

DualRefine: Self-Supervised Depth and Pose Estimation Through Iterative Epipolar Sampling and Refinement Toward Equilibrium

Antyanta Bangunharcana¹, Ahmed Magd², Kyung-Soo Kim¹

¹Mechatronics, Systems, and Control Laboratory, ²Vehicular Systems Design and Control Lab
 Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

{antabangun, a.magd, kyungsoo}@kaist.ac.kr

Abstract

Self-supervised multi-frame depth estimation achieves high accuracy by computing matching costs of pixel correspondences between adjacent frames, injecting geometric information into the network. These pixel-correspondence candidates are computed based on the relative pose estimates between the frames. Accurate pose predictions are essential for precise matching cost computation as they influence the epipolar geometry. Furthermore, improved depth estimates can, in turn, be used to align pose estimates.

Inspired by traditional structure-from-motion (SfM) principles, we propose the DualRefine model, which tightly couples depth and pose estimation through a feedback loop. Our novel update pipeline uses a deep equilibrium model framework to iteratively refine depth estimates and a hidden state of feature maps by computing local matching costs based on epipolar geometry. Importantly, we used the refined depth estimates and feature maps to compute pose updates at each step. This update in the pose estimates slowly alters the epipolar geometry during the refinement process. Experimental results on the KITTI dataset demonstrate competitive depth prediction and odometry prediction performance surpassing published self-supervised baselines¹.

1. Introduction

The optimization of the coordinates of observed 3D points and camera poses forms the basis of structure-from-motion (SfM). Estimation of both lays the foundation for robotics [34, 35, 75], autonomous driving [20], or AR/VR applications [60]. Traditionally, however, SfM techniques are susceptible to errors in scenes with texture-less regions, dynamic objects, *etc.* This has motivated the development of deep learning models that can learn to predict depth from monocular images [14, 15, 18, 48, 50]. These models can

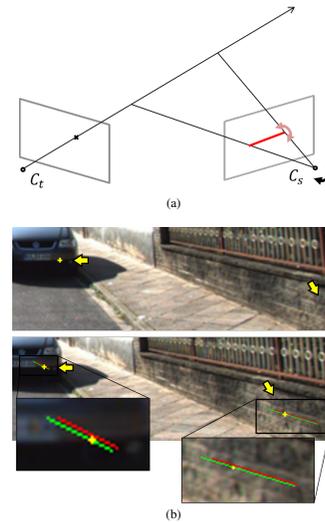


Figure 1. (a) The estimated pose of a camera affects the epipolar geometry. (b) The epipolar line in the source image, calculated from yellow points in the target image, for the PoseNet-based [43] initial pose regression (red) and our refined pose (green). The yellow point in the source image is calculated based on our final depth and pose estimates.

accurately predict depth based solely on image cues, without requiring geometric information.

In recent years, self-supervised training of depth and pose models has become an attractive method, as it alleviates the need for ground truth while demonstrating precision comparable to those of supervised counterparts [7, 19, 22, 23, 26, 28, 30, 61, 70, 74, 83, 87, 98, 106, 108]. Such an approach uses depth and pose predictions to synthesize neighboring images in a video sequence and enforce consistency between them. As the image sequence is also available at test time, recent self-supervised methods also study the use of multiple frames during inference [91]. These typically involve the construction of cost volumes from multiple views to compute pixel correspondences, bearing similarities to (multi-view) stereo models [4, 44, 77]. By incorporating multi-frame data, geometric information is integrated to make depth predic-

¹<https://github.com/antabangun/DualRefine>

tions, improving the performance as well as the robustness. In such a multi-frame matching-based model, the accuracy of matching costs computation is essential. Recent work in DepthFormer [29] demonstrates its importance, as they designed a Transformer [84]-based module to improve matching costs and achieve state-of-the-art (SoTA) depth accuracy. However, their approach came with a large memory cost.

Unlike stereo tasks, the aforementioned self-supervised multi-frame models do not assume known camera poses and use estimates learned by a teacher network, typically a PoseNet [43]-based model. This network takes two images as input and regresses a 6-DoF pose prediction. As the estimated pose affects the computation of epipolar geometry (Fig. 1(a)), the accuracy of the pose estimates is crucial to obtain accurate correspondence matches between multiple frames. However, as noted in recent studies [72], pure learning-based pose regression generally still lags behind its traditional counterpart, due to the lack of geometric reasoning. By refining the pose estimates, we can improve the accuracy of the matching costs, potentially leading to better depth estimates as well. In Fig. 1(b), we show that the epipolar lines calculated from the regressed poses do not align with our refined estimates. Conversely, a better depth prediction may lead to a better pose prediction. Thus, instead of building the cost volume once using regressed poses, we choose to perform refinements of both depth and pose in parallel and sample updated local cost volumes at each iteration. This approach is fundamentally inspired by traditional SfM optimization and is closely aligned with feedback-based models that directly couple depth and pose predictions [27].

In this work, we propose a depth and pose refinement model that drives both towards an equilibrium, trained in a self-supervised framework. We accomplish this by making the following contributions: **First**, We introduce an iterative update module that is based on epipolar geometry and direct alignment. We sample candidate matches along the epipolar line that evolves based on the current pose estimates. Then the sampled matching costs are used to infer per-pixel confidences that are used to compute depth refinements. The updated depth estimates are then used in direct feature-metric alignments to refine the pose updates towards convergence. As a result, our model can perform geometrically consistent depth and pose updates. **Second**, These updates refine the initial estimates made by the single-frame model. By doing so, we do not rely on full cost volume construction and base our updates only on local cost volumes, making it simpler, more memory efficient, and more robust. **Lastly**, we design our method within a deep equilibrium (DEQ) framework [3] to implicitly drive the predictions towards a fixed point. Importantly, DEQ allows for efficient training with low training memory, improving upon the huge memory consumption of previous work. With our proposed novel design, we show improved depth estimates through experiments that are com-

petitive with the SoTA models. Furthermore, our model demonstrates improved global consistency of visual odometry results, outperforming other learning-based models.

2. Related Work

2.1. Depth from a single image

The depth prediction problem for a single image is ill-posed due to the possibility of different 3D scenes projecting onto the same 2D image [32]. Nonetheless, humans can predict depth from a single image through experience. Motivated by this, numerous supervised neural network models were proposed to solve the monocular depth estimation task, starting with Eigen *et al.*'s paper [15]. Subsequently, performance was improved by modifying the model architecture [14, 48, 64, 94], training on large dataset [13, 53, 95, 100], designing robust loss functions [49, 103], and transforming the problem into a classification task [52]. However, supervisory depth estimation requires ground-truth depth maps, which are difficult to collect in large quantities and of high quality. This challenge is one of the main reasons why researchers are exploring semi-supervised training, where the model expects weak supervision, such as providing relative depth [12], camera poses [101], or utilizing synthetic data for training [1, 47, 62].

The need for weak supervision still presents limitations in generalizability and scalability, among other aspects. To address these constraints, research on self-supervised training techniques is gaining momentum. These techniques involve using geometry in stereo matching [19, 22] or with a sequence of single-camera images, as initially proposed by [106]. Monodepth2 [23] refined the idea of exploiting image sequences for training by using auto-masking and minimum reprojection losses to address occlusion and ego-motion issues. Further improvements were made by defining the problem as a classification task [25, 42], modifying the architecture [28], feature-based loss for regions with low texture [74, 101], or reducing artifacts from moving objects [11, 46, 80].

Our work is based on self-supervised monocular depth and employs them as a teacher and initial estimate.

2.2. Depth from multiple frames

Relying on single frames at test-time requires the model to make several assumptions about the scene's geometrical details. In contrast, multi-frame approaches, which leverage available temporal information and incorporate multi-view geometry, reduce the need for such assumptions.

Multi-frame depth prediction is closely related to stereo depth estimation, where neural networks convert input stereo images into depth maps, as demonstrated by [54, 63, 82]. Kendall *et al.* [44] achieved a significant improvement by constructing a plane-sweep stereo cost volume. Generally,

multi-view stereo (MVS) research is more relevant to our work, as it utilizes an unstructured collection of scene images, meaning that the pose between different images is not fixed. In studies like [36, 38, 58, 90, 97], it was common to combine the previously mentioned cost volumes with ground truth depth and camera poses for guidance. Consequently, these works require camera poses during inference [56, 92]. Although some research managed to relax the pose requirement during inference, training still necessitates this supervision [77, 82].

The methods discussed above, along with advances in other domains such as scale domain adaptation [31, 69, 104], view synthesis [23, 93], and other supervisory techniques [26, 30], have significantly impacted the performance of self-supervised approaches. ManyDepth [91], which is similar to our work, focuses on the drawbacks of using a single image during inference and proposes a flexible model that takes advantage of multiple frames at test time, if available. ManyDepth reduces artifacts from moving objects and temporarily stationary cameras by using a single-frame model as a teacher, resulting in improved depth map accuracy. More recently, DepthFormer [29] achieved substantial accuracy improvements by using Transformers [84] to obtain improved pixel matching costs.

Inspired by traditional bundle adjustments, we design a model that simultaneously solves for both depth and pose while incorporating many of the advancements mentioned earlier. Our approach is most similar to DRO [27]. However, compared to their approach, our model tightly integrates multi-view geometry into the iterative updates formulation and bases our refinements on the local epipolar geometry.

2.3. Iterative refinements

Iterative refinement has been employed to improve prediction quality in various learning tasks, including object detection [6, 24], optical flow estimation [37, 78], semantic segmentation [68, 102], and others [40, 71]. Some recent research has attempted to iterate the refinement process using deep convolutional networks [9, 21, 55, 57]. Other works train the same network repeatedly by utilizing the results of the previous iteration [99, 102]. In particular, RAFT [78] found success with its iterative refinement procedure for flow estimation. DEQ-flow [2] employed a deep equilibrium (DEQ) [3] framework to reduce the memory consumption of RAFT during training while maintaining accuracy.

A key component of our model is inspired by the iterative updates of RAFT and DEQ-flow. Instead of optical flow, our model refines depth and pose estimates in parallel. We design our refinement module to tightly couple the two predictions, considering the epipolar geometry of adjacent frames. With every update, the epipolar geometry is refined, which also results in a more accurate matching costs computation of pixel correspondences.

2.4. Pose estimation

Pose estimation is a crucial component in self-supervised monocular depth models. In many works, PoseNet-based models [43] take a pair of adjacent image frames and output a 6 DoF pose estimation. This class of models is straightforward, but often less accurate than their traditional counterparts [16, 17, 65] due to the absence of geometry constraints. Recent work in deep learning-based localization has adopted differentiable geometrically inspired designs within their models by using direct alignments [72, 85, 86] or geometric alignments based on optical flow [79]. These models demonstrate better generalization properties and accuracy. The use of self-supervised monocular depth has also been used to improve traditional odometry [8, 96]. However, our interest lies in learning to refine both depth and pose in the self-supervision pipeline. To improve the accuracy of pose estimation, we integrate direct alignment within the recurrent module, ensuring geometrically consistent prediction between depth and pose.

3. Method

Our model comprises two primary sub-modules. The first is a single frame self-supervised depth and pose estimator, building upon previous frameworks [23, 105], which we revisit in Sec. 3.1. This network serves as both a teacher and an initializer for the second sub-module, our proposed multi-frame network, presented in detail in Sec. 3.2.

3.1. Self-supervised depth and pose

We begin by describing the canonical self-supervised monocular depth estimation pipeline [23], which serves as the foundation for our approach. This depth training method assumes that a monocular camera with an intrinsic parameter K captures an image sequence of a scene. In this process, two networks are trained in parallel to estimate the per-pixel depth map of the images D and the relative poses between adjacent image frames. By warping neighboring images towards a shared target frame using these two predictions, self-supervised training can be performed by enforcing photometric consistency between the frames.

Given the depth map D of a target image and its relative pose with a source image $T_{t \rightarrow s}$, we can calculate the projection of each pixel $u = (x, y)$ of the target image onto the source image as follows:

$$z' u' = z' \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = K T_{t \rightarrow s} \left(D[u] K^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \right), \quad (1)$$

using the estimated depth at that pixel $D[u]$. The source images can then be warped towards the target frame by sampling the pixel values at the calculated projection

$$I_{s \rightarrow t}[u] = I_s \langle u' \rangle, \quad (2)$$

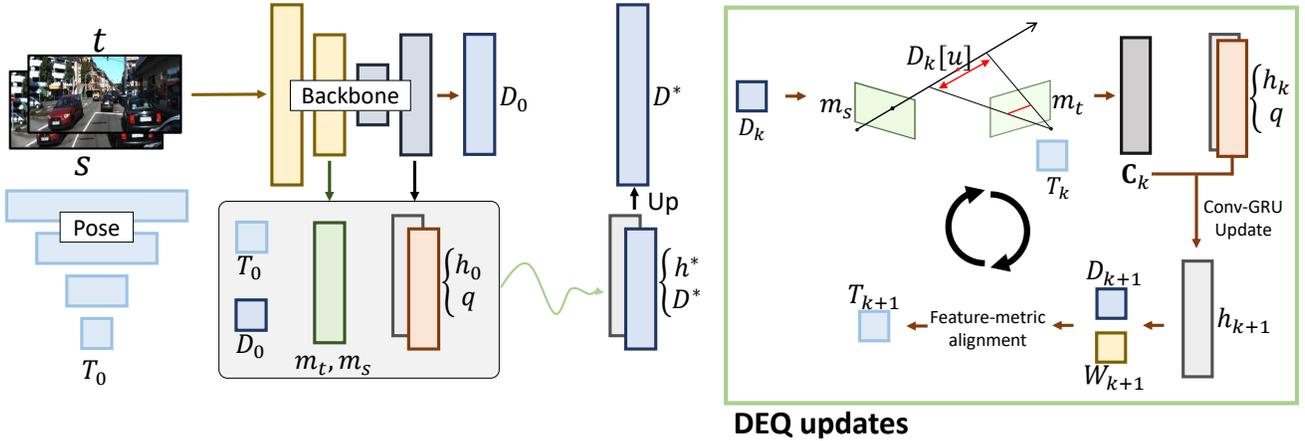


Figure 2. (a) The overall pipeline of the model. Given a pair of source and target images, the teacher model predicts an initial depth D_0 and pose T_0 , as well as initial hidden states that will be updated. DEQ-based alignments are then performed to find the fixed point and output the final predictions. (b) Each iteration in the update step takes the current depth and pose estimates. Matching costs are sampled along the current epipolar lines that evolves based on the pose estimates. The updates are computed by Conv-GRU. Then feature-metric alignment is used to obtain a geometrically consistent pose update.

where $\langle \cdot \rangle$ indicates bilinear interpolation, implemented using the spatial transformer network (STN) [39].

The self-supervised loss is calculated as a combination of the photometric error and the edge-aware smoothness loss:

$$\mathcal{L}_{self-sup} = \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s, \quad (3)$$

where the photometric error between the warped source image and the target image is calculated using the structural similarity loss, and the minimum error is taken between multiple warped images to account for occlusions. Interested readers can refer to [23]. The self-supervised loss is typically computed at multiple scales to stabilize the training.

In this paper, we train a monocular depth estimator and a pose estimation network using this pipeline, serving two purposes. First, following ManyDepth [91], we use these models as a teacher to constrain the multi-frame predictions in the presence of dynamic objects. Additionally, we employ them as an initializer for the multi-frame alignment network.

3.1.1 Monocular model

We build our monocular depth estimation model based on DIFFNet [105], a SoTA self-supervised single-frame estimator. We extract feature maps at multiple scales s from the target image using the HRNet architecture [88]. In accordance with DIFFNet, feature maps from multiple stages are accumulated in $F^{(1/2^s)}$. Then, we employ disparity decoders to make disparity predictions at scales $s = \{2, 3\}$. The pose estimation network follows the canonical PoseNet [43] architecture, taking two input images and outputting 6-DoF values, with a ResNet18 [33] backbone. The predicted disparity and pose estimation from these networks are used as a teacher to train our alignment sub-module, as was done in ManyDepth [91].

3.2. Deep equilibrium alignments

In our alignment sub-module, we assume that additional input from source image(s) is available, which can be used to refine the depth and pose estimates. In this work, we focus on using the image from the previous frame in the image sequence as our source image.

Our alignment module is formulated as a deep equilibrium model [3] that updates the hidden states, depth, and pose estimates to a fixed point. Specifically, at the fixed point z^* ,

$$(h^*, D^*, T^*) = z^* = U(z^*, x), \quad (4)$$

where z is composed of a hidden state h , the depth prediction D , and the pose prediction T . U represents our update function, refining the depth and pose alternatively. x is an input to the update module, obtained based on the epipolar geometry at each step, which we discuss in the next subsection. We perform these iterative updates at scale $s = 2$.

From the feature extraction output, we compute an initial hidden state for our recurrent updates $h^{[0]} = \tanh(H(F^{(1/4)}))$ and a context feature $q = Q(F^{(1/4)})$. Both H and Q are composed of a single residual block [33] followed by a convolutional layer.

3.2.1 Depth updates around local neighborhood

Local epipolar sampling. As discussed previously, our refinement is based on the pixel matching between the target image and the source image. We use the first two blocks of the HRNet feature extractor to extract unary features from the source image $m_s^{(1/4)}$ and the target image $m_t^{(1/4)}$, which we will use to calculate the matching costs.

Similarly to RAFT [78], our aim is to compute the matching values for candidate correspondences around the current

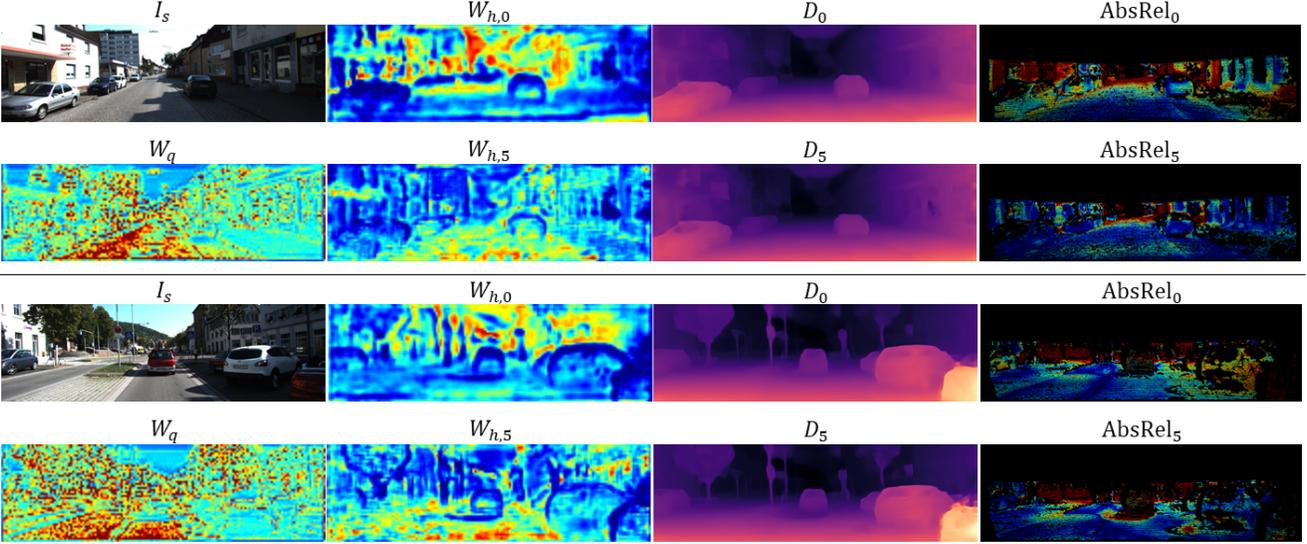


Figure 3. Qualitative results on KITTI data. I_s : input image; W_q , $W_{h,0}$, and $W_{h,5}$: confidence weights; D_0 , D_5 : disparity estimates; The Abs Rel error for the depth estimates.

prediction and use them as input to our update module at the update step k . Unlike RAFT, however, we perform matching along an epipolar line based on the pose estimate T_k . This is done by computing the projected coordinates u'_k in the source image obtained from (1) at the depths of interest. The depth candidates are computed to be the neighborhood of the current prediction. Instead of computing the all-pair correlation as was done in RAFT, we compute feature matching on the fly.

At each pixel u , we compute the local depth candidates as $D[u]_k \pm (i \times c \times n)$, with $i \in \mathbb{Z}$ and $i \leq r$. We set the sampling radius hyperparameter $r = 8$ in our experiments. In depth estimation, the error typically grows with distance. To account for this, we define $c = D[u]/C$ as a function of depth to make the sampling range dependent on the depth, where we set C as a trainable parameter. Following RAFT, to collect the matching information from a larger neighborhood, we sampled at multiple levels $n = \{1, 2, 3\}$. At each level, we bilinearly resize the matching feature map of the source image $m_s^{(1/4)}$ with a scale of $1/2^n$. Then, the matching features are sampled at the calculated corresponding set of coordinates u'_k , and the absolute differences with the target feature

$$\mathbf{C}_k[u] = |m_t[u] - m_s\langle u'_k \rangle| \quad (5)$$

are calculated and gathered. This provides us with a map \mathbf{C}_k that contains $n \times (2 \times r + 1)$ matching cost values at the corresponding depth candidates. We then encode these matching costs along with the depth using a two-layer convolutional neural network (CNN) to compute the input for the update module

$$x_k = [\text{CNN}_{\mathbf{C}}(\mathbf{C}_k), \text{CNN}_{D_x}(D_k), q], \quad (6)$$

where $[\cdot, \cdot]$ represents the concatenation.

Depth update. The update function calculates an updated hidden state h_{k+1} using the Conv-GRU block [5, 41, 78]. h_{k+1} is used to compute the depth updates. To stabilize training, we use the activation function $\tanh(\cdot)$ to bound the absolute update values for the depths to be within $r \cdot c$:

$$D_{k+1} = D_k + r \cdot c \cdot \tanh(\text{CNN}_{D_U}(h_{k+1})). \quad (7)$$

These updates are performed in an alternating fashion with the pose updates to reach the fixed point D^* . Using h^* , we compute a convex upsampling to obtain the final depth estimate at the input resolution.

3.2.2 Feature-metric pose alignments

In (1), the accuracy of pose estimation affects the calculation of the coordinates of pixels. Hence, a refined pose estimate would also improve the reliability of the matching costs. To refine the pose while being geometrically consistent, we perform our pose updates based on direct feature alignments [16, 17, 87]. These updates δ_k can be calculated by solving $H_k \delta_k = b_k$, where

$$H_k = \mathcal{J}_k^T \text{diag}(W_k) \mathcal{J}_k \text{ and } b_k = -\mathcal{J}_k^T \text{diag}(W_k) r_k \quad (8)$$

$$r_k[u] = m_s\langle u'_k \rangle - m_t[u],$$

and \mathcal{J} is the Jacobian with respect to the pose.

To compute this pose update, one could assign uniform confidence to every pixel in the image. However, in Eq. (8), additional confidence weights per pixel W can also be integrated. This is done for two reasons. First, the solution for the pose updates can be affected by dynamic objects

as well as inaccurate feature alignments that may occur in region with repeated textures. To account for this, the confidence weighting of the input context feature map W_q can be computed [72]. Specifically, we computed a confidence map $W_{t,s} = 1/(1 + \text{ReLU}(\text{CNN}_{W_q}(F_{t,s})))$ for both target and source images. The source confidence is warped towards the target frame using Eqs. (1) and (2). Finally, the confidence is computed as $W_q = W_t \cdot W_{s,\text{warped}}$. This confidence weight is only computed once to assign per-pixel confidence for the input images.

Second, the accuracy of pose updates would depend on the accuracy of the depth estimates. To obtain more accurate pose updates, we would like to assign more alignment confidence to the region with higher depth accuracy. We use a neural network to infer this confidence from the matching costs. Since the hidden states h have a history of these matching costs, another confidence weights $W_{h,k} = \text{CNN}_{W_h}(h_k)$ can be computed. Unlike the previous confidence map, this one is computed and evolves at every update step. By doing so, the network can use the depth predictions to guide the pose estimates towards convergence using the depth information. In our experiments, we investigate the use of each confidence weighting and combination of both $W_k = W_q W_{h,k}$.

Finally, we can compute the updated pose as $T_{k+1} = \exp(\delta_k^\wedge) T_k$. These operations are designed to be differentiable to enable end-to-end training.

3.2.3 DEQ training

We adopt a DEQ framework, wherein the above steps are repeated until the depth and pose values reach a fixed point, at which the update value is minimal. In our implementation, the fixed points of depth and pose are chosen separately, and it is possible for both to be selected from different update steps. Finally, the training gradient is computed using the chosen depth and pose fixed points. Operations prior to the fixed point do not require saving gradients in memory, which allows memory-efficient training.

As noted in Eqs. (1) to (3), the self-supervision losses can be computed given a depth and pose prediction. At the fixed point, we compute two additional self-supervision losses. The first performs source-to-target image warping through the pairing of the final depth D^* and the initial pose T_0 , where we detach T_0 from the computational graph to act as a teacher. This loss is used so that the output of the teacher and refinement modules shares the same scaling. The second performs the warping with the pairing of D^* and T^* . This loss helps drive refined estimation of both depth and pose towards the optimal value. Detailed experiments for this choice of loss pairings are provided in the supplementary.

In both losses, we also apply the consistency loss between D_0 and D^* , similar to ManyDepth [91], to account for dynamic objects or occluded regions. Specifically, we extract

coarse depth predictions from the raw feature matchings and mask regions where large disagreements occur, and enforce consistency with the teacher depth. Unlike ManyDepth, however, our method does not explicitly construct a cost volume. To obtain the coarse depth, we search for the lowest matching cost around the neighborhood of the teacher depth, similar to Sec. 3.2.1, but with a larger neighborhood range. This approach offers an additional advantage compared to the cost volume-based method, as we do not need to rely on an estimated minimum and maximum depth or know the scale of the estimates. Moreover, since the computation of coarse depth depends on the accuracy of feature matching costs, it can be improved with more accurate pose estimates (Tab. 2).

4. Experiments

4.1. Dataset and metrics

For depth estimation experiments, we use the Eigen train/test split [15] from the KITTI dataset [20]. To evaluate the estimated depths, we scale them by a scalar to match the scale of the ground truth. We employ standard depth evaluation metrics [14, 15], including absolute and squared relative error (Abs Rel, Sq Rel), root mean square error (RMSE, RMSE log), and accuracy under threshold (δ_1 , δ_2 , and δ_3), with a maximum depth set at 80m. Lower values are better for the first four metrics, while higher values are better for the remaining three.

For the visual odometry experiments, we use the KITTI odometry dataset and follow the same training and evaluation sequences (Seq. 00-08 for training and Seq. 09-10 for evaluation) as in previous work [106]. Since our estimation relies on a monocular camera, the estimated trajectories are aligned with the ground truth using the 7 DoF Umeyama alignment [81]. We use standard odometry evaluation metrics such as translation (t_{err}) and rotation (r_{err}) error [20], and absolute trajectory error (ATE) [76].

4.2. Implementation details

We conduct our experiments using PyTorch [66] on an RTX 3090 GPU with a batch size of 12. Following [23], we apply color and flip augmentations and resize input images to a resolution of 640×192 . For our high resolution experiments, we resize the images to 960×288 . We train the entire network for 15 epochs with a learning rate of 10^{-3} , at which point we freeze the teacher depth and pose models. We then continue to train the network with a learning rate of 10^{-4} . Adam optimizer [45] is employed with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

As mentioned earlier, the depth backbone is based on the HRNet architecture [88]. For the teacher pose model, we adhere to the standard design, using the first five layers of a ResNet18 initialized with ImageNet pre-trained weights as

	Method	Test frames	Semantics	$W \times H$	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
Low & mid res	Ranjan <i>et al.</i> [70]	1		832×256	0.148	1.149	5.464	0.226	0.815	0.935	0.973
	EPC++ [59]	1		832×256	0.141	1.029	5.350	0.216	0.816	0.941	0.976
	Struct2depth (M) [10]	1	•	416×128	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	Videos in the wild [26]	1	•	416×128	0.128	0.959	5.230	0.212	0.845	0.947	0.976
	Guizilini <i>et al.</i> [30]	1	•	640×192	0.102	0.698	4.381	0.178	0.896	0.964	0.984
	Johnston <i>et al.</i> [42]	1		640×192	0.106	0.861	4.699	0.185	0.889	0.962	0.982
	Monodepth2 [23]	1		640×192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	Packnet-SFM [28]	1		640×192	0.111	0.785	4.601	0.189	0.878	0.960	0.982
	Li <i>et al.</i> [51]	1		416×128	0.130	0.950	5.138	0.209	0.843	0.948	0.978
	DIFFNet [105]	1		640×192	0.102	0.764	4.483	0.180	0.896	<u>0.965</u>	<u>0.983</u>
	DualRefine-initial (D_0)	1		640×192	0.103	0.721	4.476	0.180	0.891	<u>0.965</u>	0.984
	Patil <i>et al.</i> [67]	N^\dagger		640×192	0.111	0.821	4.650	0.187	0.883	0.961	0.982
	Wang <i>et al.</i> [89]	2 (-1, 0)		640×192	0.106	0.799	4.662	0.187	0.889	0.961	0.982
ManyDepth (MR) [91]	2 (-1, 0)		640×192	<u>0.098</u>	0.770	4.459	0.176	0.900	<u>0.965</u>	<u>0.983</u>	
DepthFormer [29]	2 (-1, 0)		640×192	0.090	<u>0.661</u>	4.149	<u>0.175</u>	<u>0.905</u>	0.968	0.984	
DualRefine-refined (D^*)	2 (-1, 0)		640×192	0.090	0.658	<u>4.237</u>	0.171	0.912	0.967	0.984	
High res	DRO [27]	2 (-1, 0)		960×320	0.088	0.797	4.464	0.212	0.899	0.959	0.980
	Wang <i>et al.</i> [89]	2 (-1, 0)		1024×320	0.106	0.773	4.491	0.185	0.890	<u>0.962</u>	<u>0.982</u>
	ManyDepth (HR ResNet50) [91]	2 (-1, 0)		1024×320	<u>0.091</u>	0.694	4.245	<u>0.171</u>	<u>0.911</u>	0.968	0.983
	DualRefine-refined (HR) (D^*)	2 (-1, 0)		960×288	0.089	<u>0.754</u>	<u>4.273</u>	0.170	0.917	0.968	0.983

Table 1. Results and comparison with other state-of-the-arts models on the KITTI [20] Eigen split. **Bold**: Best, Underscore: Second best. \dagger : evaluated on whole sequences

Pose Updates	Consistency mask	Abs Rel	Sq Rel	RMSE	δ_1	δ_2
no update	T_0	0.097	0.713	4.462	0.898	0.964
no weights	T_0	0.090	0.658	4.237	0.912	0.967
no $W_{h,k}$	T_0	0.090	0.645	4.273	0.908	0.966
no W_q	T_0	0.091	0.713	4.295	0.910	0.967
W_q and $W_{h,k}$	T_0	0.089	0.652	4.282	0.909	0.967
no weights	T^*	0.091	0.645	4.201	0.910	0.968

Table 2. Ablation experiment for the effect of pose updates on the KITTI [20] Eigen split. **Bold**: Best.

an encoder, followed by a decoder that outputs 6-DoF pose estimates. At test time, depth estimates are made using the current frame and the previous frame when available. When the previous frame is unavailable, we skip the refinement module and simply use the initial estimates.

4.3. Ablation

Pose updates. We analyze the effect of pose refinement toward depth and present the findings in Table 2. Our model that does not perform pose updates has the worst accuracy.

Evolving confidence weighting. We also show the impact of confidence weightings. Interestingly, similar performance can be observed for all models that perform pose updates, even when no confidence weighting is used to guide the pose computation.

DEQ iteration. In addition to studying the effect of pose refinements, we also investigate aspects of DEQ iterative updates and present the findings when we vary the number of iterations during training and at test time in Table 3. The results suggest that 6 iterations is sufficient for the model to find the fixed point. We speculate that the initial estimate provides a reliable starting point and hence the fast convergence. On our machine, the 6 update iterations increase the

DEQ # iters	Abs Rel	Sq Rel	RMSE	δ_1	Time (ms)
3→3	0.097	0.777	4.431	0.906	53
6→3	0.098	0.695	4.318	0.905	53
6→6	0.090	0.658	4.237	0.912	68
12→3	0.102	0.732	4.428	0.900	53
12→6	0.095	0.708	4.336	0.907	68
12→12	0.092	0.707	4.316	0.909	99

Table 3. Ablation experiment for the DEQ iterations on the KITTI [20] Eigen split. $a \rightarrow b$ represents a : # iters at training and b : # iters at test time. **Bold**: Best.

baseline model’s inference time from 37.90 ms by ~ 31 ms to a total of 68.49 ms, running at almost 15 fps. However, we only use PyTorch basic functions in our implementation, and further optimization of the code could be made.

4.4. Depth results

Table 1 shows the comparison of our depth estimation with SoTA self-supervised models. We compare with models that train on monocular video. Our model outperforms most previous models and is competitive with the Transformer [84]-based DepthFormer [29] model. Specifically, our model shows a significant improvement in δ_1 , suggesting highly accurate inliers. Furthermore, compared to DepthFormer that requires 16GB training memory per batch, ours only consumes 24GB of memory for 12 batches, around $1/8 \times$ the memory requirement. This is mainly because our method refines 2D hidden states based on local sampling, while [29] refines 2D feature maps and a 3D feature volume using self/cross-attention along every depth bin.

Fig. 3 displays qualitative outputs for the disparity and error map of our model. We can observe an improvement to the error map of the refined depth. We also additionally dis-

Methods	Seq 9			Seq 10		
	$t_{err}(\%) \downarrow$	$r_{err}(\circ/100m) \downarrow$	ATE (m) \downarrow	$t_{err}(\%) \downarrow$	$r_{err}(\circ/100m) \downarrow$	ATE (m) \downarrow
ORB-SLAM2 [65] (w/o LC)	9.67	0.3	44.10	4.04	0.3	6.43
ORB-SLAM2 [65]	3.22	0.4	8.84	4.25	0.3	8.51
SfMLearner [106]	19.15	6.82	77.79	40.40	17.69	67.34
GeoNet [98]	28.72	9.8	158.45	23.90	9.0	43.04
DeepMatchVO [73]	9.91	3.8	27.08	12.18	5.9	24.44
Monodepth2 [23]	17.17	3.85	76.22	11.68	5.31	20.35
DW [26]-Learned	-	-	20.91	-	-	17.88
DW [26]-Corrected	-	-	19.01	-	-	14.85
SC-Depth [7]	7.31	3.05	23.56	7.79	4.90	12.00
Zou <i>et al.</i> [107]	3.49	1.00	11.30	5.81	1.8	11.80
P-RGBD SLAM [8]	5.08	1.05	13.40	4.32	2.34	7.99
DualRefine-initial (T_0)	9.06	2.59	39.31	9.45	4.05	15.13
DualRefine-refined (T^*)	3.43	<u>1.04</u>	5.18	<u>6.80</u>	1.13	<u>10.85</u>

Table 4. Results on Seq. 09 and Seq. 10 of the KITTI odometry data. We provide a comparison with other state-of-the-art self-supervised depth and odometry methods. ORB-SLAM2 is included as a representative non-learning based method. **Bold**: Best, Underscore: Second best.

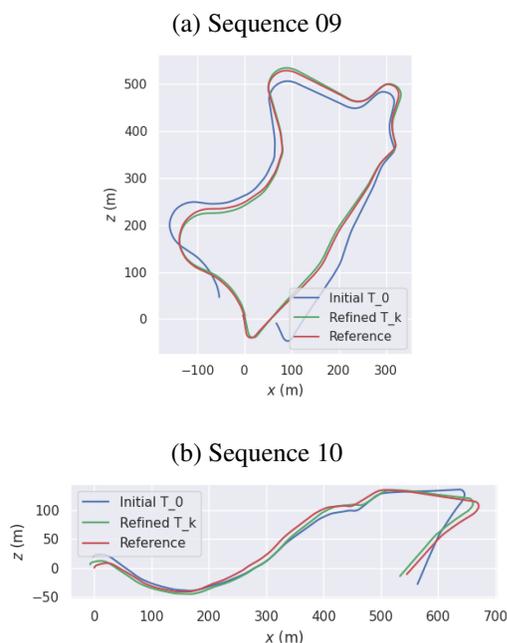


Figure 4. Estimated trajectory by the initial pose estimator and the refined trajectory using our pose refinement module on (a) Seq. 09 and (b) Seq. 10 of KITTI odometry data. The refined pose estimate improves the global trajectory, even without explicitly training for global consistency.

play confidence weight outputs obtained by the model that computes them. We observe that the confidence W_q , which is calculated once, assigns the high confidence sparsely. Interestingly, the confidence weights that evolve with each iteration initially assign high confidence to far-away points and moves towards closer points with increasing iterations.

Limitation: we note an increase in error for the moving vehicle in the lower image set. The worse RMSE of our model compared to DepthFormer also indicates higher outlier predictions. This could be due to repeated operation of

our iterative updates, which further exacerbate the outlier problem. Even with consistency masking, the model we propose displays limitations with dynamic objects. However, the results of the last row of Tab. 2 show that using the refined pose to extract the consistency mask for training slightly improves Sq Rel, RMSE, and δ_2 . We leave further the discussion of this issue for future study.

4.5. Odometry results

We present the results for visual odometry of the teacher model and the refinement model in Table 4. We also present the results of previous models that were trained on monocular videos. Our refinement module drastically improves the initial odometry results, as shown in Fig. 4. Although the goal of our study was to improve the estimation of local poses for accurate matching, we outperformed most of the other models in all metrics. Even without explicit training to ensure scale consistency, as in [7, 8], our refined output demonstrates a globally consistent odometry prediction. Additionally, unlike Zou *et al.* [107] which infers pose from long-term geometry, this result is achieved with only two input frames to infer pose estimates. Our model also achieves an ATE that is on par with the traditional ORB-SLAM2, which performs global geometric optimization, although our results in r_{err} still lag behind.

5. Conclusions

In this paper, we introduced a self-supervised pipeline for multi-frame depth and pose estimation and refinement. By leveraging the combined power of neural network representation and geometric constraints to refine both depth and pose, our approach achieved state-of-the-art performance in both tasks. Our method also demonstrates greater efficiency than competing methods, with potential for further improvement. Nevertheless, we still observed poorer depth accuracy in dynamic scenes.

References

- [1] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2800–2810, 2018. **2**
- [2] Shaojie Bai, Zhengyang Geng, Yash Savani, and J Zico Kolter. Deep equilibrium optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 620–630, 2022. **3**
- [3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019. **2, 3, 4**
- [4] Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3542–3548. IEEE, 2021. **1**
- [5] Antyanta Bangunharcana, Soohyun Kim, and Kyung-Soo Kim. Revisiting the receptive field of conv-gru in droid-slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1906–1916, 2022. **5**
- [6] Sayanti Bardhan. Salient object detection by contextual refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 356–357, 2020. **3**
- [7] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32:35–45, 2019. **1, 8**
- [8] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision*, 129(9):2548–2564, 2021. **3, 8**
- [9] Arantxa Casanova, Guillem Cucurull, Michal Drozdal, Adriana Romero, and Yoshua Bengio. On the iterative refinement of densely connected representation levels for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 978–987, 2018. **3**
- [10] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8001–8008, 2019. **7**
- [11] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. **2**
- [12] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. **2**
- [13] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. **2**
- [14] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. **1, 2, 6**
- [15] David Eigen, Christian Puhersch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. **1, 2, 6**
- [16] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. **3, 5**
- [17] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. **3, 5**
- [18] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. **1**
- [19] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016. **1, 2**
- [20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. **1, 6, 7**
- [21] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European conference on computer vision*, pages 519–534. Springer, 2016. **3**
- [22] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. **1, 2**
- [23] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. **1, 2, 3, 4, 6, 7, 8**
- [24] Jicheng Gong, Zhao Zhao, and Nic Li. Improving multi-stage object detection via iterative proposal refinement. In *BMVC*, page 223, 2019. **3**
- [25] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *Advances in Neural Information Processing Systems*, 33:12626–12637, 2020. **2**
- [26] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised

- monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019. 1, 3, 7, 8
- [27] Xiaodong Gu, Weihao Yuan, Zuoqiuo Dai, Chengzhou Tang, Siyu Zhu, and Ping Tan. Dro: Deep recurrent optimizer for structure-from-motion. *arXiv preprint arXiv:2103.13201*, 2021. 2, 3, 7
- [28] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Rantos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 1, 2, 7
- [29] Vitor Guizilini, Rareş Ambruş, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 160–170, 2022. 2, 3, 7
- [30] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020. 1, 3, 7
- [31] Vitor Guizilini, Jie Li, Rareş Ambruş, and Adrien Gaidon. Geometric unsupervised domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8537–8547, 2021. 3
- [32] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [34] Sungchul Hong, Antyanta Bangunharcana, Jae-Min Park, Minseong Choi, and Hyu-Soung Shin. Visual slam-based robotic mapping method for planetary construction. *Sensors*, 21(22):7715, 2021. 1
- [35] Sungchul Hong, Pranjay Shyam, Antyanta Bangunharcana, and Hyuseoung Shin. Robotic mapping approach under illumination-variant environments at planetary construction sites. *Remote Sensing*, 14(4):1027, 2022. 1
- [36] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018. 3
- [37] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 3
- [38] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 3
- [39] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015. 4
- [40] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4713–4722, 2022. 3
- [41] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021. 5
- [42] Adrian Johnston and Gustavo Carneiro. 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*, pages 4756–4765, 2020. 2, 7
- [43] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 1, 2, 3, 4
- [44] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 1, 2
- [45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [46] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020. 2
- [47] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2656–2665, 2018. 2
- [48] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 1, 2
- [49] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2
- [50] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. 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*, pages 1119–1127, 2015. 1
- [51] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes. *arXiv preprint arXiv:2010.16404*, 2020. 7
- [52] RuiBo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust depth prediction. In *Asian Conference on Computer Vision*, pages 663–678. Springer, 2018. 2

- [53] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2
- [54] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2811–2820, 2018. 2
- [55] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 3
- [56] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. 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*, pages 10986–10995, 2019. 3
- [57] Jiawei Liu, Jing Zhang, and Nick Barnes. Confidence-aware learning for camouflaged object detection. *arXiv preprint arXiv:2106.11641*, 2021. 3
- [58] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. In *European Conference on Computer Vision*, pages 640–657. Springer, 2020. 3
- [59] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2624–2641, 2019. 7
- [60] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020. 1
- [61] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018. 1
- [62] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126(9):942–960, 2018. 2
- [63] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 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*, pages 4040–4048, 2016. 2
- [64] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 611–619. IEEE, 2016. 2
- [65] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 3, 8
- [66] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [67] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don’t forget the past: Recurrent depth estimation from monocular video. *IEEE Robotics and Automation Letters*, 5(4):6813–6820, 2020. 7
- [68] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European conference on computer vision*, pages 75–91. Springer, 2016. 3
- [69] Koutilya PNVR, Hao Zhou, and David Jacobs. Sharingan: Combining synthetic and real data for unsupervised geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13974–13983, 2020. 3
- [70] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019. 1, 7
- [71] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv:2104.07636*, 2021. 3
- [72] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3247–3257, 2021. 2, 3, 6
- [73] Tianwei Shen, Zixin Luo, Lei Zhou, Hanyu Deng, Runze Zhang, Tian Fang, and Long Quan. Beyond photometric loss for self-supervised ego-motion estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6359–6365. IEEE, 2019. 8
- [74] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 1, 2
- [75] Pranjay Shyam, Antyanta Bangunharcana, and Kyung-Soo Kim. Retaining image feature matching performance under low light conditions. In *2020 20th International Conference on Control, Automation and Systems (ICCAS)*, pages 1079–1085. IEEE, 2020. 1
- [76] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 6
- [77] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018. 1, 3

- [78] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3, 4, 5
- [79] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems*, 34:16558–16569, 2021. 3
- [80] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4654–4665, 2020. 2
- [81] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 6
- [82] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 2, 3
- [83] Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Wolfram Burgard, Greg Shakhnarovich, and Adrien Gaidon. Neural ray surfaces for self-supervised learning of depth and ego-motion. In *2020 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2020. 1
- [84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3, 7
- [85] Lukas Von Stumberg, Patrick Wenzel, Qadeer Khan, and Daniel Cremers. Gn-net: The gauss-newton loss for multi-weather relocalization. *IEEE Robotics and Automation Letters*, 5(2):890–897, 2020. 3
- [86] Lukas Von Stumberg, Patrick Wenzel, Nan Yang, and Daniel Cremers. Lm-reloc: Levenberg-marquardt based direct visual relocalization. In *2020 International Conference on 3D Vision (3DV)*, pages 968–977. IEEE, 2020. 3
- [87] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018. 1, 5
- [88] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 4, 6
- [89] Jianrong Wang, Ge Zhang, Zhenyu Wu, XueWei Li, and Li Liu. Self-supervised joint learning framework of depth estimation via implicit cues. *arXiv preprint arXiv:2006.09876*, 2020. 7
- [90] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International conference on 3d vision (3DV)*, pages 248–257. IEEE, 2018. 3
- [91] Jamie Watson, Oisín Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021. 1, 3, 4, 6, 7
- [92] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *European conference on computer vision*, pages 230–247. Springer, 2020. 3
- [93] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 3
- [94] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6101–6108. IEEE, 2019. 2
- [95] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 2
- [96] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. 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*, pages 1281–1292, 2020. 3
- [97] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 3
- [98] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 1, 8
- [99] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K Fishman, and Alan L Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8280–8289, 2018. 3
- [100] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2
- [101] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 340–349, 2018. 2
- [102] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019. 3
- [103] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 2
- [104] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019. 3
- [105] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. *arXiv preprint arXiv:2110.09482*, 2021. 3, 4, 7
- [106] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 1, 2, 6, 8
- [107] Yuliang Zou, Pan Ji, Quoc-Huy Tran, Jia-Bin Huang, and Manmohan Chandraker. Learning monocular visual odometry via self-supervised long-term modeling. In *European Conference on Computer Vision*, pages 710–727. Springer, 2020. 8
- [108] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 36–53, 2018. 1