camera motion. Our evaluation in Section 5 indicates, that depth prediction works surprisingly well for dynamic objects, even though a static scene is assumed during this explicit modelling. However, we demonstrate quantitatively and qualitatively that reconstructions of dynamic objects are systematically worse than the static scene on real world autonomous driving data, e.g. the KITTI dataset. Our findings show, that the root cause is mostly a lack of supervision from the lidar ground truth. Lidar density is significantly lower on moving objects than on the rest of the scene and is missing completely near depth discontinuities or on transparent surfaces, e.g. windows.

We extend the supervised DeepV2D framework [70] by learning an additional uncertainty and show how this can identify gross outliers in these regions. An overview of our DeepV2cD framework can be seen in Figure 1: based on a sequence of images we compute a) a camera pose graph b) the key frame depth and c) an uncertainty estimate. Because the depth network utilizes a cost volume based architecture, this extension comes with no additional overhead. In fact, any cost volume based approach main-

tains a probability volume, such that an uncertainty measure can be readily computed [47,86]. However, our experiments indicate, that by adding a separate uncertainty head to our network and learning a measure, we achieve superior performance compared to the Shannon entropy as done in related work [78,47,41]. Training this uncertainty head requires an additional loss term, which we adapt from [86]. Our training strategy results in slightly better final accuracy and has the benefit of an uncertainty.

An alternative approach to dealing with missing groundtruth in real data, is the use of synthetic data. Advances in datasets [52,15,51,7] allow training with a dense and accurate ground truth before finetuning unsupervised on sparse real data [83]. As a final contribution we show, that the Deepv2D networks generalize well to the real world even when only trained on virtual data. The domain gap can be further reduced by finetuning with a semi-supervised loss on the real data without the need for expensive depth ground truth. We make a comparison of current unsupervised, weakly supervised and supervised SotA on the KITTI and Cityscapes dataset.

2 Related work

Reconstruction from monocular image sequences has a large body of literature. We refer interested readers to an in-depth survey with focus on dynamic scenes [63], that is beyond the scope of this paper. We distinguish between *traditional* methods and *learning* based systems. Learned methods can be distinguished into loosely coupled individual networks with *single-view* depth prediction and true *multi-view* systems. Finally, we give an overview of methods for *uncertainty* prediction for this problem.

Traditional. Early work in SfM worked only on a small collection of images [49]. Traditionally, the problem is solved by finding correspondences with hand-crafted features, then solving for motion and structure. Today, several mature full pipelines exist [77,67,65], that improved considerably w.r.t scalability, robustness and accuracy compared to early systems. The reconstruction of monocular images is only possible to a common relative scale, due to a missing absolute scale. Traditional SfM relies on a static scene assumption. By filtering out dynamic points as outliers, they do not perform a full scene reconstruction. Several non-learning approaches target these shortcomings: Avidan et al. [4,66,3] coined the term trajectory triangulation by extending the epipolar constraint to the general case of a moving point with a specific trajectory. Park et al. [2,55,56] show how it is possible to decompose motions into a linear combination of basis trajectories using the Discrete Cosine Transform (DCT). However, these approaches only perform sparse reconstruction and require knowing the dynamic/static assignment to be known in advance. Another line of work relies on the piecewise-planar assumption and dense reconstruction based on optical flow [60,37,38]. Still, especially colinearly moving objects cannot be distinguished from observed background and result in wrong reconstructions [84].

Learning geometry and motion. Early work on learning reconstructions used e.g. Markov Random Fields (MRF's) [64,46] or non-linear classifiers [40]. With the advances of deep neural networks, the single-view reconstruction problem can be effectively approached with Convolutional Neural Networks (CNN's) [17,48,44,42,18,32]. Camera

poses can also be inferred by deep neural networks. [33,30] learn camera poses from single images, while [74,82,10] learn from image pairs. By learning the poses from whole sequences [12], [74], [79], a more accurate camera trajectory can be estimated. An important trend is the development of self-supervised learning objectives for learning both geometry and motion [19,21,89,80,22,29,24], which alleviates the need of expensive 3D ground truth data. In this context, dynamic objects were addressed with semantics [8,43,35] or general motion parametrizations and object masks [73,45,61]. However, self-supervised training cannot fully resolve the inherent ambiguities in dynamic scenes. Furthermore, they rely on single-view reconstruction networks, which may not generalize well from one dataset to another [48,14] or create the need for massive datasets [59]. Ranftl et al. [58] apply the transformer architecture to supervised single-view depth prediction and achieve SotA results within this setting, but the accuracy is still behind multi-view approaches on our target datasets. Another line of work focuses on improving temporal depth consistency in dynamic scenes by finetuning existing single-view networks [36,87]. While they improve the accuracy considerably, they rely on a complicated processing pipeline and use an offline camera pose estimation. Other works learn geometry from multi-view information based on cost volumes [81,47,16,68,5], but assume known camera poses and no object motion. Wimbauer et al. [76] leverage a cost volume in a semi-supervised approach. They learn a motion mask for dynamic objects and use stereo supervision in these areas.

Combinations with traditional SfM. One of the first deep SfM systems was published by [72]. Since then, a series of frameworks combine multi-view image information for inferring camera motion and scene geometry [88,11,69,70,23,75]. While most works rely on generic network architectures, few combine learning with a traditional geometric optimization [70,69,11]. We base our model on DeepV2D [70], which couples supervised training of depth based on a cost volume architecture with a geometric pose graph optimization. These choices allow the networks to exhibit a strong cross-dataset generalization, which can be seen from the results in [70] and our experiments. Recent work Manydepth focuses on training unsupervised multi-view cost volumes [75], but unsupervised training results in larger errors on moving objects due to the before-mentioned problems. Recent work DRO [23] has focused on creating a fully recurrent architecture both for structure and motion estimation. While DRO collapses the cost volume in order to achieve a lightweight architecture, we exploit the uncertainty information that is maintained inside the volume. We focus on depth prediction, but our approach can also be leveraged in downstream tasks, e.g. scene flow estimation [6] or mapping/fusion [47,88,41].

Uncertainty estimation. Uncertainty can be distinguished into *aleatoric* and *epistemic* uncertainty [31]. While epistemic uncertainty is the uncertainty over model parameters, aleatoric uncertainty is over the model outputs. Several uncertainty estimation strategies exist for regression problems: 1. Learn parameters of probability distribution as output [31,78,6]. 2. Maintain a full discrete probability volume [34,70,47,86,41]. 3. Bayesian neural network with distributions over model parameters 4. Dropout variational inference [31]. Another related work is [57], which investigates various uncertainty techniques for unsupervised single-image depth prediction. Our approach falls in the second

category and learns an aleatoric uncertainty. The idea of using cost volumes for depth reconstruction originates from stereo disparity estimation [34, 9, 86, 85]. The uncertainty was not yet investigated and exploited with more recent deep SfM-frameworks in the multi-view setting. We realize, that using the Shannon entropy naively [47, 78] within our framework does not result in optimal uncertainty estimates. We instead follow a similar strategy as [86], and introduce an uncertainty head into the network.

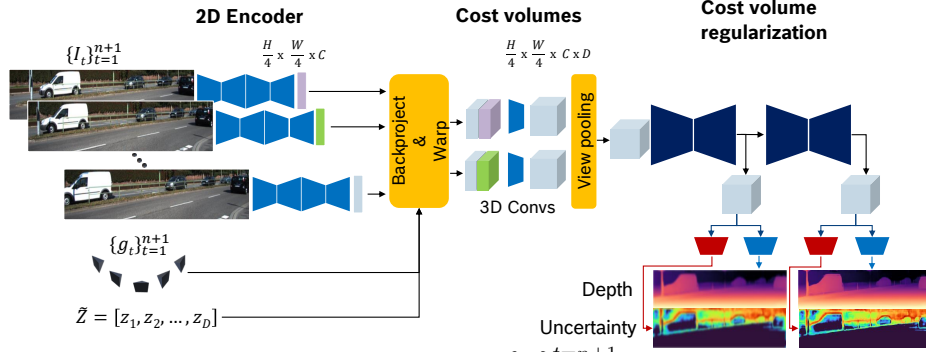


Fig. 2: Depth network of DeepV2cD. Based on $\{I_t\}_{t=1}^{t=n+1}$ and current estimated camera poses $\{g_t\}_{t=1}^{t=n+1}$ the network builds a cost volume with learned features for each frame. 3D hourglasses regularize the cost volume and output intermediate volumes. We train a separate uncertainty head to learn an uncertainty σ for the depth estimate.

3 Approach

3.1 Problem Statement

Given an ordered sequence of images $\{I_t\}_{t=1}^{2n+1}$ from a calibrated camera with intrinsics \mathbf{K} , compute the set of extrinsic camera parameters $\{g_t\}_{t=1}^{2n+1}$ and scene depth Z_{n+1} . This can be considered as windowed bundle adjustment problem. For ease of notation, we use simply Z to denote the key frame depth.

3.2 Notation

Based on a pinhole camera model, projection and backprojection are defined as:

$$\pi(\mathbf{X}) = \left[f_x \frac{X}{Z} + c_x, f_y \frac{Y}{Z} + c_y \right] \quad (1)$$

$$\pi^{-1}(\mathbf{x}, Z) = \left[Z \frac{u - c_x}{f_x}, Z \frac{v - c_y}{f_y}, Z, 1 \right]^T \quad (2)$$

with camera coordinates $\mathbf{x} = [u, v, 1]$, 3D coordinates $\mathbf{X} = [X, Y, Z, 1]^T$ and camera intrinsics $[f_x, f_y, c_x, c_y]$. The camera motion is modelled using the rigid body

motion $\mathbf{g}_{ij} \in SE_3$. Given two views i and j , the relative coordinate transformation is defined as:

$$\begin{aligned} \mathbf{x}^j &= \pi(\mathbf{g}_{ij}\mathbf{X}^i) = \pi(\mathbf{g}_j\mathbf{g}_i^{-1}\pi^{-1}(\mathbf{x}^i, Z)) \\ &= \Psi(\mathbf{g}_{ij}, \mathbf{x}^i, Z), \end{aligned} \quad (3)$$

where Ψ is the reprojection operator that transforms a pixel coordinate from camera i into camera j . Given an element of the Lie group \mathbf{g} , the logarithm map $\xi = \log \mathbf{g} \in se_3$ maps to the Lie algebra. Vice versa the exponential map is defined as $\mathbf{g} = e^\xi \in SE_3$.

3.3 Networks

This section introduces the DeepV2cD framework. After restating the base networks [70], we introduce our uncertainty head and the explored learning strategies.

Motion Network. We use the motion network from [70]. We briefly summarize the architecture and idea in the following. The motion network predicts a set of camera poses $\{\mathbf{g}_t\}_{t=1}^{2n+1}$ based on Z and $\{\mathbf{I}_t\}_{t=1}^{2n+1}$. It is a recurrent network that instead of regressing a final pose set, outputs a sequence of l sets $[\{\mathbf{g}\}_1, \{\mathbf{g}\}_2, \dots, \{\mathbf{g}\}_l]$. The camera poses $\{\mathbf{g}_t\}_{t=1}^{n+1}$ act as state variables, that determine the alignment of learned features and are updated based on an internal geometric pose graph optimization. Relative camera poses between frames are initialized based on a generic network [89]. A feature encoder produces dense features $F^i \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 32}$ for each image I^i . Based on the current state $\{\mathbf{g}_{t,i}\}_{t=1}^{2n+1}$ and input scene geometry Z features are aligned in a canonical coordinate frame and concatenated. A 2D hourglass network [54] estimates a pairwise residual optical flow \mathbf{R} and confidence map \mathbf{W} , depending on the variable state across recurrences. We observe, that the motion network learns to ignore regions in the image, that violate the static scene assumption via the confidence \mathbf{W} . It only keeps regions with good optical flow for pose optimization. Given a state $\{\mathbf{g}_t\}_{t=1}^{2n+1}$, Z , \mathbf{W} and \mathbf{R} , an update ξ can be computed via Gauss-Newton optimization [70].

Depth Network. We extend the depth architecture with an additional prediction head and restate the core principles in the following. An overview of the depth network can be seen in Figure 2. The network design is conceptually similar to PSMNet [9], but gives only two instead of three intermediate predictions due to memory constraints in our setting. This network predicts a dense depth map Z based on $\{\mathbf{I}_t\}_{t=1}^{2n+1}$ and $\{\mathbf{g}_t\}_{t=1}^{2n+1}$. The motion of dynamic objects is not given as input to the network.

For each image $I^i \in \mathbb{R}^{H \times W \times 3}$, a 2D hourglass encoder [54] learns features $F^i \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times L}$ with feature dimension $L = 32$. The 2D features for each time frame i are reprojected based on the static scene geometry and camera motion defined by (Z, \mathbf{g}_{ij}) and aligned in a canonical coordinate frame j , that is the keyframe. Given the correct depth Z^* and camera motions \mathbf{g}_{ij}^* , the features should be matched well as long as the scene is static. A cost volume \mathbf{C} is constructed by backprojecting 2D features into the key frame over a discrete 1D depth interval with $\tilde{Z} = [z_1, z_2, \dots, z_D]$. The depth range in the scene is discretized into $D - 1$ bins. When reprojecting features into this common

reference frame based on Ψ , they are bilinearly interpolated with the sampler $\Phi(\cdot)$ [28]. For each point \mathbf{x}_k^i in frame i and depth $z_d \in \tilde{Z}$:

$$\begin{aligned} C_{kd}^j &= \Phi[\Psi_{ij}(\mathbf{x}_k^i, z_d)] \\ &= \Phi[\pi(\mathbf{g}_j \mathbf{g}_i^{-1} \pi^{-1}(\mathbf{x}_k^i, z_d))] \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times D \times L}. \end{aligned} \quad (4)$$

We use a discretization of $D = 32$ bins, ranging to 80m in practice due to memory constraints. The reprojection operator $\Psi_{ij}: i \mapsto j$, is only dependent on camera motion and scene geometry. All non-keyframe features are concatenated with the backprojected keyframe features to form a final pairwise cost volume $\mathbf{C}^{ij} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times D \times 2L}$. Afterwards feature matching is learned with a series of 3D convolutions on the volume for each pair. The information across all pairs of images is globally aggregated in a pooling layer, so that a single cost volume $\mathbf{C} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times D \times L}$ remains. The cost volume is then processed by a series of 3D hourglass modules [54], which make up most of the networks parameters. This step is commonly referred to as cost volume regularization in the stereo literature [34, 9] and we believe it is crucial to achieve good performance in dynamic environments since up to this point the learned features have been matched solely on a static scene assumption. Each hourglass outputs an intermediate regularized cost volume that is run through a depth head to give a final depth estimate $Z_i \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 1}$.

The depth head produces a probability volume by a series of $1 \times 1 \times 1$ convolutions and a softmax operator over the depth dimension. The prediction is estimated using the differentiable argmax function [34], thus giving the expected depth of the probability distribution. All predictions are upsampled to final resolution $H \times W$ with naive bilinear upsampling. The design is similar to the stereo network PSMNet [9], but only gives two intermediate predictions instead of three in PSMNet due to memory constraints in the multi-view setting.

Similar to related work [78, 41, 47], we can estimate an uncertainty based on the maintained probability volume. Given the probability distribution P over \tilde{Z} , we can compute the Shannon entropy:

$$H = \sum_{d=1}^D P(z_d) \log P(z_d) \quad (5)$$

as a measure of uncertainty [78]. However, due to the limited discrete depth hypothesis space and the soft-argmax operation, the network does not necessarily learn to estimate uni-modal distributions since an infinite number of distributions can produce the same expected depth value. The entropy is not necessarily a good measure of uncertainty. Instead of taking the entropy naively, we learn a confidence $f \in [0, 1]^{M \times N \times 1}$ with a separate network head similar to [86]. This head performs a series of $3 \times 3 \times 3$ convolutions on the regularized cost volume to check for hard-to-match pixels; we use 4 convolution layers. The uncertainty σ is then defined as the linearly scaled $\sigma = s \cdot (1 - f) + \epsilon$, which avoids numerical issues; we set $s = 2$ and $\epsilon = 0.25$ empirically. For training the uncertainty head we adopt the regularization loss from [86] and define an

unimodal groundtruth distribution:

$$P^*(z_d) = \text{softmax}\left(-\frac{|z_d - Z^*|}{\sigma}\right) \quad (6)$$

over all values in the depth hypothesis space z_d and with the estimated uncertainty σ from the uncertainty head. The sharpness of the peak is dependent on the uncertainty predicted by the network.

Iterative inference by coupling motion and depth. As the two networks are functions partially dependent on the output of each other, it is alternated between them during inference. They converge to a final optimum after few iterations as shown in [70]. At the start of the inference, the depth is initialized to a constant depth map for all of our experiments, thus requiring no warm start.

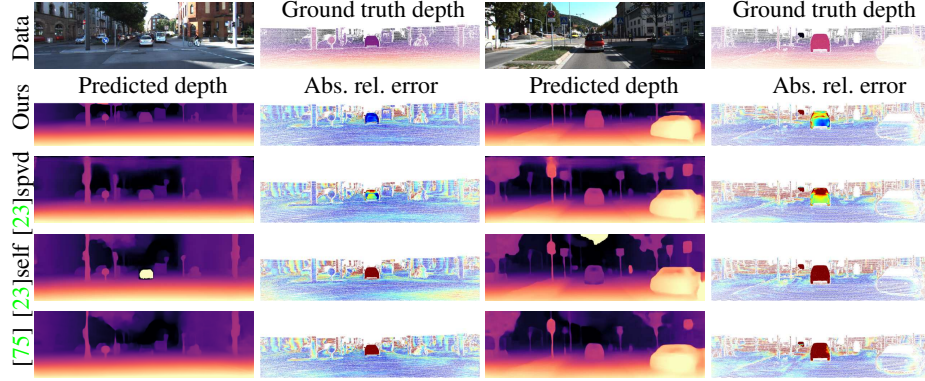


Fig. 3: Qualitative results on dense KITTI 2015 training split for 5-frame predictions. Self-supervised training struggles with objects moving colinear to the camera. While supervised frameworks do not suffer from this, they have error patterns due to sparse lidar groundtruth (e.g. cars and windows). Moving objects are usually not fully covered.

4 Training

We tightly follow the training protocol of [70]. Each network outputs a sequence of predictions, resulting in \mathbf{Z}_i and $\{\mathbf{g}\}_i$. Since the model output is a sequence of m predictions the final loss is defined as the weighted sum:

$$L = \sum_{s=1}^m \gamma^{m-s} L_s \quad (7)$$

with $\gamma = 0.5$.

4.1 Losses

We experiment with a supervised and semi-supervised setting and compare to related work:

$$\begin{aligned} L_{spvd} &= \lambda_1 L_1 + \lambda_2 L_{smooth} + \lambda_3 L_{focal} + \lambda_4 L_{flow} \\ L_{semi} &= \hat{\lambda}_1 L_{d, photo} + \hat{\lambda}_2 L_{smooth} + \hat{\lambda}_3 L_{m, photo} + \hat{\lambda}_4 L_{se3} \end{aligned} \quad (8)$$

Supervised. As is common, depth is supervised with an l_1 loss:

$$L_1(Z, Z^*) = \frac{1}{HW} \sum_k |Z(\mathbf{x}_k) - Z^*(\mathbf{x}_k)| \quad (9)$$

The camera pose graph is supervised by a reprojection error, based on the induced optical flow from the camera motion:

$$L_{flow}(\mathbf{G}, \mathbf{G}^*) = \frac{1}{HW} \sum_k \|\Psi(G, \mathbf{x}_k, Z(\mathbf{x}_k)) - \Psi(\mathbf{G}^*, \mathbf{x}_k, Z(\mathbf{x}_k))\|_1.$$

Since the ground truth is sparse on real data, an additional smoothness loss is used when no ground truth is defined. The binary mask M denotes missing ground truth:

$$L_{smooth} = \frac{1}{HW} \sum_k M(\mathbf{x}_k) \odot |\partial_x Z(\mathbf{x}_k) + \partial_y Z(\mathbf{x}_k)|. \quad (10)$$

It is common practice in monocular depth prediction to use edge-aware smoothing, mainly popular in unsupervised learning [21,23]:

$$L_{smooth} = \frac{1}{HW} \sum_k M \odot |\partial_x Z| e^{-|\partial_x \mathbf{I}|} + |\partial_y Z| e^{-|\partial_y \mathbf{I}|}. \quad (11)$$

This assumes, that edges in the depth map are usually a subset of the image edges, thus sharing a spatial location. Since the loss can be readily used for supervised learning as well, we use it for both our semi-supervised and DeepV2cD experiments.

In order to train our uncertainty, we use a depth focal loss [86]:

$$L_{focal} = \frac{1}{HW} \sum_k \sum_{d=1}^D (1 - P^*(\mathbf{x}_k))^{-\delta} \cdot (-P(\mathbf{x}_k) \cdot \log P^*(\mathbf{x}_k)),$$

with hyperparameter δ ; we set $\delta = 2$. Depth supervision on real data is costly to obtain, but camera poses are often readily available. We consider a semi-supervised setting, because it can help to resolve the scale ambiguity that arises with photometric information only. We use a weighted geodesic distance on the camera poses similar to [33,50] for supervising camera pose estimation and average this over the pose graph:

$$L_{se3}(\mathbf{G}^*, \mathbf{G}) = d(\mathbf{T}^*, \mathbf{T}) + \beta \cdot d(\mathbf{R}^*, \mathbf{R}) \quad (12)$$

with $\beta = 1.0$.

Self-supervised. Self-supervised learning uses the image data for creating a training signal. Based on the estimated depth Z and a relative pose \mathbf{g} , a warped image \mathbf{I}' can be constructed with Ψ from the input image \mathbf{I} for each key-reference image pair. The similarity is measured by a weighted sum of structural similarity index (SSIM) and $L1$ loss [21,22,24]:

$$L_{photo}(\mathbf{I}_i, \mathbf{I}'_i) = \alpha \frac{1 - SSIM(\mathbf{I}_i, \mathbf{I}'_i)}{2} + (1 - \alpha) \|\mathbf{I}_i - \mathbf{I}'_i\|_1. \quad (13)$$

We adopt the minimum fusion strategy from [22] for fusing the photometric losses across view pairs. Similar to the supervised case, we can enforce edge-aware smoothness for the depth predictions, but with M being all in-view pixels.

Weighting. We adopt hyperparameters for balancing the original loss terms from [70] and determine others empirically. For our results, we set $\lambda = [1.0, 0.02, 0.002, 1.0]$ and $\hat{\lambda} = [10.0, 0.02, 10.0, 1.0]$.

4.2 Implementation details

We extend on the published code of Teed et al. [70], but upgrade their implementation to *Tensorflow 2* [1]. All components from the network are trained from scratch with RMSProp [71] if not stated otherwise. We use standard color and flip augmentation similar to [70,23] and small random camera pose perturbation as [70]. Our input and output resolution is 192×1088 , which is reasonable for the limited lidar coverage. We use $n = 5$ frames as input as this turns out to be enough temporal information to achieve optimal performance, see [70].

Training can be divided into two stages: 1. the motion module is trained alone with interpolated groundtruth depth as input. 2. the motion network gets a cached depth prediction with increasing likelihood from the previous epoch as input instead of the depth groundtruth and the depth network takes the predicted camera poses. Both networks are trained jointly on the combined loss and therefore depend on each other already during training. We train for approx. [5, 15] epochs with batch sizes [12, 3] for the respective stages like the baseline from [70] if not stated otherwise. During inference we run 5 iterations of the networks compared to 8 in the reference implementation, because we found that this already matches the reported performance.

5 Experiments

We test DeepV2cD and our own *Tensorflow 2* DeepV2D implementation on multiple automotive datasets that include dynamic objects. Our primary focus lies on the depth prediction performance and we include both multi-view and single-view comparisons. As common in deep monocular multi-view SfM we report results for the ground truth scale aligned depth [72,69,70]. We compare with current SotA multi-view frameworks and report results for their publicly available models.

5.1 Datasets

KITTI. The KITTI dataset [20] contains image and lidar sequences from a moving car and is widely established for evaluating depth estimation. We follow the Eigen train/test split and evaluation protocol [17] with the official ground truth maps. The original dataset has only sparse depth ground truth and misses moving object labels. To address these shortcomings, the KITTI 2015 dataset was published [53] with dense ground truth and labels for 200 frames. We evaluate on the KITTI 2015 training split to show the performance gap between the static scene and dynamic objects. During evaluation we also use the Eigen crop as on KITTI 2012.

Virtual KITTI 2. Virtual KITTI [7] is a synthetic dataset, consisting of several driving sequences from the real KITTI dataset imitated with the unity engine. It contains diverse weather and lighting conditions and moving objects. Even though it does not cover as many driving scenes as the original real data, we use it to investigate the zero-shot generalization performance of our backbone network DeepV2D and improved performance for our supervised DeepV2cD model. We generated our own 95/5 train/test split in similar fashion to the Eigen split. We noticed, that some frames have misaligned ground planes and objects that appear/vanish in the middle of the scene. By removing such frames and ones with no camera motion, we achieve a split of approx. 14.3k and 754 train/test frames. We train for [9, 20] epochs and we train sequentially when combining multiple datasets. We believe, that with a better domain adaption strategy [25] results can be improved further than sequential training on the multiple datasets.

Cityscapes. The Cityscapes dataset [13] is an autonomous driving dataset collected from various cities in Germany. The dataset contains many moving object classes, such as cars and pedestrians. It is mainly used for object detection, but also contains 1500 frames with Semi-Global Matching (SGM) [26] depth ground truth. We evaluate on this dataset to test zero-shot generalization performance and follow the evaluation protocol of [8, 75].

5.2 Accuracy of multi-view networks

Table 1 shows the current SotA for monocular depth prediction on the KITTI Eigen split. It can be seen that DeepV2D and DeepV2cD achieve the best reconstruction accuracy in the 5-view setting on the KITTI Eigen split. All multi-view frameworks achieve better results than single-view ones. While DRO [23] and Manydepth [75] are significantly faster than DeepV2D and DeepV2cD, they are less accurate in a 5-view setting. Depending on the training, not all multi-view models achieve higher accuracy with more views added compared to the 2-view setting. Our cost volume regularization and uncertainty strategy gives better results than the DeepV2D baseline [70] while needing fewer inference iterations.

All supervised multi-view frameworks except DeepV2cD can be trained with self-supervision as well. We did not manage to train DeepV2D in this setting. However, we can show, that it achieves good performance by simply training on the small Virtual KITTI (VK) dataset. Current unsupervised multi-view approaches perform best

Table 1: Results on KITTI Eigen split with improved ground truth. Multi-view frameworks have higher accuracy than single-view. We achieve the best results when regularizing the cost volume of our backbone network and training with more data.

Method	Views Supervised		$W \times H$	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$
MonoDepth [22]	1	✗	1024×320	0.091	0.531	3.742	0.135	0.916	0.984
Kuznetsov et al. [39]	1	✓	621×187	0.089	0.478	3.610	0.138	0.906	0.98
Packnet-SfM [24]	1	(✓)	640×192	0.078	0.420	3.485	0.121	0.931	0.986
DORN [18]	1	✓	513×385	0.069	0.300	2.857	0.112	0.945	0.988
BANet [69]	5	✓	-	0.083	-	3.640	0.134	-	-
Manydpepth [75]	2	✗	1024×320	0.055	0.313	3.035	0.094	0.958	0.990
Manydpepth [75]	5	✗	1024×320	0.055	0.312	3.034	0.094	0.958	0.991
DRO [23]	2	✗	960×320	0.057	0.342	3.201	0.123	0.952	0.989
DRO [23]	5	✗	960×320	0.064	0.381	3.262	0.120	0.951	0.988
DRO [23]	2	✓	960×320	0.046	0.210	2.674	0.083	0.969	<u>0.993</u>
DRO [23]	5	✓	960×320	0.047	0.212	2.711	0.084	0.968	0.994
DeepV2D (ours) (VK)	5	✗	1088×192	0.060	0.423	3.302	0.110	0.950	0.984
DeepV2D (ours) (VK + K)	5	(✓)	1088×192	0.058	0.669	3.246	0.100	0.960	0.985
DeepV2D [70]	2	✓	1088×192	0.064	0.350	2.946	0.120	0.946	0.982
DeepV2D [70]	5	✓	1088×192	<u>0.037</u>	0.174	2.005	0.074	0.977	<u>0.993</u>
DeepV2cD (ours)	5	✓	1088×192	<u>0.037</u>	<u>0.167</u>	<u>1.984</u>	<u>0.073</u>	<u>0.978</u>	0.994
DeepV2cD* (ours) (K + VK)	5	✓	1088×192	0.035	0.158	1.877	0.071	0.980	0.994

when trained on the target dataset. The gap can be further closed after finetuning semi-supervised (✓) without expensive depth ground truth for just 5 epochs. Manydepth utilizes a test-time refinement, which we did not experiment with for DRO and DeepV2D.

Table 2: Results on KITTI 2015 train split. Metrics are slightly worse for all frameworks compared to the KITTI 2012 Eigen split.

Method	Supervised	ARE Dyn. ↓	ARE Static ↓	ARE All ↓	$\delta < 1.25 \uparrow$
DRO [23]	✗	449.1	0.081	0.276	0.795
Manydpepth	✗	0.177	0.0641	0.090	0.914
DRO [23]	✓	0.069	0.050	0.056	0.953
DeepV2D (ours)	✓	0.152	<u>0.042</u>	0.071	<u>0.959</u>
DeepV2cD* (ours)	✓	<u>0.127</u>	0.039	<u>0.062</u>	0.965

Table 2 shows results on the KITTI 2015 training split to emphasize the difference between static and dynamic scene. In Figure 3, we show some qualitative examples. DeepV2cD achieves the best reconstruction of the static scene. All frameworks perform systematically worse on moving objects. Unsupervised training cannot resolve the inherent ambiguity resulting from the independent object motion and thus results in gross errors when not taking care of regularization. Manydepth resolves this problem partially by regularizing their network with a single-view teacher network compared to the self-supervised DRO during training. We observe, that even though no supervised framework addresses the object motion directly in their reprojection operator, the reconstruction works in most cases. We believe, that this is due to cost volume regularization in DeepV2D and DeepV2cD and the temporal information of the recurrent network in

Table 3: Zero shot cross-dataset generalization of multi-view frameworks. We evaluate models trained on KITTI (K) on Cityscapes (CS).

Method	Supervised	Dataset	Abs Rel \downarrow	Sq Rel \downarrow	RMSE \downarrow	RMSE log \downarrow	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Manydepth [75]	\times	CS	0.114	1.193	6.223	0.170	0.875	0.967	0.989
DRO [23]	\times	K \rightarrow CS	0.328	7.348	11.656	0.597	0.548	0.747	0.822
DRO [23]	\checkmark	K \rightarrow CS	0.157	2.228	10.306	0.299	0.777	0.900	0.948
Manydepth [75]	\times	K \rightarrow CS	0.162	1.697	8.230	0.229	0.764	0.935	0.979
DeepV2D	\checkmark	K \rightarrow CS	0.109	1.479	6.7633	0.1842	0.876	0.952	0.978
DeepV2cD*	\checkmark	K + VK \rightarrow CS	0.104	1.325	6.7328	0.1792	0.883	0.955	0.979
DeepV2cD* filtered 80%	\checkmark	K + VK \rightarrow CS	0.070	0.559	4.188	0.124	0.930	0.975	0.989

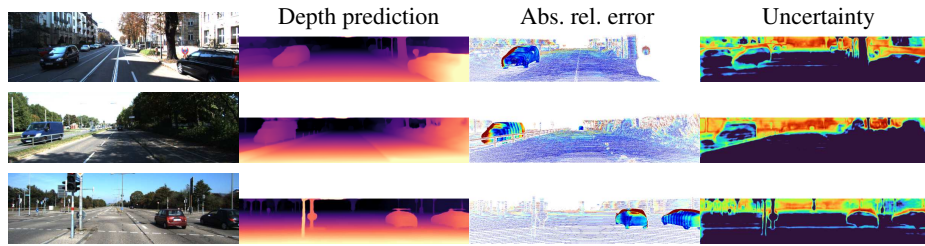
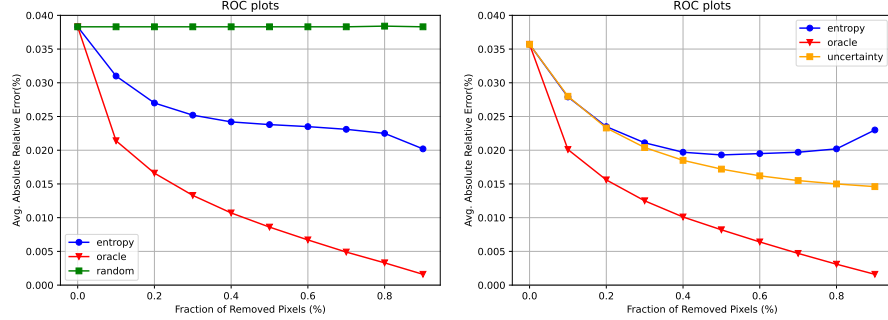


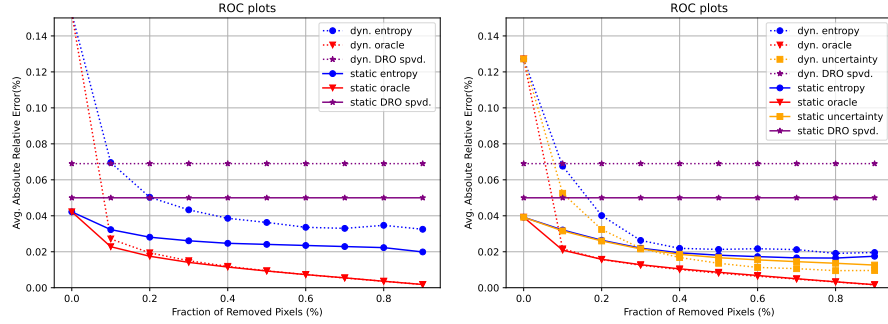
Fig. 4: Qualitative results on three samples of KITTI 2015 training split. Warm colors indicate high error and uncertainty. DeepV2cD makes errors in areas of occlusion, missing ground truth, reflective surfaces or thin structures. The uncertainty head learns to predict these areas reliably with no overhead. Most problematic for reconstruction of dynamic objects is missing lidar supervision, noticeable by the cut off roofs and holes at windows. DeepV2cD reconstructs these regions as part of the background. The uncertainty is strongest in these areas and in the back of the scene at far depths.

DRO. Since we could not observe these errors on virtual data, we believe that these errors are due to missing supervision with the sparse lidar. While the systematic model errors of DeepV2cD and DRO on the dynamic parts look qualitatively similar, DRO has a significantly lower gap to the static reconstruction. We observe, that DeepV2cD sometimes does not reconstruct the car shapes as well and produces more gross outliers. A possible explanation for this is the longer training time of DRO compared to DeepV2cD (50 epochs compared to 15) and the pre-trained ImageNet [62] weights. Pretraining on Virtual KITTI does not resolve this issue completely.

Table 3 shows the zero-shot cross-dataset generalization performance of current SotA multi-view frameworks. DeepV2D and DeepV2cD can generalize well to other datasets. We generalize best from KITTI to Cityscapes and outperform unsupervised single-view networks trained on the target. This could be explained with the explicit geometric pose graph optimization for motion estimation and the cost volume. The learned uncertainty transfers well to Cityscapes. When keeping 80% of the pixels, we effectively reach 50% the error of other frameworks in this setting.



(a) DeepV2D on KITTI 2012 Eigen test split (b) DeepV2cD* on KITTI 2012 Eigen test split



(c) DeepV2D on KITTI 2015 train split (d) DeepV2cD* on KITTI 2015 train split

Fig. 5: Sparsification plots on KITTI 2012 and 2015: We sort the pixels in ascending order of uncertainty. The x-axis represents the percentage of pixels, that are included in the evaluation and the y-axis represents the absolute relative error (ARE). An oracle acts as upper bound for identifying errors. We also include a random selection baseline for the Eigen split. We can distinguish between errors on moving objects and static background on KITTI 2015. We include supervised DRO [23] for comparison. While the Shannon entropy is already useful for filtering, the learned uncertainty correlates better with the actual error. By rejecting few percent of the pixels, we can reliably filter out gross outliers and achieve SotA performance. The filtered predictions of DeepV2cD converge to similar reconstruction performance for both static and dynamic parts of the scene.

5.3 Uncertainty

The previous experiments have shown, that with an uncertainty head and an improved regularization strategy DeepV2cD achieves SotA depth prediction accuracy on 5-frame videos on KITTI. In Figure 4, we show qualitative examples of our learned uncertainty on the KITTI dataset. It correlates with hard to match pixels, occlusions, reflective surfaces and areas of missing ground truth. In general it is higher at far away pixels in the scene. Since the uncertainty head can be run in parallel to the depth prediction head, this does not come with an additional computational overhead.

In the next experiments, we investigate the quality of the learned uncertainty for identifying errors made by the model. This implicates, that sorting the pixels by uncertainty results in the same ordering as when we sort by errors. We use sparsification plots [27,86,78] and compare the learned uncertainty with three baselines: 1. Shannon entropy 2. Random filtering 3. Oracle. Furthermore, we compare DeepV2cD to the inherent uncertainty of DeepV2D. Figure 5 shows, that DeepV2cD can learn a reliable aleatoric uncertainty while achieving SotA accuracy. We beat all baselines and can achieve cleaner reconstructions after filtering with no significant overhead. When throwing away 20% of the pixels, the avg. abs. rel. error is below 2.5%. Related work DRO [23] achieves SotA accuracy by collapsing the cost volume, but cannot detect the model errors inherent to the sparse supervision. We argue that this property is very useful for downstream tasks, such as map building or scene flow estimation [6].

6 Conclusion

In this paper, we investigated the performance of deep SfM frameworks on several autonomous driving datasets. We show improved results for a cost volume based architecture due to better loss supervision and an additional uncertainty head. Our results indicate, that supervised models make errors on real datasets mainly due to a lack of supervision and that they are able to learn more accurate depths when provided with high quality dense data on virtual datasets. Missing supervision can be compensated by considering the aleatoric uncertainty. Instead of taking the Shannon entropy inside the cost volume, a learned uncertainty showed better performance at identifying outliers.

7 ACKNOWLEDGEMENT

Research presented here has been supported by the Robert Bosch GmbH. We thank our colleagues Jan Fabian Schmidt, Annika Hagemann and Holger Janssen for fruitful discussions and proof reading.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016) 10
2. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Nonrigid structure from motion in trajectory space (2008) 3
3. Avidan, S., Shashua, A.: Trajectory triangulation of lines: Reconstruction of a 3d point moving along a line from a monocular image sequence. In: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149). vol. 2, pp. 62–66. IEEE (1999) 1, 3
4. Avidan, S., Shashua, A.: Trajectory triangulation: 3d reconstruction of moving points from a monocular image sequence. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(4), 348–357 (2000) 3

5. Božič, A., Palafox, P., Thies, J., Dai, A., Nießner, M.: Transformerfusion: Monocular rgb scene reconstruction using transformers. arXiv preprint arXiv:2107.02191 (2021) [4](#)
6. Brickwedde, F., Abraham, S., Mester, R.: Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2780–2790 (2019) [4](#), [15](#)
7. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020) [3](#), [11](#)
8. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Unsupervised monocular depth and ego-motion learning with structure and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) [4](#), [11](#)
9. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5418 (2018) [5](#), [6](#), [7](#)
10. Chen, Y., Schmid, C., Sminchisescu, C.: Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7063–7072 (2019) [4](#)
11. Clark, R., Bloesch, M., Czarnowski, J., Leutenegger, S., Davison, A.J.: Learning to solve nonlinear least squares for monocular stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 284–299 (2018) [4](#)
12. Clark, R., Wang, S., Wen, H., Markham, A., Trigoni, N.: Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017) [4](#)
13. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) [11](#)
14. Dijk, T.v., Croon, G.d.: How do neural networks see depth in single images? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2183–2191 (2019) [1](#), [4](#)
15. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017) [3](#)
16. Duzceker, A., Galliani, S., Vogel, C., Speciale, P., Dusmanu, M., Pollefeys, M.: Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15324–15333 (2021) [4](#)
17. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. arXiv preprint arXiv:1406.2283 (2014) [1](#), [3](#), [11](#)
18. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018) [1](#), [3](#), [12](#)
19. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European conference on computer vision. pp. 740–756. Springer (2016) [2](#), [4](#)
20. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012) [11](#)
21. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 270–279 (2017) [2](#), [4](#), [9](#), [10](#)
22. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838 (2019) [2](#), [4](#), [10](#), [12](#)

23. Gu, X., Yuan, W., Dai, Z., Tang, C., Zhu, S., Tan, P.: Dro: Deep recurrent optimizer for structure-from-motion. arXiv preprint arXiv:2103.13201 (2021) [1](#), [2](#), [4](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)
24. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2485–2494 (2020) [2](#), [4](#), [10](#), [12](#)
25. Guizilini, V., Li, J., Ambrus, R., Gaidon, A.: Geometric unsupervised domain adaptation for semantic segmentation. arXiv preprint arXiv:2103.16694 (2021) [11](#)
26. Hirschmüller, H.: Semi-global matching-motivation, developments and applications. Photogrammetric Week 11 pp. 173–184 (2011) [11](#)
27. Hu, X., Mordohai, P.: A quantitative evaluation of confidence measures for stereo vision. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2121–2133 (2012) [15](#)
28. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. arXiv preprint arXiv:1506.02025 (2015) [7](#)
29. Johnston, A., Carneiro, G.: Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4756–4765 (2020) [4](#)
30. Kendall, A., Cipolla, R.: Modelling uncertainty in deep learning for camera relocation. In: 2016 IEEE international conference on Robotics and Automation (ICRA). pp. 4762–4769. IEEE (2016) [4](#)
31. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977 (2017) [4](#)
32. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018) [3](#)
33. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocation. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946 (2015) [4](#), [9](#)
34. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 66–75 (2017) [4](#), [5](#), [7](#)
35. Klingner, M., Termöhlen, J.A., Mikolajczyk, J., Fingscheidt, T.: Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: European Conference on Computer Vision. pp. 582–600. Springer (2020) [4](#)
36. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1611–1621 (2021) [4](#)
37. Kumar, S., Dai, Y., Li, H.: Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4649–4657 (2017) [3](#)
38. Kumar, S., Dai, Y., Li, H.: Superpixel soup: Monocular dense 3d reconstruction of a complex dynamic scene. IEEE transactions on pattern analysis and machine intelligence (2019) [3](#)
39. Kuznetsov, Y., Stuckler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6647–6655 (2017) [12](#)
40. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 89–96 (2014) [3](#)
41. Laidlow, T., Czarnowski, J., Nicastro, A., Clark, R., Leutenegger, S.: Towards the probabilistic fusion of learned priors into standard pipelines for 3d reconstruction. In: 2020 IEEE

- International Conference on Robotics and Automation (ICRA). pp. 7373–7379. IEEE (2020) [3](#), [4](#), [7](#)
42. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016) [1](#), [3](#)
 43. Lee, S., Im, S., Lin, S., Kweon, I.S.: Instance-wise depth and motion learning from monocular videos. arXiv preprint arXiv:1912.09351 (2019) [4](#)
 44. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: 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. pp. 1119–1127 (2015) [1](#), [3](#)
 45. Li, H., Gordon, A., Zhao, H., Casser, V., Angelova, A.: Unsupervised monocular depth learning in dynamic scenes. arXiv preprint arXiv:2010.16404 (2020) [4](#)
 46. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1253–1260. IEEE (2010) [3](#)
 47. Liu, C., Gu, J., Kim, K., Narasimhan, S.G., Kautz, J.: Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10986–10995 (2019) [3](#), [4](#), [5](#), [7](#)
 48. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence **38**(10), 2024–2039 (2015) [1](#), [3](#), [4](#)
 49. Longuet-Higgins, H.C.: A computer algorithm for reconstructing a scene from two projections. Nature **293**(5828), 133–135 (1981) [3](#)
 50. Mahendran, S., Ali, H., Vidal, R.: 3d pose regression using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 2174–2182 (2017) [9](#)
 51. Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A., Brox, T.: What makes good synthetic training data for learning disparity and optical flow estimation? International Journal of Computer Vision **126**(9), 942–960 (2018) [3](#)
 52. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4040–4048 (2016) [3](#)
 53. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3061–3070 (2015) [11](#)
 54. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016) [6](#), [7](#)
 55. Park, H.S., Shiratori, T., Matthews, I., Sheikh, Y.: 3d reconstruction of a moving point from a series of 2d projections. In: European conference on computer vision. pp. 158–171. Springer (2010) [3](#)
 56. Park, H.S., Shiratori, T., Matthews, I., Sheikh, Y.: 3d trajectory reconstruction under perspective projection. International Journal of Computer Vision **115**(2), 115–135 (2015) [3](#)
 57. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3227–3237 (2020) [4](#)
 58. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021) [1](#), [4](#)

59. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341* (2019) 1, 4
60. Ranftl, R., Vineet, V., Chen, Q., Koltun, V.: Dense monocular depth estimation in complex dynamic scenes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4058–4066 (2016) 3
61. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12240–12249 (2019) 4
62. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**(3), 211–252 (2015) 13
63. Saputra, M.R.U., Markham, A., Trigoni, N.: Visual slam and structure from motion in dynamic environments: A survey. *ACM Computing Surveys (CSUR)* **51**(2), 1–36 (2018) 3
64. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence* **31**(5), 824–840 (2008) 3
65. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4104–4113 (2016) 3
66. Shashua, A., Avidan, S., Werman, M.: Trajectory triangulation over conic section. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. vol. 1, pp. 330–336. *IEEE* (1999) 3
67. Snavely, K.N.: Scene reconstruction and visualization from internet photo collections. University of Washington Washington, DC (2008) 3
68. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15598–15607 (2021) 4
69. Tang, C., Tan, P.: Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807* (2018) 1, 2, 4, 10, 12
70. Teed, Z., Deng, J.: Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605* (2018) 1, 2, 4, 6, 8, 10, 11, 12
71. Tieleman, T., Hinton, G., et al.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* **4**(2), 26–31 (2012) 10
72. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5038–5047 (2017) 1, 4, 10
73. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804* (2017) 4
74. Wang, S., Clark, R., Wen, H., Trigoni, N.: Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 2043–2050. *IEEE* (2017) 4
75. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1164–1174 (2021) 1, 2, 4, 8, 11, 12, 13
76. Wimbauer, F., Yang, N., von Stumberg, L., Zeller, N., Cremers, D.: Monorec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. *arXiv preprint arXiv:2011.11814* (2020) 4

77. Wu, C., et al.: Visualsfm: A visual structure from motion system (2011) [1](#), [3](#)
78. Yang, G., Hu, P., Ramanan, D.: Inferring distributions over depth from a single image. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6090–6096. IEEE (2019) [3](#), [4](#), [5](#), [7](#), [15](#)
79. Yang, N., Stumberg, L.v., Wang, R., Cremers, D.: D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1281–1292 (2020) [4](#)
80. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) [4](#)
81. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 767–783 (2018) [4](#)
82. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1983–1992 (2018) [4](#)
83. Yoon, J.S., Kim, K., Gallo, O., Park, H.S., Kautz, J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5336–5345 (2020) [3](#)
84. Yuan, C., Medioni, G., Kang, J., Cohen, I.: Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. IEEE transactions on pattern analysis and machine intelligence **29**(9), 1627–1641 (2007) [2](#), [3](#)
85. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 185–194 (2019) [5](#)
86. Zhang, Y., Chen, Y., Bai, X., Yu, S., Yu, K., Li, Z., Yang, K.: Adaptive unimodal cost volume filtering for deep stereo matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12926–12934 (2020) [3](#), [4](#), [5](#), [7](#), [9](#), [15](#)
87. Zhang, Z., Cole, F., Tucker, R., Freeman, W.T., Dekel, T.: Consistent depth of moving objects in video. ACM Transactions on Graphics (TOG) **40**(4), 1–12 (2021) [4](#)
88. Zhou, H., Ummenhofer, B., Brox, T.: Deeptam: Deep tracking and mapping. In: Proceedings of the European conference on computer vision (ECCV). pp. 822–838 (2018) [4](#)
89. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017) [4](#), [6](#)