Do What You Can, With What You Have: Scale-aware and High Quality Monocular Depth Estimation Without Real World Labels

Kunal Swami, Amrit Muduli, Uttam Gurram, Pankaj Bajpai
Camera Solutions Group,
Samsung Research India Bangalore
{kunal.swami,amrit.muduli,uttam.g,pankaj.b}@samsung.com

Abstract

Learning robust and scale-aware monocular depth estimation (MDE) requires expensive data annotation efforts. Self-supervised approaches use unlabelled videos but, due to ambiguous photometric reprojection loss and no labelled supervision, produce inferior quality relative (scale ambiguous) depth maps with over-smoothed object boundaries. Approaches using synthetic training data suffer from the non-trivial domain adaptation problem; despite complicated unsupervised domain adaptation (UDA) techniques, these methods still do not generalize well to real datasets.

This work presents a novel and effective training methodology to combine self-supervision from unlabelled monocular videos and dense supervision from the synthetic dataset synergistically without complicated UDA techniques. With our method, geometry and semantics are learned from monocular videos, whereas scale-awareness and qualitative attributes, e.g., sharp and smooth depth variations, that are crucial for practical use cases are learned from the synthetic dataset. Our method outperforms self-supervised, semi-supervised, and all the domain adaptation methods on standard benchmark datasets while being competitive with fully supervised methods.

Furthermore, our method leads to qualitatively superior depth maps, which increases its practical utility compared to existing methods. We demonstrate this by applying our method to develop an MDE model for a real life application—DSLR-like shallow depth-of-field effect on smartphones. The new high quality synthetic depth dataset that we generate for this task will be available to the community.

1. Introduction

Depth information is used in 3D reconstruction, augmented reality, autonomous driving [11–13] and computational photography, such as shallow depth-of-field effect [43]. Compared to time-of-flight (ToF) and stereo cameras, monocular depth estimation (MDE) can enable widespread vision applications in a low-cost effective manner. MDE also works with images and videos stored in a media.

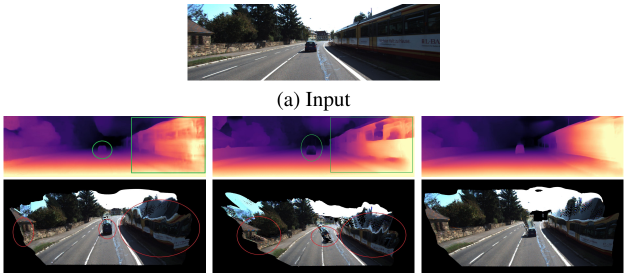

(a) Input
(a) Monodepth2 [14] (b) PackNet-SfM [16] (c) Ours

Figure 1. Comparison of our qualitatively superior and scale-aware depth estimation. The third row shows that the reconstructed 3D point cloud using our accurate scale-aware depth preserves shapes of objects much better than state-of-the-art methods.

With advancements in deep learning, MDE has witnessed significant progress [2, 8–10, 26, 28, 29, 31, 32, 49]. However, majority of earlier works focused on building supervised MDE models which demand immense data collection and filtering efforts [12,40]. Also, based on the capture setup, for e.g., LiDAR [12,38], it may not be possible to acquire ground-truth in all desired scenarios due to sensor or setup limitations. *Training data collection is an activity that is periodically required to maintain and adapt a learning based model to evolving test scenarios, building real world RGB-D supervised MDE models needs labor intensive data creation as well as maintenance costs.*

To solve this problem, researchers proposed self-supervised approaches [11, 13, 14, 16, 54, 57] which use unlabelled monocular videos [14, 16, 54, 57] to jointly learn depth and ego-motion. The idea is that a target video frame can be recreated from temporally adjacent frames using the estimated depth and ego-motion, and the image (or photometric) reconstruction error between original and recreated images can be used as supervision. However, this photometric consistency can be easily violated in real world scenarios because of occlusions, illumination changes, textureless areas, dynamic and non-rigid objects, which hinder the optimization process [14, 46]. *As a result, self-supervised approaches produce inferior quality depth maps with blur and over-smoothed object boundaries and gaps* [14, 16, 46]. *Additionally, due to lack of ground-truth supervision these methods generate relative, i.e., scale ambiguous depth.*

Another popular approach is to simulate large-scale diverse synthetic RGB-D pairs with dense pixel-perfect ground-truth using computer graphics software [3, 19]. However, due to the domain gap between synthetic and real world images, which is majorly attributed to differences in appearance, physics, and rendering styles [19], models trained on synthetic datasets do not generalize well to real world images. Recently, it has been found [55] that multiple networks (e.g., generative models for style transfer) and optimization losses (e.g., adversarial losses) used in unsupervised domain adaptation (UDA) methods compete against each other and do not contribute to the optimization of the depth estimation task. These methods [6, 35, 53] also use additional modules, for e.g., real-to-synthetic style transfer, during inference time. *Therefore, not only are these approaches hard to train, they perform worse than methods trained on real datasets [55] and also incur additional inference load.*

Motivation and Contributions. Self-supervised approaches require the cheapest and widely available source of training data, viz., monocular videos. However, their low quality of output depth makes these approaches practically less attractive, especially when depth information is required for image enhancement effects [43]. Whereas synthetic datasets provide dense ground-truth for supervision, but the performance of domain adaptation methods hardly matches that of fully supervised methods, mainly because their UDA techniques do not optimize the depth estimation task [55]. Motivated by these observations, we wonder if we can utilize both monocular videos and synthetic datasets together to train an MDE model that *generalizes to real world data, is scale-aware, and also estimates depth maps with the following qualitative attributes—sharp edges and smooth depth variations*. At first, it might seem to be a trivial task of training an MDE model jointly using self-supervised learning on monocular videos and pixel-wise depth regression task on the synthetic dataset. However, our experiments find that the task is not simple and demands a carefully designed training strategy.

This work presents a novel and effective training methodology to combine self-supervision from unlabelled monocular videos and dense supervision from the synthetic dataset synergistically without complicated domain adaptation techniques. In our method, real world geometry and semantics are learned from monocular videos through self-supervision, whereas scale-awareness and qualitative depth attributes, like sharp and smooth depth variations, that are crucial for practical applications [43] are acquired from synthetic dataset training. To achieve the best depth estimation accuracy in our method, we further propose to disentangle the task of relative depth estimation with qualitative depth attributes from the scale-aware depth estimation task.

Following are the **major contributions** of this work:

1. Novel training methodology to learn geometry and semantics from monocular videos and scale-awareness and qualitative depth attributes, like sharp and smooth depth variations from the synthetic dataset.
2. In our approach, we disentangle scale, and qualitative depth attributes from the synthetic dataset to achieve state-of-the-art results without complicated domain adaptation techniques.
3. Our method sets a new state-of-the-art among the only two (to the best of our knowledge) scale-aware self-supervised, self-supervised, and all domain adaptation based MDE methods in the literature.
4. We showcase the practical utility of our method by applying it to develop an MDE model for *Portrait Mode* effect on smartphones. The new human-centric synthetic depth dataset that we generate for this task will be available to the community, which will be helpful for ongoing human-centric vision research [41].

2. Related Work

Self-Supervised MDE. Garg *et al.* [11] pioneered this approach using calibrated stereo image pairs for training, Zhou *et al.* [57] generalized this technique for joint depth and ego-motion estimation with monocular videos. GeoNet [52] extended this by additionally learning optical flow to explcitly handle motion by dynamic objects. Several methods [14, 15, 45, 46, 54] have improved upon this approach by incoporating additional geometric [23] and loss [14] constraints to compensate for violations of photometric consistency. However, the structure-from-motion (SfM) [39] constraints allow these methods to learn depth and ego-motion only upto an unknown scale, thus, these method scale their estimates using ground-truth for evaluation [57] on KITTI Eigen test split [12, 57]. Recently, Guizilini *et al.* [16] used camera velocity and Chawla *et al.* [4] used GPS information to supervise ego-motion, leading to scale-aware depth estimation. However, velocity or GPS information is not always available during training [16], it also requires additional hardware that is often prone to operational noise. In this work, we extend self-supervised learning techniques to estimate more accurate scale-aware depth using supervision from synthetic datasets, which can be generated efficiently. When compared to these existing methods, our scale-aware depth is also qualitatively superior.

Synthetic-Real Domain Adaptation. Existing domain adaptation methods [1, 7, 24, 35, 53, 56] assume that the synthetic-real domain bias problem can be solved by learning domain invariant features [24, 35] or using synthetic-to-real style transfer [1, 53, 56]. Specifically, these methods perform unsupervised domain adaptation using additional generative models, such as CycleGAN [59] and adversarial discriminator networks to judge whether the learned feature space or style transferred images are indistinguishable across domains. A recent method [7, 55] shifts away from this philosophy, it assumes availability of small quantity of real world RGB-D data, whereas [7] simply fine-tunes a synthetic dataset pre-trained model on real world data. Another recent method [17] adopts a multi-task multi-domain geometric unsupervised domain adaptation primarily for semantic segmentation. It also improved the depth estimation task; however, it requires a multi-task training setup and multi-task labels. In contrast, we propose a novel joint training methodology on monocular videos and synthetic datasets to learn high quality scale-aware depth estimation without any complicated domain adaptation technique.

3. Proposed Methodology

A diagram of the proposed training methodology for MDE is shown in Fig. 2. The proposed method involves training following networks: a MDE network $\Phi : I \rightarrow d$, that takes an input image I and outputs a depth map d; pose estimation network $\Omega : \{I_a \odot I_b\} \rightarrow \hat{\mathcal{T}}_{a \rightarrow b}$, that takes a pair of input images ($\odot$ denotes channel-wise concatenation) and estimates a 6DOF relative pose $\hat{\mathcal{T}}$ between them; ScaleNet $\delta : f_\Phi \rightarrow s$, that takes coarsest high level feature maps f_Φ of our depth encoder and outputs a global scene scale estimate s.

The proposed training framework involves three training stages (see Fig. 2). In Stage 1, we pre-train (follow red path) the MDE model Φ on monocular videos dataset using self-supervision. This is to ensure that Φ learns feature representations that are specific to real world images, which also constitute our target test dataset. Subsequently, in Stage 2, we jointly train (follow red and blue path) Φ on monocular videos and synthetic datasets, the joint training is important to prevent synthetic dataset domain bias. In each iteration in Stage 2, we sample separate mini-batches from both datasets in the ratio 2:1 (real:synthetic) and pass these batches sequentially (which constitutes our one forward pass) to generate corresponding outputs and losses. In Stage 3, we train (follow green path) ScaleNet δ, depth and pose networks (Φ and Ω) are freezed in this stage. ScaleNet δ is trained only on the synthetic dataset, an input image I_{syn} is passed through Φ and the coarsest level feature maps f_Φ are input to δ which estimates a global scene scale. We freeze Φ in this stage because training only on synthetic data for scene scale estimation task will make its feature representations specific to the synthetic domain.

The self-supervised and synthetic dataset losses are computed on output batches corresponding to respective datasets. The overall training loss function L for each iteration is computed as follows:

$$L = \alpha * L_{ss} + \beta * L_{syn} \tag{1}$$

Here, the parameters α and β are set in such a way that the losses L_{ss} and L_{syn} are comparable in magnitude. In Stage 1, β is set to 0, whereas in Stage 3, α is set to 0. The synthetic dataset specific losses that we compute in Stage 2 and Stage 3, which constitute L_{syn} are described in subsequent subsections.

3.1. Self-supervision from Monocular Videos

The self-supervised approach constrains the MDE model Φ to reconstruct the view of a target (input) image from the view of a source (temporally adjacent) image using the estimated depth and ego-motion. The view (or photometric) reconstruction error is used as supervision to train Φ. We assume that the camera intrinsics ($\mathcal{K}$) are known for all the videos in the training dataset, $\mathcal{K}$ can also be computed (with decent approximation) using the open-source SfM pipeline COLMAP [39]. During training, an input image I_t is sampled along with its two temporally adjacent frames $I_{t\pm1}$ (see Fig. 2). I_t is referred to as target, which is fed to the MDE model Φ to estimate depth d_t, whereas source images $I_{t\pm1}$ are used to perform view reconstruction based supervision.

3.1.1 Self-supervised Loss

The pose network Ω takes $\{I_t \odot I_{t-1}\}$ and $\{I_t \odot I_{t+1}\}$ as input and estimates 6DOF relative poses $\hat{\mathcal{T}}_{t\rightarrow t-1}$ and $\hat{\mathcal{T}}_{t\rightarrow t+1}$ respectively. With the camera intrinsics ($\mathcal{K}$), relative poses ($\hat{\mathcal{T}}_{t\rightarrow t\pm1}$) and estimated depth information (d_t), a homogeneous pixel $p_t \in I_t$ can be mapped to a homogeneous pixel $p_{t\pm1} \in I_{t\pm1}$ as follows:

$$\hat{p}_{t\pm1} \sim \mathcal{K}\hat{\mathcal{T}}_{t\rightarrow t\pm1} d_t(p_t) \mathcal{K}^{-1} p_t \tag{2}$$

Since $\hat{p}_{t\pm1}$ are in continuous coordinates, differentiable bilinear image warping [21] is applied to compute $I_{t\pm1}(\hat{p}_{t\pm1})$ which are the reconstructed target images deonted by $\hat{I}_{t-1\rightarrow t}(p_t)$ and $\hat{I}_{t+1\rightarrow t}(p_t)$.

A combination of L_1-norm error and SSIM score [14] is used as photometric reconstruction error (pe) for training the MDE model Φ as follows:

$$\begin{aligned} pe = \alpha & \frac{1 - SSIM(I_t - \hat{I}_{t\pm1\rightarrow t})}{2} \\ & + (1-\alpha)|I_t - \hat{I}_{t\pm1\rightarrow t}| \end{aligned} \tag{3}$$

Here, α is set to 0.85. Following [14], auto-masking is used to minimize the effects of dynamic, non-rigid objects and occlusions during training. An edge-aware smoothness term ($\mathcal{S}$) is also added in the loss function to encourage depth predictions that respect object boundaries.

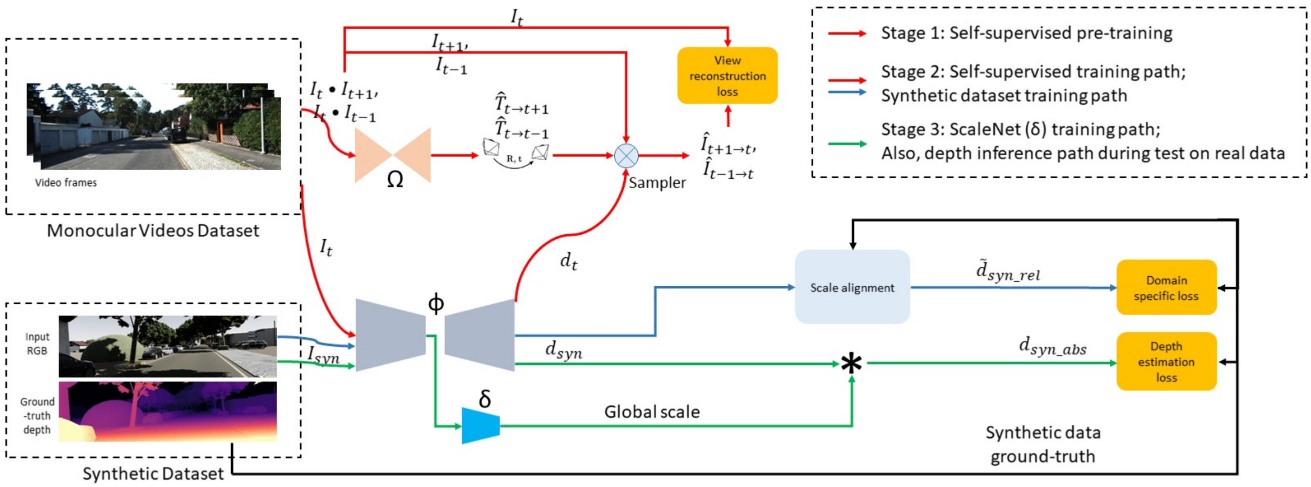


Figure 2. Diagram explaining the proposed method to effectively combine real world self-supervision from monocular videos and dense supervision from synthetic dataset to achieve scale-aware depth estimation with qualitative depth attributes.

$$\mathcal{S} = |\partial_x \hat{d}_t| e^{-|\partial_x I_t|} + |\partial_y \hat{d}_t| e^{-|\partial_y I_t|} \tag{4}$$

Here, $\hat{d}_t$ is the mean-normalized inverse depth to avoid shrinking of estimated depth [14]. The overall self-supervised loss L_{ss} is defined as follows:

$$L_{ss} = pe + \lambda \mathcal{S} \tag{5}$$

3.1.2 Relative Depth and Ego-motion

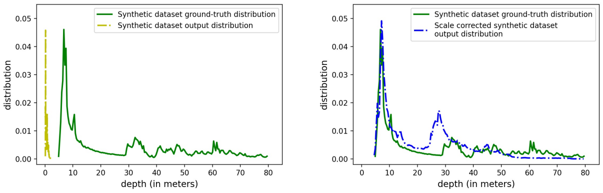
The photometric reconstruction loss in Eq. 3 is scale ambiguous to the joint depth and ego-motion prediction [44]. During training, with any two depth maps which differ only in scale and with corresponding poses also updated (scaled), the pixel mapping in Eq. 2 will be the same. Hence, the photometric reconstruction error will also be the same. Essentially, the depth and pose networks learn to co-adapt their scales during training to minimize the loss in Eq. 3 Therefore, the self-supervised training loss is scale-invariant in this sense. Hence, self-supervised MDE methods rely on the ground-truth LiDAR depth to artificially scale their depth estimates for evaluation [57].

3.2. Dense Supervision from Synthetic Dataset

A trivial approach to train the MDE model Φ jointly with self-supervision is to perform pixel-wise depth regression using a L_1-norm (or L_2-norm) loss as follows:

$$L_{syn} = \|d_{syn\_abs} - \hat{d}_{syn}\| \tag{6}$$

Here, d_{syn_abs} denotes the estimated scale-aware (absolute) depth for the input image I_{syn}. However, since the MDE model, Φ pre-trained using self-supervision outputs relative depth, jointly training the model using Eq. 6 as loss function incentivizes the model to estimate absolute depth in the synthetic domain. This deranges the jointly converged state of depth and pose networks; with the modified scale of output depth, the photometric loss increases because the pose network is not adapted to modified scales of output depth. The magnitude of the scale change required for our model is graphically shown in Fig. 3(a). This leads to a training scenario where two losses in Eq. 3 and 6 compete against each other, and the MDE model converges to a suboptimal state. We show in our experiments that such trivial approaches and their modifications do not achieve optimal results, and the model shows inferior performance.



(a) Default synthetic dataset output and ground-truth distribution (b) Synthetic dataset output distribution after scale alignment correction

Figure 3. Scale alignment for synthetic dataset training

In our training approach, we propose to disentangle *scale-awareness* and *qualitative depth attributes* from synthetic dataset training. We propose that qualitative depth attributes can be learned without deranging the converged state of depth and pose networks by applying a scale transformation on output depth to match its distribution with ground-truth, then computing an appropriate loss function that captures these attributes. This approach does not penalize relative depth estimates of the depth network trained using self-supervision. We show in our experiments that this approach improves the quality of the estimated relative depth and makes it more accurate than the baseline MDE model trained using self-supervision. Next, we propose learning a global scene scale estimation network ScaleNet δ, trained on the synthetic dataset to transform our accurate relative depth predictions into absolute depth predictions.

3.2.1 Accurate Relative Depth Estimation

We use a scale alignment module (see Fig. 2) in Stage 2 joint training to transform an output depth from a synthetic image to match its ground-truth distribution. Our scale alignment module performs a median scaling of the output depth, which is defined as follows:

$$\tilde{d}_{syn\_rel} = \frac{median(\hat{d}_{syn})}{median(d_{syn})} * d_{syn} \tag{7}$$

Since we want the self-supervised pre-trained MDE model to learn qualitative depth attributes, viz. sharp and smooth depth variations from synthetic dataset ground-truth, we compute a loss function in the gradient domain that captures such features in an image. It is shown ablation study in Section 5.3 that this domain specific loss on synthetic dataset leads to slightly better results compared to the standard loss function in Eq. 6. The loss function in the gradient domain denoted by L_{syn_grad} is computed using scale aligned output depth $\tilde{d}_{syn\_rel}$:

$$L_{syn\_grad} = |\nabla_x \tilde{d}_{syn\_rel} - \nabla_x \hat{d}_{syn}| + |\nabla_y \tilde{d}_{syn\_rel} - \nabla_y \hat{d}_{syn}| \tag{8}$$

3.2.2 ScaleNet: Global Scene Scale Estimation

In Stage 2 joint training, the MDE model learns to output accurate relative depth d_{syn} with desired qualitative depth attributes. However, now it is not trivial to train the MDE model for scale-aware depth. For this task, we design a light-weight global scene scale estimation network ScaleNet δ (see Fig. 2) which takes the coarsest level feature maps from depth encoder as input and estimates a single scale factor s which is multipled with relative depth d_{syn}. We then apply a standard pixel-wise depth regression loss [9] using ground-truth $\hat{d}_{syn}$ to train ScaleNet.

$$d_{syn\_abs} = s * d_{syn} \tag{9}$$

Note that in Stage 3, we freeze the MDE model Φ to prevent synthetic dataset domain bias, as ScaleNet is trained using only synthetic dataset supervision. Our experiments show that ScaleNet generates accurate scale-aware depth even on unseen real world testing dataset (see supplementary).

4. Experimental Setup

4.1. Datasets

KITTI. We use the standard KITTI dataset [12] as our unlabelled monocular videos dataset for training and testing. More specifically, we use KITTI Eigen split [9] containing $39,810$ training, 4424 validation and 697 test frames with LiDAR ground-truth used only for evaluation. Following the standard practice in literature [14, 34, 35, 52–54, 57], we cap the depth range to 80m during evaluation.

Virtual KITTI 2. We use VKITTI2 [3] as our synthetic dataset for training. It is the virtual world simulation of scenarios in the KITTI dataset. It contains $21,260$ RGB images with dense and pixel-perfect ground-truth. We use $20,000$ RGB-D pairs for training, while the remaining $1,260$ samples form our validation dataset. Like existing domain adaptation methods [1, 7, 35, 53], our method also requires that the synthetic dataset with desired depth distribution is created by acquiring limited amount of real world sensor data (seed data) for the purpose of calibration [3].

4.2. Implementation Details

This work uses PyTorch framework for implementation. For all our experiments, we use Adam optimization [22] with momentum term 0.9. We set an initial learning rate of 2x10^{-4} for all pre-trainings and 2x10^{-5} for all joint trainings and fine-tuning. A polynomial learning rate decay policy was applied with power term set to 0.9. Data augmentation in the form of random crop, brightness, gamma, and color shift was performed randomly on the fly.

Network Architecture. Since the main focus of this work is to develop a training methodology, we adopt the light-weight network architecture from Monodepth2 [14] as our MDE and pose estimation model. More specifically, we use ResNet18 [18] as encoder and DispNet [33] as decoder for our MDE model (Φ).

5. Results and Discussion

We compare the results of our model with MDE methods that estimate relative depth as well as methods that estimate scale-aware depth. For fair comparison against relative depth estimation methods, we ignore the scale predicted by our ScaleNet network and scale the estimated relative depth with median ground-truth LiDAR information during evaluation [57].

5.1. Relative Depth Estimation

Table 1 shows the detailed quantitative comparison of our method with the latest state-of-the-art self-supervised, stereo self-supervised, semi-supervised, fully supervised, and domain adaptation methods on the KITTI Eigen test split. Our model outperforms other methods for relative depth estimation in almost all the metrics. The performance of our model is best among all self-supervised, monocular, and stereo self-supervised as well as semi-supervised methods trained using pseudo depth ground-truth [46]. Our model has 3% better Abs Rel error, 19% better Sq Rel error, 7% better RMSE error, 8% better RMSE log error

Method	**Year**	**Data**	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [14]	ICCV'19	M	0.115	0.882	4.701	0.190	0.879	0.961	0.982
PackNet-SfM [16]	CVPR'20	M	0.107	0.802	4.538	0.186	0.889	0.962	0.981
SCSI [45]	ICCV'21	M	0.109	0.779	4.641	0.186	0.883	0.962	0.982
WaveletMonodepth [36]	CVPR'21	S	0.105	0.797	4.732	0.203	0.869	0.952	0.977
Monodepth2 [14]	ICCV'19	M+S	0.106	0.806	4.630	0.193	0.876	0.958	0.980
D3VO [50]	CVPR'20	M+S	0.099	0.763	4.485	0.185	0.885	0.958	0.979
R-MSFM6 [58]	ICCV'21	M+S	0.108	0.753	4.469	0.185	0.888	0.963	0.982
DVSO [51]	ECCV'18	M+D*	**0.097**	0.734	4.442	0.187	0.888	0.958	0.980
Depth Hints [46]	ICCV'19	MS+D*	0.098	0.702	4.398	0.183	0.887	0.963	0.983
pRGBD-Refined [42]	ECCV'20	M+D*	0.113	0.793	4.655	0.188	0.874	0.960	0.983
Ours (relative)	-	M+V	0.103	**0.654**	**4.300**	**0.178**	**0.891**	**0.966**	**0.984**
Eigen [9]	NeurIPS'14	D	0.203	1.548	6.307	0.282	0.702	0.890	0.890
Liu [32]	CVPR'15	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Kuznietsov [25]	CVPR'17	S+D	0.113	0.741	4.621	0.189	0.862	0.960	**0.986**
AdaDepth [24]	CVPR'18	M+D+V (DA)	0.167	1.257	5.578	0.237	0.771	0.922	0.971
T^2Net [56]	ECCV'18	M+V (DA)	0.174	1.410	6.046	0.253	0.754	0.916	0.966
GASDA [53]	CVPR'19	S+V (DA)	0.120	1.022	5.162	0.215	0.848	0.944	0.974
S^3Net [7]	ECCV'20	M+V (DA)	0.124	0.826	4.981	0.200	0.846	0.955	0.982
ARC [55]	CVPR'20	D+V (DA)	0.143	0.927	4.679	0.246	0.798	0.922	0.968
SharinGAN [35]	CVPR'20	S+V (DA)	0.116	0.939	5.068	0.203	0.850	0.948	0.978
S2R-DepthNet [6]	CVPR'21	V (DA)	0.165	1.351	5.695	0.236	0.781	0.931	0.972
GUDA [17]	ICCV'21	M+V (DA)	0.114	0.875	4.808	-	0.871	-	-
PackNet-SfM [16]	CVPR'20	M+Vel	**0.107**	0.803	4.566	0.197	**0.876**	0.957	0.979
Chawla [4]	ICRA'21	M+GPS	0.109	0.844	4.774	0.194	0.869	0.958	0.981
Ours (absolute)	-	M+V	0.109	**0.702**	**4.409**	**0.185**	**0.876**	**0.962**	0.984

Table 1. **Quantitative comparison on the KITTI Eigen test split [9]. M**: methods trained on monocular videos using self-supervision, **S**: methods trained on stereo images using self-supervision, **D***: auxiliary depth supervision, **D**: KITTI ground-truth supervision, **V**: methods trained on synthetic dataset, **DA**: methods which use domain adaptation techniques, **Vel**: velocity [16] and **GPS**: GPS [4] based supervision for scale-aware depth estimation. PackNet-SfM [16] and Chawla [4] are the only two scale-aware self-supervised MDE methods. Above red line: relative depth comparison, below absolute depth comparison. **Bold** denotes best and underline denotes second best performance.

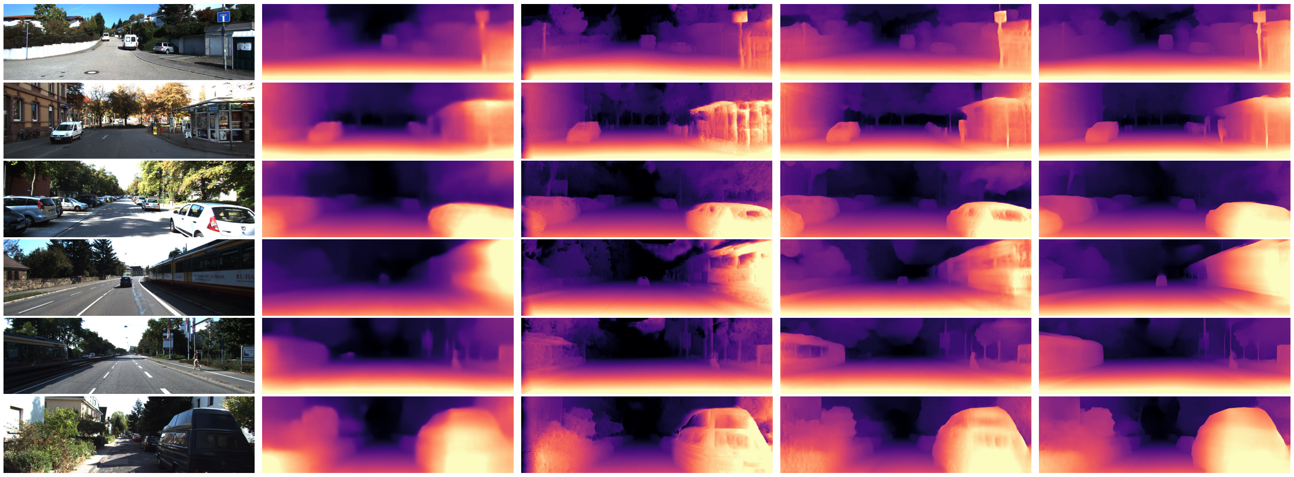


Figure 4. **KITTI results.** Qualitative comparison with state-of-the-art. The proposed method leads to edge-consistent depth estimation, smooth depth variations, no bleeding object edges and no holes within objects (particularly on reflective surfaces).

compared to Monodepth2 [14] which is the baseline model for our method because we use their architecture for our training.

5.2. Scale-aware (or Absolute) Depth Estimation

In comparison to methods that estimate absolute depth, our model again performs significantly better than all domain adaptation methods, and the velocity supervision based PackNet-SFM [16]. To ensure a fair comparison of our method against GUDA [17], we have included results of the model trained without the partially supervised photometric loss. Also, it must be noted that PackNet-SfM uses a more complex network with 128 million parameters compared to our ResNet18 based network with 14 million parameters. Even so, our model outperforms PackNet-SfM on almost all metrics with 12% better Sq Rel error and 6% better RMSE log error. Therefore, with improved architecture, the performance of our method can increase significantly. To the best of our knowledge, there are only two scale-aware self-supervised methods [16], [4] and our model outperforms both by a significant margin. Our model also significantly outperforms the semi-supervised method by Kuznietsov *et al.* [25] which is trained using ground-truth depth information.

The quantitative results in Fig. 4 show that the proposed method generates visually more accurate, sharp, and smooth depth maps compared to other methods, which have blur boundaries, holes in reflective surfaces, and missing thin structures (for e.g., the distant pole in first row image).

We also performed an additional evaluation (without re-training) on the Make3D [38] dataset to check the generalization capability of our method, which is included in the supplementary.

5.3. Ablation Studies

All the experiments were evaluated on the KITTI Eigen test split. Table 2 shows the result of our ablation study.

For relative depth estimation: **Baseline SS** denotes the model trained using self-supervision. **Joint (PT SS)** denotes the model pre-trained using self-supervision and then jointly trained using standard pixel-wise regression loss (Eq. 6) on the synthetic dataset. It can be seen that this model performs inferior compared to baseline, mainly because of two competing loss functions as explained in Section 3.2. It must be noted that this inferior performance is on relative depth, **Ours** L_1 denotes the model trained using the proposed method till Stage 2, using a standard L_1-norm loss instead of gradient domain loss in Eq. 8. **Ours** L_1 performs significantly better than baseline, however, the model trained with gradient domain loss in Eq. 8, i.e., **Ours Grad** achieves best performance.

For absolute depth estimation: **Baseline Syn** denotes the model trained only on the synthetic dataset; it can be seen that it does not generalize well on real world images. **Joint (Naïve)** denotes the model jointly trained using self-supervision and pixel-wise regression loss on the synthetic dataset from scratch, whereas **Joint (PT Syn)** and **Joint (PT SS)** perform pre-training on synthetic and monocular videos respectively. **Ours** denotes **Ours Grad** model trained in Stage 3, which is the only model that learns to estimate scale-aware depth. Due to space limitations, a comprehensive ablation study with detailed discussion and ScaleNet analysis is included in the supplementary.

Experiment	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	$\delta_{1.25}$
Baseline SS	0.115	0.882	4.701	0.190	0.879
Joint (PT SS)	0.430	8.967	12.203	0.462	0.481
Ours (L_1)	0.106	0.713	4.369	0.181	0.888
Ours (Grad)	**0.103**	**0.654**	**4.300**	**0.178**	**0.891**
Baseline Syn	0.200	1.588	6.853	0.323	0.663
Joint (Naïve)	0.548	5.147	11.391	0.833	0.007
Joint (PT Syn)	0.588	5.820	12.041	0.930	0.005
Joint (PT SS)	0.941	14.147	18.408	3.094	0.001
Ours	**0.109**	**0.702**	**4.409**	**0.185**	**0.876**

Table 2. **Ablation study results.** Quantitative results of our ablation study to demonstrate the efficacy of the proposed method.

6. Practical Application of Our Method

We additionally demonstrate the practical usefulness of our method by developing an MDE model for the task of applying DSLR-like synthetic depth-of-field effect, popular as Portrait Mode on smartphones. The following paper [43] serves as an excellent reference to understand the use-case in general. The use-case requires accurate relative depth that preserves sharp boundaries and gaps and has smooth depth variations for a pleasing effect.

In literature, supervised MDE methods that attempt robust depth estimation, [5, 27, 30, 47, 48] advocate creating large amounts of RGB-D data by crowd-sourcing web stereo images or stereo pairs from 3D movies to extract disparity [27]. These approaches require immense data collection and ground-truth depth filtering efforts. Besides, these datasets have biases and limitations, such as the limited quality of web images. Most importantly, copyright or license restrictions prohibit commercial usage of models developed using such data.

However, with our method, we can collect our own monocular videos dataset and generate a synthetic depth dataset (see Fig. 6 and Sec. 6.1) using computer graphics software [19] with relatively less efforts. We then train a lightweight MDE model based on MobileNetV2 [37] (due to our computational requirements, a detailed discussion is omitted due to space limitations) using our training procedure. Fig. 5 shows the results of our MDE model on some test images. Our model generates accurate depth maps with sharp edges and gaps, for e.g., see gaps between leaves in Fig. 5(b), Fig. 5(h) and gaps between fingers in Fig. 5(d).

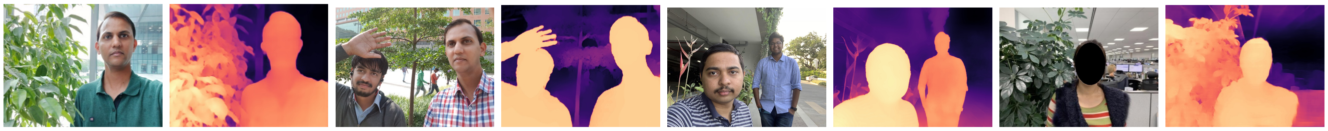

(a) Input 1 (b) Output 1 (c) Input 2 (d) Output 2 (e) Input 3 (f) Output 3 (g) Input 4 (h) Output 4

Figure 5. Qualitative results of our custom MDE solution.

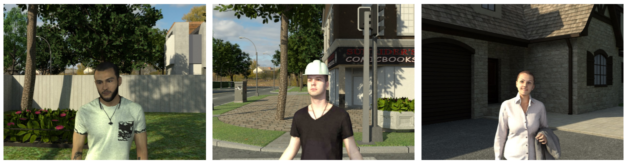


Figure 6. Representative input images from our in-house synthetic depth dataset.

We also analyze Lasinger *et al.*'s supervised MDE method on our test dataset. The method is trained on ≈ 2 million RGB-D pairs, most of which are derived from 3D movies which contain ample human subjects; thus, our test images are not unfamiliar with their model. Fig. 7 shows the result of [27] on selected images. It can be seen that the model has generalization issues in Fig. 7(b) and blur boundaries in Fig. 7(d). We do not intend to compare our method with [27] one-to-one because a fair comparison is not possible due to differences in training datasets. However, we want to highlight that our method is practically more relevant. Training data collection is an activity that is required periodically to adapt models to evolving test scenarios, extending the training data with our method is more practical compared to fully supervised methods. Furthermore, while relaxing training data collection efforts, our method makes no compromise with the output quality, rather it achieves high quality depth estimation.

6.1. Human-centric Synthetic Depth Dataset

In the literature, the standard benchmark datasets for MDE are mainly focused on autonomous driving [12], and indoor scenarios [40]. In this work, we generate synthetic RGB-D pairs for our work, which will be available to the community, and we expect it to be helpful in human-centric vision research, including learning portrait (bokeh) effect [20], human segmentation, and human depth estimation [41]. In Fig.6, we show representative images from our in-house synthetic depth dataset that we created using computer graphics software Blender [19]. Additional representative images and corresponding dense depth maps are provided in the supplementary.

For generating the dataset, a virtual scene was set up in Blender [19] and to capture semantically correct images, we manually defined human model positions, pose, gestures, and camera trajectories in the scene. Each rendering iteration applied one of the predefined valid position, pose, and gesture settings. As a result, the dataset contains different human models posing with varying gestures of hand and props (e.g., towel, toy, hat, etc.). The human models also have random poses relative to the camera. The RGB image rendering also includes natural sunshine and shadow cast by the light source (primarily the sun) on humans as well as other objects in the scene (see Fig. 6). The dataset has images captured in outdoor scenarios, such as roads, parks, and footpaths, with a high degree of realism.

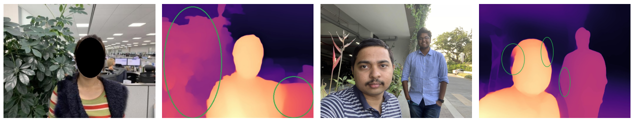

(a) Input 1 (b) Output 1 (c) Input 2 (d) Output 2

Figure 7. Assessing performance of Lasinger *et al.*'s [27] method trained on ≈ 2 million diversely curated RGB-D pairs

The final rendered dataset contains 3000 RGB-D pairs with 832x640 resolution with pixel-perfect dense ground-truth depth. The ground-truth depth ranges from 0.1 up to maximum 50 meters.

7. Conclusion

This work proposed a metrically accurate, sharp, and smoothly varying monocular depth estimation method without using real-world labels. It leverages a novel training methodology to synergistically combine positive aspects of two easy to obtain and scalable datasets, viz., monocular videos and synthetic dataset. The proposed method learns geometry and semantics from monocular videos, whereas scale-awareness and qualitative depth attributes, viz., sharp and smooth depth variations, are acquired from the synthetic dataset. To achieve this, a novel method was proposed that disentangles relative depth estimation with qualitative depth attributes from the task of scale-aware depth estimation. Despite any real world data labels and domain adaptation techniques, the proposed approach significantly outperforms the state-of-the-art on two challenging benchmark datasets while setting a new state-of-the-art in self-supervised and domain adaptation based monocular depth estimation. Further, we show that using the proposed method, an easily scalable and superior quality monocular depth estimation solution can be created for real life applications without real world data labelling efforts. Finally, to overcome the unavailability of high quality human-centric depth datasets, we will open our new synthetic depth dataset to aid ongoing human-centric vision research.

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 *CVPR*, 2018. 2, 3, 5

[2] Shubhra Aich, Jean Marie Uwabeza Vianney, Md. Amirul Islam, Mannat Kaur, and Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *ICRA*, 2021. 1

[3] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 2, 5

[4] Hemang Chawla, Arnav Varma, Elahe Arani, and Bahram Zonooz. Multimodal scale consistency and awareness for monocular self-supervised depth estimation. In *ICRA*, 2021. 2, 6, 7

[5] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *CVPR*, 2019. 7

[6] Xiaotian Chen, Yuwang Wang, Xuejin Chen, and Wenjun Zeng. S2r-depthnet: Learning a generalizable depth-specific structural representation. In *CVPR*, 2021. 2, 6

[7] Bin Cheng, Inderjot Singh Saggu, Raunak Shah, Gaurav Bansal, and Dinesh Bharadia. S^3net: Semantic-aware self-supervised depth estimation with monocular videos and synthetic data. In *ECCV*, 2020. 2, 3, 5, 6

[8] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 1

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 1, 5, 6

[10] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 1

[11] Ravi Garg, B. G. Vijay Kumar, Gustavo Carneiro, and Ian D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 1, 2

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Res.*, 2013. 1, 2, 5, 8

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 1

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6, 7

[15] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *ICCV*, 2019. 2

[16] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 1, 2, 6, 7

[17] Vitor Guizilini, Jie Li, Rares Ambrus, and Adrien Gaidon. Geometric unsupervised domain adaptation for semantic segmentation. In *ICCV*, 2021. 3, 6, 7

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[19] Roland Hess. *Blender Foundations: The Essential Guide to Learning Blender 2.6*. Focal Press, 2010. 2, 7, 8

[20] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *Workshops*, pages 418–419, 2020. 8

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, 2015. 3

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[23] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *ECCV*, 2020. 2

[24] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *CVPR*, 2018. 2, 3, 6

[25] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, 2017. 6, 7

[26] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision 3DV*, 2016. 1

[27] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 7, 8

[28] Jaehan Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *CVPR*, 2019. 1

[29] Jun Li, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single RGB images. In *ICCV*, 2017. 1

[30] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 7

[31] Ce Liu, Shuhang Gu, Luc Van Gool, and Radu Timofte. Deep line encoding for monocular 3d object detection and depth prediction. In *BMVC*, 2021. 1

[32] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, 2015. 1, 6

[33] 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 *CVPR*, 2016. 5

[34] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *ICRA*, 2019. 5

[35] Koutilya PNVR, Hao Zhou, and David Jacobs. Sharingan: Combining synthetic and real data for unsupervised geometry estimation. In *CVPR*, 2020. 2, 3, 5, 6

[36] Michaël Ramamonjisoa, Michael Firman, Jamie Watson, Vincent Lepetit, and Daniyar Turmukhambetov. Single image depth prediction with wavelet decomposition. In *CVPR*, 2021. 6

[37] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 7

[38] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009. 1, 7

[39] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 3

[40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 1, 8

[41] Feitong Tan, Hao Zhu, Zhaopeng Cui, Siyu Zhu, Marc Pollefeys, and Ping Tan. Self-supervised human depth estimation from monocular videos. In *CVPR*, 2020. 2, 8

[42] Lokender Tiwari, Pan Ji, Quoc-Huy Tran, Bingbing Zhuang, Saket Anand, and Manmohan Chandraker. Pseudo RGB-D for self-improving monocular SLAM and depth prediction. In *ECCV*, 2020. 6

[43] Neal Wadhwa, Rahul Garg, David E. Jacobs, Bryan E. Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T. Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Trans. Graph.*, 2018. 1, 2, 7

[44] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 4

[45] Lijun Wang, Yifan Wang, Linzhao Wang, Yunlong Zhan, Ying Wang, and Huchuan Lu. Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner? In *ICCV*, 2021. 2, 6

[46] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *ICCV*, 2019. 2, 5, 6

[47] 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 *CVPR*, 2018. 7

[48] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *CVPR*, 2020. 7

[49] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 1

[50] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *CVPR*, 2020. 6

[51] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *ECCV*, 2018. 6

[52] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 2, 5

[53] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *CVPR*, 2019. 2, 3, 5, 6

[54] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *CVPR*, 2020. 1, 2, 5

[55] Yunhan Zhao, Shu Kong, Daeyun Shin, and Charless C. Fowlkes. Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation. In *CVPR*, 2020. 2, 3, 6

[56] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *ECCV*, 2018. 2, 3, 6

[57] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 1, 2, 4, 5

[58] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In *ICCV*, 2021. 6

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3