

DRO: Deep Recurrent Optimizer for Video to Depth

Xiaodong Gu^{1*}, Weihao Yuan^{1*}, Zuozechuo Dai¹, Siyu Zhu¹, Chengzhou Tang², Zilong Dong¹ and Ping Tan^{1,2}

Abstract—There are increasing interests of studying the video-to-depth (V2D) problem with machine learning techniques. While earlier methods directly learn a mapping from images to depth maps and camera poses, more recent works enforce multi-view geometry constraints through optimization embedded in the learning framework. This paper presents a novel optimization method based on recurrent neural networks to further exploit the potential of neural networks in V2D. Specifically, our neural optimizer alternately updates the depth and camera poses through iterations to minimize a feature-metric cost, and two gated recurrent units iteratively improve the results by tracing historical information. Extensive experimental results demonstrate that our method outperforms previous methods and is more efficient in computation and memory consumption than cost-volume-based methods. In particular, our self-supervised method outperforms previous supervised methods on the KITTI and ScanNet datasets. Our source code will be made public.

Index Terms—Deep Learning for Visual Perception; Computer Vision for Transportation; Visual Learning

I. INTRODUCTION

VIDEO to depth (V2D) is a fundamental task in computer vision and essential for numerous applications such as robotics, autonomous driving, augmented reality, and 3D reconstruction. Given a sequence of images, V2D methods optimize depth maps and camera poses to recover the 3D structure of a scene. Traditional methods solve the Bundle-Adjustment (BA) problem, where the re-projection error between reprojected 3D scene points and 2D image feature points are minimized iteratively.

Recently, deep-learning-based methods have dominated most benchmarks and demonstrated advantages over traditional methods [1–6]. Earlier learning-based methods [5, 7–9] directly regress the depth maps and camera poses from the input images. To combine the strength of neural networks and traditional geometric methods, more recent works formulate the geometric-based optimization as differentiable layers and embed them in a learning framework [3, 4, 10].

We follow the approach of combining neural networks and optimization methods with some novel insights. Firstly, previous methods [3, 4, 11] adopt gradient-based optimization such as Levenberg-Marquardt or Gauss-Newton methods. However, the gradients could be noisy and misleading especially for the high-dimensional optimization problem in dense depth map

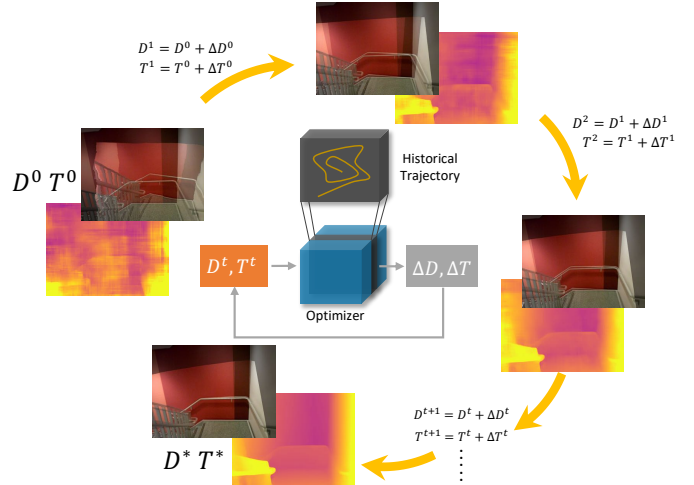


Fig. 1: A gated recurrent network (indicated as the optimizer in the middle) iteratively updates a depth map and the relative motion between two images by minimizing a feature-metric cost. At each iteration, we show a color-coded depth map and a superimposed image generated from the two input images according to the depth map and camera motion. Over the iterations, the superimposed image becomes gradually sharper while the depth map improves.

computation. Traditional methods like LSD-SLAM [12] and DSO [13] focus on edge pixels to make the optimization problem feasible. More recent learning-based methods design learned regularization such as depth bases [3] or manifold embeddings [14, 15] to address this problem. However, this learned regularization might have difficulty to be generalized to unseen scenes. Furthermore, a multi-resolution strategy is needed to gradually compute the solution from coarse to fine. In comparison, we employ a gated recurrent neural network for optimization as inspired by [16]. An illustration is shown in Fig. 1. Remarkably, our method is gradient-free and works on the high resolution image directly without any regularization which might limit the algorithm generalization.

Secondly, some methods [4, 6, 10, 17] build cost volumes to solve dense depth maps. Similar cost volumes are also employed in [16] to compute optical flow. A cost volume encodes the errors of multiple different depth values at each pixel. It evaluates the result quality within a large spatial neighborhood in the solution space in a discrete fashion. While cost volumes have been demonstrated effective in computing depth maps [6, 18, 19], they are inefficient in time and space because they exhaustively evaluate results in a large spatial neighborhood. Differently, in this work, we use a gated recurrent network [20] to minimize the feature-metric error to compute dense depth without computing such a cost volume. A gated recurrent network updates depth maps only by evaluating the current solution (i.e. a single point in the solution space)

Manuscript received: October, 18, 2022; Revised January, 20, 2023; Accepted March, 6, 2023.

This paper was recommended for publication by Editor Cadena Lerma, Cesar upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Alibaba Group

*Equal contribution. Authors are with ¹ Alibaba Group and ² Simon Fraser University. dadong.gxd@alibaba-inc.com, qianmu.ywh@alibaba-inc.com

Digital Object Identifier (DOI): see top of this page.

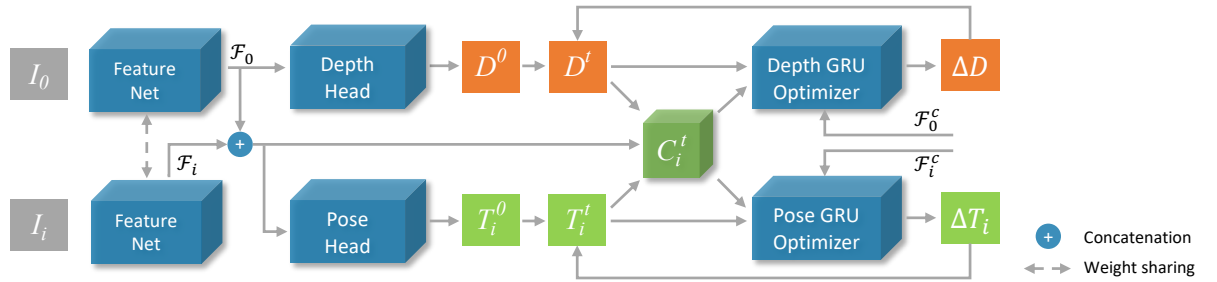


Fig. 2: The overview of our framework. The reference image I_0 and context images I_i are first fed into a feature network with shared parameters to extract feature maps. The extracted features are then mapped to an initial depth map D^0 and an initial relative camera pose T_i^0 by the depth head and the pose head, respectively. Afterwards, the recurrent optimizer update the depth and pose iteratively. During this optimization, we first compute a cost map C_i^t from the current solution D^t and T_i^t . The GRU optimizer then computes the solution updates ΔD and ΔT_i accordingly. Eventually, the depth map and pose gradually converge to the optimum, D^* and T_i^* .

and some previous solutions over time. In spirit, our learned recurrent network exploits temporal information during iterations, while gradient based methods [3] or cost volume based methods [6] rely only on spatial information. In this way, our method has the potential of better running time and memory efficiency.

In experiments, we show that our recurrent optimizer reduces the feature-metric cost over iterations and produces gradually improved depth maps and camera poses. Our method demonstrates better accuracy than previous methods in both indoor and outdoor data, under both supervised and self-supervised settings. In particular, our self-supervised method outperforms previous methods by 14.5% on KITTI and by 23.2% on ScanNet.

Our contributions can be summarized as follows:

- 1) We propose a novel recurrent optimizer for joint depth and pose optimization where gradients or cost volumes are not involved for better memory and computation efficiency.
- 2) The depths and poses are alternately updated to uncouple the mutual influence by the GRU module for effective optimization.
- 3) Our optimizer produces better results than previous methods in both supervised and self-supervised settings.

II. RELATED WORK

Supervised Deep V2D. Deep neural networks can learn to solve the V2D problem directly from data [5, 6, 17, 21]. With the ground-truth information, DeMoN [5] trains two network branches to regress depths and motions separately with an auxiliary flow prediction task to exploit feature correspondences. Some methods adopt a discrete sampling strategy to achieve high-quality depth maps [4, 17]. They generate depth hypotheses and utilize multiple images to construct a cost volume. Furthermore, the pose volume is also introduced in [6]. They take the feature maps to build two cost volumes and employ 3D convolutions to regularize. There are also methods to directly regress scene depth from a single input image [1, 7, 22], which is an ill-posed problem.

Self-supervised Deep V2D. Supervised methods, nevertheless, require collecting a large number of training data with ground-truth depth and camera poses. Recently, many self-supervised works [2, 23–33] have been proposed to train a depth and pose estimation model from only monocular

RGB images. They employ the predicted depths and poses to warp neighbor frames to the reference frame, such that a photometric constraint is created to serve as a self-supervision signal. In this case, the dynamic objects is a problem and would generate errors in the photometric loss. To address this, semantic mask [34] and optical flow [35–37] are proposed to exclude the influence of moving objects. Another challenge is the visibility between different frames. To deal with this, a minimum re-projection loss is designed in [2, 24] to handle the occlusion. Despite these efforts, there is still a gap between self-supervised and supervised methods.

Learning to Optimize. Traditional computer vision methods usually formulate the tasks as optimization problems according to the first principles such as photo-consistency and multi-view geometry. Many of recent works try to combine the strength of neural networks and traditional optimization-based methods. There are mainly two approaches in learning to optimize. One approach [3, 4, 38, 39] employs a network to predict the inputs or parameters of an optimizer, which is implemented as some layers in a large neural network for end-to-end training. On the contrary, the other approach directly learn to update optimization variables from the data [11, 16, 40–42]. There, however, are some problems in previous methods. Methods of the first approach need to explicitly formulate the solver and are limited to problems where the objective functions are easily defined [3, 4, 38, 39]. Furthermore, the methods in [3, 11] need to explicitly evaluate gradients of the objective function, which is hard in many problems. Besides, the methods in [4, 16] adopt cost volumes, which make the model heavy to apply.

In comparison, our method does not require gradients computation or cost volume aggregation. It only evaluates the result quality at a single point in the solution space at each step. By accumulating temporal evidence from previous iterations, our GRU module learns to minimize the objective function. Besides, two updaters in our framework, one for depth and the other one for pose, are alternately updated, which is inspired by the traditional bundle adjustment.

III. DEEP RECURRENT OPTIMIZER

A. Overview

Given a reference image I_0 and N neighboring images $\{I_i\}_{i=1}^N$, our method outputs the depth D of the reference

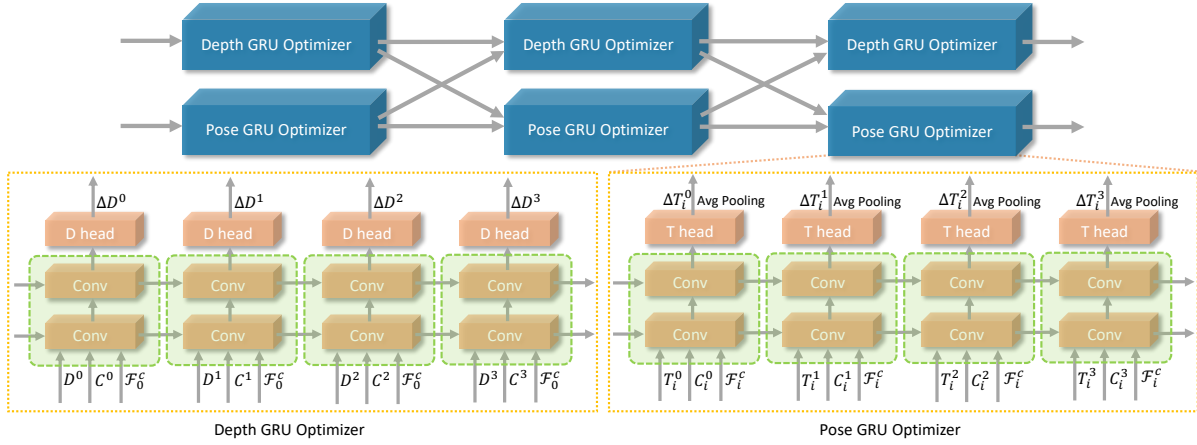


Fig. 3: Working flow of the optimizer. Updating of the depth and the pose are separated in each stage, where 4 updates of the depth are followed by 4 updates of the pose. We adopt 3 stages in our framework. For each update for the depth, the predicted depth \mathbf{D}^t , cost \mathbf{C}^t , and contextual feature map \mathcal{F}_0^c are fed in, then the update $\Delta \mathbf{D}^t$ is predicted based on the inputs and historical information. Afterwards, the depth is updated by $\mathbf{D}^{t+1} = \mathbf{D}^t + \Delta \mathbf{D}^t$.

image and the relative camera poses $\{\mathbf{T}_i\}_{i=1}^N$ for images $\{\mathbf{I}_i\}_{i=1}^N$ as shown in Fig. 2. Images first go through a shared feature extraction module to produce feature maps \mathcal{F}_i . A depth head and a pose head then take in these features and output an initial depth map \mathbf{D}^0 and initial relative poses \mathbf{T}_i^0 . Finally, these initial results are iteratively refined by the depth and the pose GRU-optimizers alternately, and converge to the final depth \mathbf{D}^* and poses \mathbf{T}_i^* .

B. Feature Extraction and Cost Construction

a) *Feature extraction.*: We use ResNet18 [43] as the backbone to extract the features $\{\mathcal{F}_i\}_{i=0}^N$. The resolution of the feature maps is 1/8 of the original input images. Likewise, the initial hidden state defined by h^0 and the contextual features $\{\mathcal{F}_i^c\}_{i=0}^N$ for the GRU optimizer come from the same-structure feature network.

b) *Cost Construction.*: Similar to BA-Net [3], we construct a photometric cost in feature space as the energy function to minimize. The cost measures the distance between aligned feature maps. Given the depth map \mathbf{D} of the reference image \mathbf{I}_0 and the relative camera pose \mathbf{T}_i of another image \mathbf{I}_i with respect to \mathbf{I}_0 , the cost is defined at each pixel x in the reference image \mathbf{I}_0 :

$$\mathbf{C}_i(x) = \|\mathcal{F}_i(\pi(\mathbf{T}_i \circ \pi^{-1}(x, \mathbf{D}(x)))) - \mathcal{F}_0(x)\|_2, \quad (1)$$

where $\|\cdot\|_2$ is the L2 norm, and $\pi(\cdot)$ is the projection of a 3D point to the image plane. Thus, its inverse function $\pi^{-1}(x, \mathbf{D}(x))$ maps a pixel x and its depth $\mathbf{D}(x)$ back to a 3D point. The transformation \mathbf{T}_i convert 3D points from the camera space of \mathbf{I}_0 to that of \mathbf{I}_i . When there are multiple neighboring images, we average multiple cost values $\{\mathbf{C}_i(x)\}_{i=1}^N$ as $\mathbf{C}(x)$ at each pixel:

$$\mathbf{C}(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{C}_i(x). \quad (2)$$

Note that the feature-metric error in BA-Net [3] will further sum the cost over all pixels as $\sum_x \mathbf{C}(x)$. However, in this work, we maintain a cost map $\mathbf{C}(x)$ with the same resolution as the feature map \mathcal{F}_i to facilitate minimization of the

feature-metric error. In the following of this paper, we refer $\mathbf{C}(x)$ as cost map instead of feature-metric error. In alternate optimization, the camera pose \mathbf{T}_i is optimized by minimizing $\mathbf{C}_i(x)$ and the depth map \mathbf{D} is optimized by minimizing the averaged cost $\mathbf{C}(x)$.

C. Iterative Optimization

We minimize the cost map \mathbf{C} in an iterative manner. At each iteration, the optimizer updates the results as

$$\mathbf{D}^{t+1} \leftarrow \mathbf{D}^t + \Delta \mathbf{D}^t, \quad \mathbf{T}_i^{t+1} \leftarrow \mathbf{T}_i^t + \Delta \mathbf{T}_i^t, \quad (3)$$

where $\Delta \mathbf{D}$ and $\Delta \mathbf{T}_i$ are the updates. Inspired by [16], we utilize a gated recurrent unit to compute these updates, since a GRU can memorize the status of previous steps and can effectively exploit temporal information during the optimization process.

1) *Initialization.*: The depth map and relative poses are initialized by two different heads on top of the feature extraction network. The depth head consists of two convolutional layers, while the pose head is further equipped with an additional average pooling layer. The hidden state of the GRU is initialized by the contextual features with the tanh function as activation.

2) *Recurrent Update.*: We design two GRU modules, one for updating the depth and the other one for updating the camera pose. Each GRU module receives the current cost map \mathbf{C}^t and the current estimated variables \mathbf{V}^t (depth map \mathbf{D}^t or camera pose \mathbf{T}^t) and outputs an incremental update $\Delta \mathbf{V}^t$ to update the results as $\mathbf{V}^{t+1} = \mathbf{V}^t + \Delta \mathbf{V}^t$.

Specifically, we first project the variable \mathbf{V}^t and the cost map \mathbf{C}^t into the feature space with \mathcal{P}_v and \mathcal{P}_c , both of which are composed of two convolutional layers. We then concatenate $\mathcal{P}_v(\mathbf{V}^t)$, $\mathcal{P}_c(\mathbf{C}^t)$, and the image contextual feature \mathcal{F}^c to form a tensor \mathbf{M}^t as the input. The structure inside each GRU unit is as follows:

$$\begin{aligned} z^{t+1} &= \sigma(\text{Conv}_{5 \times 5}([h^t, \mathbf{M}^t], W_z)) \\ r^{t+1} &= \sigma(\text{Conv}_{5 \times 5}([h^t, \mathbf{M}^t], W_r)) \\ \tilde{h}^{t+1} &= \tanh(\text{Conv}_{5 \times 5}([r^{t+1} \odot h^t, \mathbf{M}^t], W_h)) \\ h^{t+1} &= (1 - z^{t+1}) \odot h^t + z^{t+1} \odot \tilde{h}^{t+1}, \end{aligned} \quad (4)$$

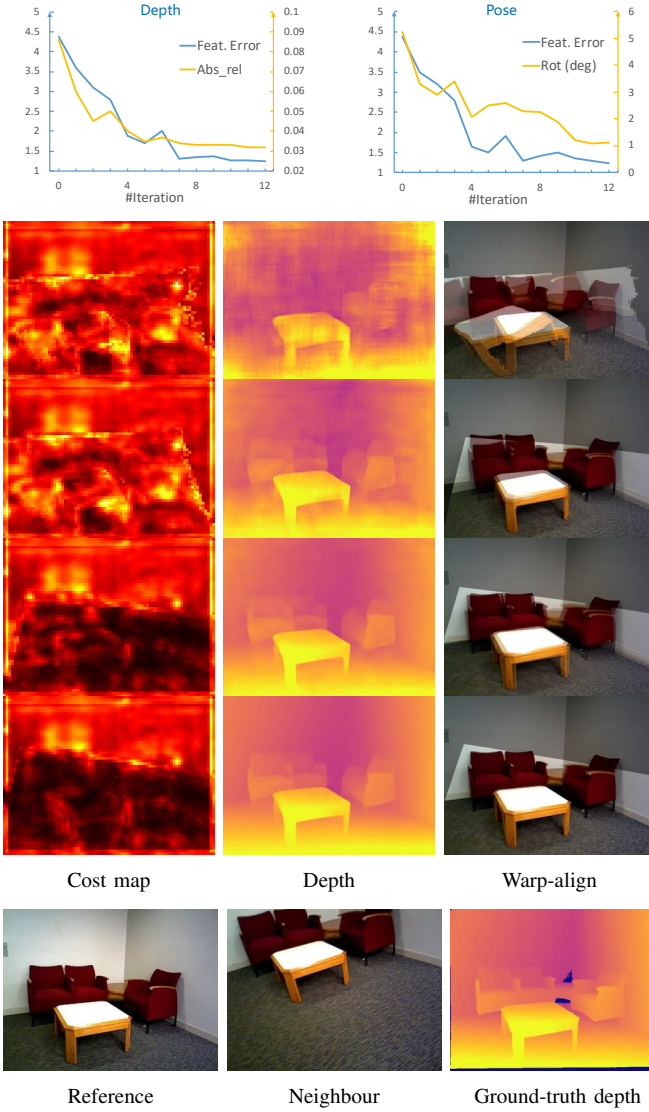


Fig. 4: The first row shows the feature-metric error (the cost), depth error, and camera pose error decrease with regards to the iterations of the GRU optimizer. In the middle we show the estimated cost map, depth map, and a warp-aligned image at each iteration of the GRU optimizer. The warp-aligned image warps a neighbor image onto the reference image using the estimated depth and camera pose. It becomes clearer at later iterations, indicating a better result.

where $\text{Conv}_{5 \times 5}$ represents a separable 5×5 convolution, \odot is the element-wise multiplication, σ and \tanh are the sigmoid and the tanh activation functions. Finally, the depth maps or the camera poses are predicted from the hidden state h^t by the same structures as the initial depth or pose head in Sec. III-C1.

With this optimizer, starting from an initial solution, the estimated depths and poses are iteratively refined by the optimization iterations and eventually converge to the final results as, $\mathbf{D}^t \rightarrow \mathbf{D}^*$ and $\mathbf{T}_i^t \rightarrow \mathbf{T}_i^*$.

3) *Alternate Optimization*: After defining the structure of the GRU module, we update the depth map \mathbf{D}^t and camera poses \mathbf{T}_i^t alternately with m stages. As shown in Fig. 3, at each stage, we first freeze the camera pose and update the depth map as $\mathbf{D}^{t+1} = \mathbf{D}^t + \Delta \mathbf{D}^t$, which is repeated by n times. We then freeze the depth map \mathbf{D} and update the camera poses with $\mathbf{T}_i^{t+1} = \mathbf{T}_i^t + \Delta \mathbf{T}_i^t$, which is also repeated by n times.

This alternative optimization leads to more stable optimization and easier training empirically. In experiments, we fix m at 3 and n at 4, unless specified otherwise.

To gain more insights into the recurrent process and demonstrate the GRU module behaves as a recurrent optimizer, we visualize how the feature-metric error decreases over the GRU iterations on the first row of Fig. 4. The blue curves are the plots of the feature-metric cost over iterations, while the yellow curves are the errors in depth map and relative camera poses compared with a known ground truth. In general, the feature-metric error decreases, but not strictly monotonically decreases. The error might increase and then decrease again, indicating our optimizer has the capability of jumping out of local minimums.

The last two rows of Fig. 4 further visualize the intermediate results over the iterations. At the bottom are the two input images and the ground truth depth map. In the middle, shown are the cost map \mathbf{C} , estimated depth map, and a warp-aligned image that is composed by warping the neighbor image according to the estimated depth and camera pose and superimpose with the reference image. From top to bottom we can see the warp-aligned image becomes sharper, indicating higher quality of estimated depth and camera pose.

D. Training Loss

Our method can be applied to both supervised and self-supervised V2D problem. In the following, we introduce the training loss for both schemes.

1) *Supervised Case*: When the ground truth is available, we supervise the training by the depth and pose errors.

Depth supervision $\mathcal{L}_{\text{depth}}$ computes the L1 distance between the predicted depth map \mathbf{D} and the ground-truth depth map $\hat{\mathbf{D}}$ in each stage.

Pose supervision $\mathcal{L}_{\text{pose}}$ is defined as the following according to the ground truth depth $\hat{\mathbf{D}}$ and pose $\hat{\mathbf{T}}_i$:

$$\mathcal{L}_{\text{pose}} = \sum_{s=1}^m \sum_x \gamma^{m-s} \|\pi(\mathbf{T}_i^s \circ \pi^{-1}(x, \hat{\mathbf{D}}(x))) - \pi(\hat{\mathbf{T}}_i \circ \pi^{-1}(x, \hat{\mathbf{D}}(x)))\|_1. \quad (5)$$

This loss computes the image re-projection of a pixel x according to the estimated camera pose \mathbf{T}_i^s and the true pose $\hat{\mathbf{T}}_i$ in each stage. The pose loss is the distance between these two projections, which is in the image space and insensitive to different scene scales.

2) *Self-supervised Case*: When the ground truth is not available, we borrow the loss defined in [47] for self-supervised training. Specifically, the supervision signal comes from geometric constraints and consists of two terms, a photometric loss and a smoothness loss.

Photometric loss $\mathcal{L}_{\text{photo}}$ measures the similarity of the reconstructed images $\{\mathbf{I}_i\}_{i=1}^N$ to the reference image \mathbf{I}_0 . Here, the reconstructed images \mathbf{I}_i are generated by warping the input image \mathbf{I}_i according to the depth \mathbf{D} and pose \mathbf{T}_i to reconstruct \mathbf{I}_0 . This similarity is measured by the structural similarity (SSIM) [48] with L1 loss as

$$\mathcal{L}_{\text{photo}} = \alpha \frac{1 - \text{SSIM}(\mathbf{I}'_i, \mathbf{I}_0)}{2} + (1 - \alpha) \|\mathbf{I}'_i - \mathbf{I}_0\|_1, \quad (6)$$

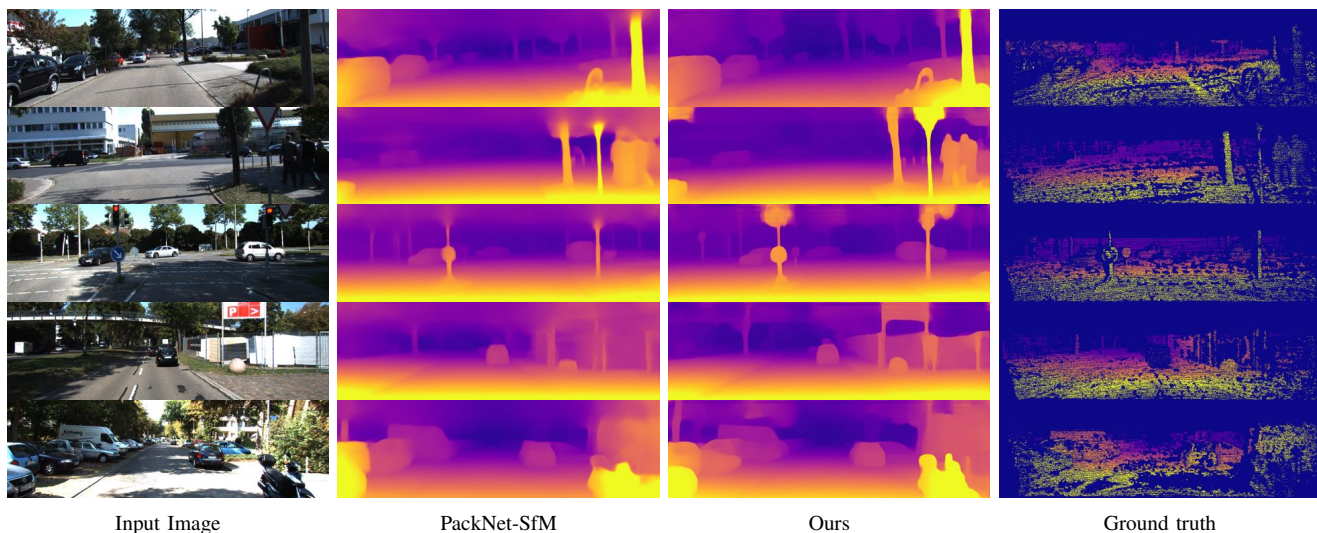


Fig. 5: Qualitative results on the KITTI dataset.

Method	Input	Supervised	GT type	Abs Rel	Sq Rel	RMSE	RMSE _{log}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
MonoDepth [24]	M→O	✗	Improved	0.090	0.545	3.942	0.137	0.914	0.983	0.995
PackNet-SfM [2]	M→O	✗	Improved	0.071	0.359	3.153	0.109	0.944	0.990	0.997
DRO (ours)	Multi	✗	Improved	0.057	0.342	3.201	0.123	0.952	0.989	0.996
Kuznetsov et al. [44]	One	✓	Improved	0.089	0.478	3.610	0.138	0.906	0.980	0.995
DORN [1]	One	✓	Improved	0.072	0.307	2.727	0.120	0.932	0.984	0.995
PackNet-SfM [2]	M→O	✓	Improved	0.064	0.300	3.089	0.108	0.943	0.989	0.997
BANet [3]	Multi	✓	Improved	0.083	—	3.640	0.134	—	—	—
DeepV2D (2-view) [4]	Multi	✓	Improved	0.064	0.350	2.946	0.120	0.946	0.982	0.991
Wang et al. [21]	Multi	✓	Improved	0.055	0.224	2.273	0.091	0.956	0.984	0.993
DRO (ours)	Multi	✓	Improved	0.047	0.199	2.629	0.082	0.970	0.994	0.998
SfMLearner [33]	M→O	✗	Velodyne	0.208	1.768	6.856	0.283	0.678	0.885	0.957
CCNet [26]	M→O	✗	Velodyne	0.140	1.070	5.326	0.217	0.826	0.941	0.975
GLNet [45]	M→O	✗	Velodyne	0.135	1.070	5.230	0.210	0.841	0.948	0.980
MonoDepth [24]	M→O	✗	Velodyne	0.115	0.882	4.701	0.190	0.879	0.961	0.982
PackNet-SfM [2]	M→O	✗	Velodyne	0.107	0.803	4.566	0.197	0.876	0.957	0.979
ManyDepth [46]	M→O	✗	Velodyne	0.093	0.715	4.245	0.172	0.909	0.966	0.983
DRO (ours)	Multi	✗	Velodyne	0.088	0.797	4.464	0.212	0.899	0.959	0.980
PackNet-SfM [2]	M→O	✓	Velodyne	0.090	0.618	4.220	0.179	0.893	0.962	0.983
DRO (ours)	Multi	✓	Velodyne	0.073	0.528	3.888	0.163	0.924	0.969	0.984

TABLE I: Quantitative results on the KITTI dataset. Eigen split is used for evaluation and seven widely used metrics are reported. Results on two ground-truth types are displayed since different methods are evaluated using different types. ‘M→O’ means multiple images are used in the training while one image is used for inference.

where α is a weighting factor. For the fusion of multiple photometric losses, we also take the strategies defined in [47], which adopts a minimum fusion and masks stationary pixels.

Smoothness loss $\mathcal{L}_{\text{smooth}}$ encourages adjacent pixels to have similar depths and is defined as:

$$\mathcal{L}_{\text{smooth}} = |\partial_x \mathbf{D}| e^{-|\partial_x \mathbf{I}_0|} + |\partial_y \mathbf{D}| e^{-|\partial_y \mathbf{I}_0|}. \quad (7)$$

IV. EXPERIMENTS

A. Implementation Details

Our work is implemented in Pytorch experimented with on Nvidia GTX 2080 Ti GPUs. The network is optimized end-to-end with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The training runs for 50 epochs with the learning rate reduced from 2×10^{-4} to 5×10^{-5} . For the supervised training, we use the

losses described in Sec. III-D1 with γ set as 0.85. For the self-supervised training, the losses are depicted in Sec. III-D2 with α set as 0.85 and λ set as 0.01.

B. Datasets

KITTI dataset. The KITTI dataset is a widely used benchmark with outdoor scenes captured from a moving vehicle. We adopt the training/testing split proposed by Eigen et al. [7] with 22,600 training images and 697 testing images. There are two types of ground-truth depth. One is the original Velodyne Lidar points which are quite sparse. The other one is the improved annotated depth map, which uses five successive images to accumulate the Lidar points and stereo images to handle moving objects. For the improved depth type there are 652 images for testing.

Method	Supervised	Abs Rel	Sq Rel	RMSE	RMSE _{log}	SI Inv	Rot (deg)	Tr (deg)	Tr (cm)	Time (s)
Photometric BA [12]	✓	0.268	0.427	0.788	0.330	0.323	4.409	34.36	21.40	–
DeMoN [5]	✓	0.231	0.520	0.761	0.289	0.284	3.791	31.626	15.50	–
BANet [3]	✓	0.161	0.092	0.346	0.214	0.184	1.018	20.577	3.39	–
DeepV2D (2-view) [4]	✓	0.069	0.018	0.196	0.099	0.097	0.692	11.731	1.902	0.95
DRO (ours)	✗	0.140	0.127	0.496	0.212	0.210	0.691	11.702	1.647	0.11
DRO (ours)	✓	0.053	0.017	0.168	0.081	0.079	0.473	9.219	1.160	0.11

TABLE II: Quantitative results on the ScanNet dataset. Five metrics of the depth and three metrics of the pose are reported.

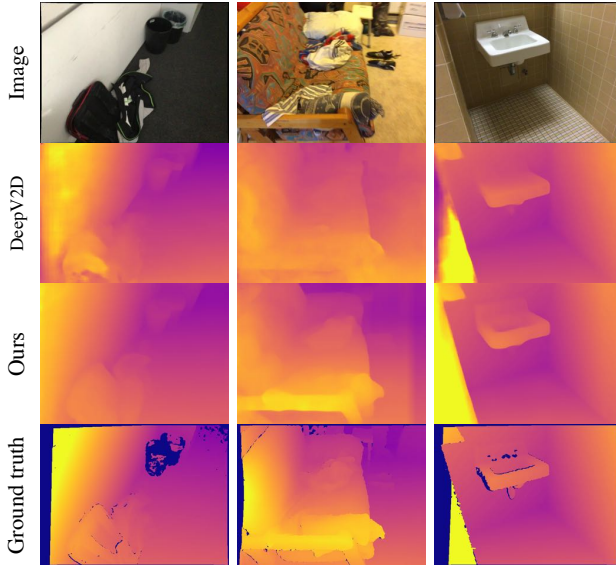


Fig. 6: Qualitative results on the ScanNet dataset.

ScanNet dataset. ScanNet [49] is a large indoor dataset with 1,513 RGB-D videos in 707 distinct environments. The raw data is captured from a depth camera. The depth maps and camera poses are obtained from RGB-D 3D reconstruction. We use the training/testing split proposed by [3] with 2,000 test image pairs from 90 scenes.

C. Evaluation

Evaluation on KITTI. For outdoor scenes, we present the results of our method and some previous methods on the KITTI dataset in TABLE I. State-of-the-art single-frame depth estimation methods [1, 44] and deep V2D methods [2–4, 24, 26, 33, 45] are listed. For a fair comparison, all V2D methods are evaluated under the two-view setting. From the results, our approach outperforms other methods by a large margin in both the supervised setting and the self-supervised setting. Also, the performance of our *self-supervised* model already surpasses most of previous *supervised* methods. The qualitative results of these outdoor scenes are shown in Fig. 5, from which we can see our approach estimates better depth for distant and small-size or thin objects, e.g., people, motorbikes, and guideposts. Also, we predict sharper edges at object boundaries. Thin structures are usually recovered by fine updates in the last few iterations.

Evaluation on ScanNet. For indoor scenes, we evaluate our method on the ScanNet dataset in TABLE II. For a fair comparison, all methods are evaluated under the two-view setting since there are only 2 images in the test split. The results of Photometric BA and DeMoN are cited from [3]. The

Setting	Abs Rel	Sq Rel	RMSE	R _{log}	1.25	1.25 ²	
w/o GRU	0.058	0.258	2.953	0.097	0.955	0.992	
w/o Alter	0.055	0.247	2.952	0.094	0.959	0.992	
w/o Cost	0.065	0.324	3.270	0.112	0.940	0.988	
Cost volume	0.049	0.214	2.804	0.086	0.966	0.994	
Full-setting	0.047	0.199	2.629	0.082	0.970	0.994	
DRO ⁺	0.042	0.151	2.294	0.072	0.978	0.996	
Infer iterations	0	0.094	0.529	4.014	0.150	0.891	0.974
	4	0.059	0.266	2.992	0.099	0.951	0.992
	8	0.049	0.208	2.687	0.084	0.968	0.994
	16	0.046	0.198	2.623	0.081	0.970	0.994
24	0.046	0.199	2.626	0.082	0.970	0.994	

TABLE III: Ablation study of the depth accuracy on KITTI dataset. The first six metrics of those used in TABLE I are reported here.

Setting	Abs Rel	Sq Rel	RMSE	Rot	Tr (deg)	Tr (cm)
w/o C, G, A	0.075	0.026	0.215	3.300	61.577	6.405
+ C	0.067	0.023	0.200	0.828	15.769	2.120
+ C + G	0.061	0.020	0.188	0.683	13.020	1.751
Full-setting	0.053	0.017	0.168	0.473	9.219	1.160
DRO ⁺	0.051	0.015	0.159	0.491	8.715	1.188
DRO _{5view}	0.049	0.015	0.157	0.463	8.871	1.139

TABLE IV: Ablation study of the depth and pose accuracy on ScanNet. “C”, “G”, and “A” denote Cost, GRU, and Alternate update. DRO⁺ refers to the high-resolution model and DRO_{5view} refers to using five images for inference.

results show that our model outperforms previous methods on both depth accuracy and pose accuracy. Our *self-supervised* model is able to predict the results that are comparable to previous *supervised* methods, especially on the pose accuracy. Among previous methods, DeepV2D performs best but it requires pre-training a complex pose solver first. Also, the inference time of their method is much longer than ours. Even using five views their performance is still not comparable to ours, with a larger depth error of 0.057. From the qualitative results shown in Fig. 6, our model predicts the finer depth of the indoor objects and is robust in clutter. Inference with 5 images further reduces the depth error from 0.053 to 0.049, as shown in TABLE IV.

D. Ablation Study

To better understand the individual components, we evaluate each module by an ablation study and present the results in TABLE III and TABLE IV.

GRU module. The core module of our method is the recurrent optimizer. To see its effectiveness in optimization, we replace the GRU block with three convolutional layers. In the training in KITTI dataset, we found the depth error decreases to 0.058 in the first a few epochs but then the network diverges.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

This demonstrates that by leveraging the historical information in a recurrent optimizer, not only a superior optimum could be reached, but also the optimization would be more stable. We think this is because the GRU modules could remember the historical updating information, e.g., it can remember the historical updating direction and magnitude, which is helpful for guiding the next update.

Alternate update. It is important to alternately update the depth and camera poses to decouple their influence in the feature-metric error. To verify this, we train a model where the optimizer predicts the updates for depth and camera poses simultaneously. The depth accuracy in this setting is reduced a lot. This proves the importance of the alternative optimization.

Cost Volume. One of the advantages of our method is that we do not need a heavy cost volume for optimization. Here, we replace the feature-metric cost map C with a $H \times W \times 64$ cost volume. The cost volume is in a cascaded structure, i.e., the depth range of the volume is dynamically adjusted over iterations for better results. From the results shown in TABLE III, the performance of using this heavy cost volume is similar to using the cost map, which proves that employing information in temporal domain can make up the lack of neighborhood information in spatial domain. Also, we test the performance of a model without the cost input. As expected, the error of the depth estimation is large since the optimizer loses the objective to minimize.

Iteration times. Until now, we only use 12 iterations of recurrent optimization for all experiments. We could also vary the number of iterations during inference. Here, we test different iteration numbers in the inference. Zero iteration means we do not update the initial depth and pose at all. According to the results in TABLE III, our optimizer already outputs decent results after 4 iterations and predicts accurate results after 8 iterations, after which the depths are further refined with more iterations. This demonstrates that our optimizer has learned to optimize the estimation step by step. A model trained with a fixed number of iterations can be applied with more iterations in real applications to obtain finer results.

Initialization. To inspect how robust our method is to the initial depth and pose, we add the noise to the initialization and see how our network performs. TABLE V shows our results with different levels of Gaussian noise in the predicted initial depth and camera poses. The parameters $N_d\{\sigma\}$, $N_t\{\sigma\}$, $N_r\{\sigma\}$ in the first column are the standard deviations of the added noise in depth, camera center, and orientation respectively. Note the scene depth range is 10m. Thus, $N_d\{\sigma=2m\}$ is a very large noise. In the last three row, we initialize with a constant depth at 5m or zero camera motion. For each case, we show the errors in the initial and optimized results. Our method always generates high-quality results despite the noisy initialization. This demonstrates our GRU-based optimization has a large convergence basin.

High-resolution. Since our method is efficient in computation and memory consumption, we can use high-resolution features in the optimization. In previous experiments, the features \mathcal{F} is of $\frac{1}{8}$ size of the input image. Here we add an FPN structure on the features generated by the encoder, to get features of $\frac{1}{4}$ size and $\frac{1}{2}$ size. Then we adopt features of $\frac{1}{4}$

Noise Setting	Initialization			After Optimization		
	AbsRel	Rot	Tr	AbsRel	Rot	Tr
w/o noise	0.109	3.215	54.257	0.053	0.473	9.219
$N_d\{1m\}N_t\{1cm\}N_r\{1^\circ\}$	0.117	3.657	59.299	0.053	0.478	9.546
$N_d\{1.5m\}N_t\{2cm\}N_r\{2^\circ\}$	0.130	4.661	65.444	0.059	0.520	10.369
$N_d\{2m\}N_t\{5cm\}N_r\{5^\circ\}$	0.157	8.699	75.944	0.089	2.101	24.500
$D_0 = 5m$	0.232	3.215	54.257	0.068	0.684	11.215
$t = r = [0, 0, 0]$	0.109	3.577	90.000	0.053	0.484	9.418
$D_0 = 5m, t = r = [0, 0, 0]$	0.232	3.577	90.000	0.068	0.677	12.056

TABLE V: Noisy initialization on ScanNet **without retraining**.

Models	Resolution	Memory	Time	Abs Rel	
DeepV2D [4]	192×1088	6.75 G	1.49 s	0.064	
DRO	iterate 12	320×960	1.16 G	0.12 s	0.047
	iterate 4	320×960	1.15 G	0.049 s	0.059

TABLE VI: Efficiency experiments on the KITTI dataset.

size in the 2nd stage, and features of $\frac{1}{2}$ size in the 3rd stage in Fig. 3. This high-resolution version obtains better performance and is denoted by DRO⁺, as is shown in TABLE III and TABLE IV.

E. Run-time and Memory Efficiency

We compare our method with DeepV2D [4] in run-time and memory efficiency in TABLE VI. All experiments are conducted on the KITTI dataset with an Nvidia GTX 2080 Ti GPU. Compared to DeepV2D, our method achieves similar depth accuracy with only 4 iterations, which takes only about 1/6 of GPU memory and 1/30 of inference time comparing with DeepV2D. Even with 12 iterations, our method is still more than 10 times faster and reduces about 26% depth errors. More experiments are shown in the supplements.

F. Limitations

Our framework works based on the correspondence between two frames, in which case the performance decreases when the overlap between the image pair is small. Also, the estimated depth on the occluded regions may be inaccurate since there is no correspondence to build.

V. CONCLUSION

We propose a deep recurrent optimizer for addressing the video-to-depth problem. Two gated recurrent units have been introduced to optimize scene structures and camera poses respectively. Our optimizer avoid computing a cost volume or gradients by exploiting temporal information during the optimization process. The experiments demonstrate our method outperforms previous methods on both outdoor and indoor datasets, in both supervised and self-supervised settings.

REFERENCES

- [1] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018, pp. 2002–2011.
- [2] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *CVPR*, 2020, pp. 2485–2494.
- [3] C. Tang and P. Tan, "Ba-net: Dense bundle adjustment networks," in *ICLR*, 2019.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

- [4] Z. Teed and J. Deng, “Deepv2d: Video to depth with differentiable structure from motion,” in *ICLR*, 2020.
- [5] B. Ummerhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, “Demon: Depth and motion network for learning monocular stereo,” in *CVPR*, 2017, pp. 5038–5047.
- [6] X. Wei, Y. Zhang, Z. Li, Y. Fu, and X. Xue, “Deepstfm: Structure from motion via deep bundle adjustment,” in *ECCV*. Springer, 2020, pp. 230–247.
- [7] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *NIPS*, 2014, pp. 2366–2374.
- [8] D. Jayaraman and K. Grauman, “Learning image representations tied to ego-motion,” in *ICCV*, 2015, pp. 1413–1421.
- [9] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, “From big to small: Multi-scale local planar guidance for monocular depth estimation,” *arXiv preprint arXiv:1907.10326*, 2019.
- [10] Z. Yu and S. Gao, “Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement,” in *CVPR*, 2020, pp. 1949–1958.
- [11] R. Clark, M. Bloesch, J. Czarnowski, S. Leutenegger, and A. J. Davison, “Ls-net: Learning to solve nonlinear least squares for monocular stereo,” *arXiv preprint arXiv:1809.02966*, 2018.
- [12] J. Engel, T. Schöps, and D. Cremers, “Lsd-slam: Large-scale direct monocular slam,” in *ECCV*. Springer, 2014, pp. 834–849.
- [13] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” *PAMI*, vol. 40, no. 3, pp. 611–625, 2017.
- [14] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, “Codeslam—learning a compact, optimisable representation for dense visual slam,” in *CVPR*, 2018, pp. 2560–2568.
- [15] M. Bloesch, T. Laidlow, R. Clark, S. Leutenegger, and A. J. Davison, “Learning meshes for dense visual slam,” in *ICCV*, 2019, pp. 5855–5864.
- [16] Z. Teed and J. Deng, “Raft: Recurrent all-pairs field transforms for optical flow,” in *ECCV*. Springer, 2020, pp. 402–419.
- [17] H. Zhou, B. Ummerhofer, and T. Brox, “Deeptam: Deep tracking and mapping,” in *ECCV*, 2018, pp. 822–838.
- [18] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, “Cascade cost volume for high-resolution multi-view stereo and stereo matching,” in *CVPR*, 2020, pp. 2495–2504.
- [19] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, “Mvsnet: Depth inference for unstructured multi-view stereo,” in *ECCV*, 2018, pp. 767–783.
- [20] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.
- [21] J. Wang, Y. Zhong, Y. Dai, S. Birchfield, K. Zhang, N. Smolyanskiy, and H. Li, “Deep two-view structure-from-motion revisited,” in *CVPR*, 2021, pp. 8953–8962.
- [22] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, “Newcrfs: Neural window fully-connected crfs for monocular depth estimation,” in *CVPR*, 2022.
- [23] J.-W. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, “Unsupervised scale-consistent depth and ego-motion learning from monocular video,” in *NIPS*, 2019.
- [24] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in *ICCV*, 2019, pp. 3828–3838.
- [25] R. Mahjourian, M. Wicke, and A. Angelova, “Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints,” in *CVPR*, 2018, pp. 5667–5675.
- [26] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, “Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation,” in *CVPR*, 2019, pp. 12 240–12 249.
- [27] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, “Learning depth from monocular videos using direct methods,” in *CVPR*, 2018, pp. 2022–2030.
- [28] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, “Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos,” in *CVPR*, 2019, pp. 8071–8081.
- [29] W. Yuan, Y. Zhang, B. Wu, S. Zhu, P. Tan, M. Y. Wang, and Q. Chen, “Stereo matching by self-supervision of multiscope vision,” in *IROS*. IEEE, 2021, pp. 5702–5709.
- [30] N. Yang, R. Wang, J. Stuckler, and D. Cremers, “Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry,” in *ECCV*.
- [31] Z. Yin and J. Shi, “Geonet: Unsupervised learning of dense depth, optical flow and camera pose,” in *CVPR*, 2018, pp. 1983–1992.
- [32] 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 *CVPR*, 2018, pp. 340–349.
- [33] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *CVPR*, 2017, pp. 1851–1858.
- [34] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, “Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance,” in *ECCV*. Springer, 2020, pp. 582–600.
- [35] Y. Zou, Z. Luo, and J.-B. Huang, “Df-net: Unsupervised joint learning of depth and flow using cross-task consistency,” in *ECCV*, 2018, pp. 36–53.
- [36] W. Zhao, S. Liu, Y. Shu, and Y.-J. Liu, “Towards better generalization: Joint depth-pose learning without posenet,” in *CVPR*, 2020, pp. 9151–9161.
- [37] J.-W. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, “Unsupervised scale-consistent depth and ego-motion learning from monocular video,” in *NIPS*, 2019.
- [38] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and Z. Kolter, “Differentiable convex optimization layers,” in *NIPS*, 2019, pp. 9558–9570.
- [39] B. Amos and J. Z. Kolter, “Optnet: Differentiable optimization as a layer in neural networks,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 136–145.
- [40] J. Adler and O. Öktem, “Solving ill-posed inverse problems using iterative deep neural networks,” *Inverse Problems*, vol. 33, no. 12, p. 124007, 2017.
- [41] Y. Chen, M. W. Hoffman, S. G. Colmenarejo, M. Denil, T. P. Lillicrap, M. Botvinick, and N. Freitas, “Learning to learn without gradient descent by gradient descent,” in *ICML*. PMLR, 2017, pp. 748–756.
- [42] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang, “End-to-end learning of motion representation for video understanding,” in *CVPR*, 2018, pp. 6016–6025.
- [43] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [44] Y. Kuznetsov, J. Stückler, and B. Leibe, “Semi-supervised deep learning for monocular depth map prediction,” in *CVPR*, 2017, pp. 2215–2223.
- [45] Y. Chen, C. Schmid, and C. Sminchisescu, “Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera,” in *ICCV*, 2019, pp. 7063–7072.
- [46] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, “Self-supervised multi-frame monocular depth,” in *CVPR*, 2021, pp. 1164–1174.
- [47] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in *ICCV*, 2019, pp. 3828–3838.
- [48] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, “Image quality assessment: from error visibility to structural similarity,” *TIP*, vol. 13, no. 4, pp. 600–612, 2004.
- [49] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in *CVPR*, 2017.