Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation

Andra Petrovai Sergiu Nedevschi
Technical University of Cluj-Napoca, Romania
{andra.petrovai, sergiu.nedevschi}@cs.utcluj.ro

Abstract

We present a novel self-distillation based self-supervised monocular depth estimation (SD-SSMDE) learning framework. In the first step, our network is trained in a self-supervised regime on high-resolution images with the photometric loss. The network is further used to generate pseudo depth labels for all the images in the training set. To improve the performance of our estimates, in the second step, we re-train the network with the scale invariant logarithmic loss supervised by pseudo labels. We resolve scale ambiguity and inter-frame scale consistency by introducing an automatically computed scale in our depth labels. To filter out noisy depth values, we devise a filtering scheme based on the 3D consistency between consecutive views. Extensive experiments demonstrate that each proposed component and the self-supervised learning framework improve the quality of the depth estimation over the baseline and achieve state-of-the-art results on the KITTI and Cityscapes datasets.

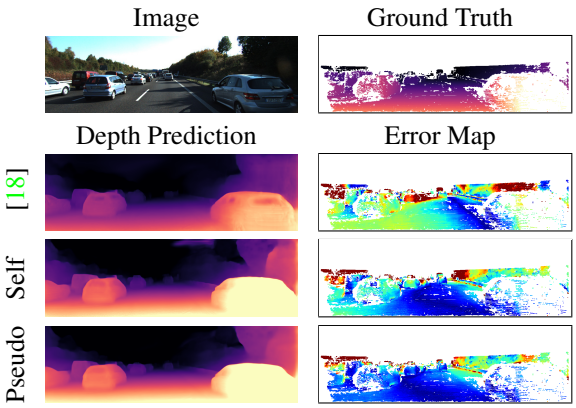

Figure 1. **SD-SSMDE, our self-distillation framework for self-supervised monocular depth estimation**. The teacher network (Self) trained in a self-supervised manner brings significant improvements over the baseline [18]. Our student model (Pseudo) is trained with pseudo labels generated with the previous model. The error is reduced, especially on the car on the right and on the entire left-hand side area of the image. In the error maps, small error is encoded with blue, while large error is encoded with red.

1. Introduction

One of the long-lasting research fields of computer vision is the accurate estimation of the 3D geometry of scenes. This includes tasks such as depth and ego motion prediction, which have major importance in the perception system of real-world applications, such as robotics and automated driving. While precise depth measurements can be directly obtained using specialized sensors such as LiDAR, they have several disadvantages such as high cost and reduced output density. As an alternative, estimating depth from images captured from a moving monocular or binocular system of cameras is attractive due to the lower cost and generally simpler setup. The stereo setup exhibits some limitations from a practical point of view as the stereo rig has to be carefully calibrated and synchronized.

Monocular depth estimation is an inherently ill-posed problem and initial results had lower performance compared to aforementioned methods. With the recent advances in deep learning, the performance gap has been reduced, especially in a supervised setting. The prohibitively large cost of collecting high-quality ground truth has led to the emergence of self-supervised monocular depth estimation, which unlocks the power of large-scale unlabeled datasets. Such approaches learn both the depth and ego motion, and embed 3D geometric constraints by using 3D reprojection models to synthesize consecutive images. More specifically, points from the target frame are back-projected in the camera coordinate system, displaced by the camera motion and reprojected onto adjacent source frames. In this way, the target image can be reconstructed from the source images, and the photometric difference between the target and synthesized image will be minimized during training.

Self-supervised monocular depth estimation relies on several assumptions that are not always true and hinder the

learning performance. It assumes that the scene is rigid and the camera is moving, that all image regions can be reconstructed from the neighboring frames and that all surfaces are Lambertian, *i.e.* have constant brightness. However, dynamic objects or a static camera, occlusions and illumination changes between consecutive views break these assumptions. Recent works address various issues [2,5,18,38] by designing masking techniques for filtering errors during training, by using stereo images or external information such as semantic segmentation and optical flow to guide the training process or improve the feature representation. However, in the self-supervised setting, propagating correct training signals is still difficult for all pixels, photometric loss can be high in occluded areas or for moving objects, and low in uniform texture areas or for repetitive structures.

In this paper, we propose a novel self-distillation based self-supervised learning framework for monocular depth estimation (**SD-SSMDE**) that leads to significant improvements when trained on monocular video and introduce the following contributions: (1) a two-stage self-distillation training strategy for monocular depth estimation: self-supervised first stage to generate high resolution pseudo labels and a supervised second stage using a similar or a more lightweight network (2) a novel architecture for the depth network for more accurate results (3) we solve scale ambiguity by incorporating the scale in pseudo labels, therefore depth predictions from the second stage are scaled and inter-frame scale-consistent (4) a filtering strategy based on 3D consistency between consecutive views to filter out large errors in pseudo labels. We perform extensive experiments on the KITTI and Cityscapes datasets and demonstrate that the proposed network and two-stage training framework yield state-of-the-art results and surpass or achieve on par results with current approaches.

2. Related Work

2.1. Supervised Monocular Depth Estimation

Depth estimation from a single image is an ill-posed problem since a 2D image can be generated from an infinity of 3D scenes. With the emergence of deep learning, Eigen *et al.* [12] formulated depth regression as a supervised learning problem. Since then, various improvements to network architectures [11,33,41,55] and loss functions [34,60] have been made. Xian *et al.* [36] models depth estimation as classification and obtains more robust results. However, the classification increases the complexity of the network and introduces challenges regarding the depth interval discretization. DORN [13] and SORD [10] propose improvements over the uniform discretization technique.

The aforementioned methods require ground truth depth, which is usually sparse depth from LiDAR scans. The difficulty to acquire ground truth has led to the development of semi-supervised methods that rely on weak labels such as relative depth [4], camera pose [59] or synthetic data [1,32,40]. Another line of research [30,37] proposes the use of conventional structure-from-motion methods [47], that are usually computationally intensive, to generate pseudo labels. Knowledge distillation from stereo depth estimates [7, 23, 42, 43, 53] has also been recently exploited for improved depth predictions. In contrast, in our framework we employ only monocular sequences and we do not rely on the availability of calibrated and synchronized stereo cameras.

2.2. Self-Supervised Monocular Depth Estimation

Early approaches on self-supervised monocular depth estimation [14, 17] were inspired by auto-encoders and employ stereo pairs during training. The SfmLearner [61] is the first solution working on monocular image sequences by jointly training a depth and pose estimation network.

Current approaches address some of the issues of self-supervised monocular learning. Monodepth2 [18] handles the lack of ego motion with an auto-masking of stationary pixels and the occlusion problem with a minimum reprojection loss. Low-texture areas are often problematic when using the photometric loss, therefore feature-based reconstruction losses [48, 59] have been proved more robust. Formulating self-supervised depth estimation as a depth classification problem has been tackled in [19, 27]. Other works improve the network architecture [21] or include test-time refinement procedures [2, 5]. Feature representation learning is guided with semantic networks or single-view reconstruction auto-encoder networks in several approaches [22,28,49]. For extracting potentially moving objects, instance or semantic segmentation have been used in [2,29,50]. Other methods [5,20,38,46,58] employ external optical-flow networks and design selective masking techniques to avoid propagating large errors in the training signal for moving objects. ManyDepth [54] uses multi-frame input at test-time for improved results.

The idea of self-distillation has been approached in several works. Compared to [39] and [44], we generate high-resolution pseudo labels, which can be further distilled by a similar or more lightweight student network that can be trained on low or high resolution images. Poggi *et al.* [44] proposes a filtering scheme by estimating the uncertainty of the depth output, while our filtering scheme relies on measuring the 3D consistency between consecutive views. Yang *et al.* [57] devises a filtering scheme based on depth reprojection error, however in the context of multi-view stereo. Compared to [57] we adopt a minimum reprojection error from multiple source images to account for pixels which are visible in the target image but are occluded in one of the source images. An important contribution of our work which has not been previously investigated is solving scale ambiguity: the student network learns from inter-

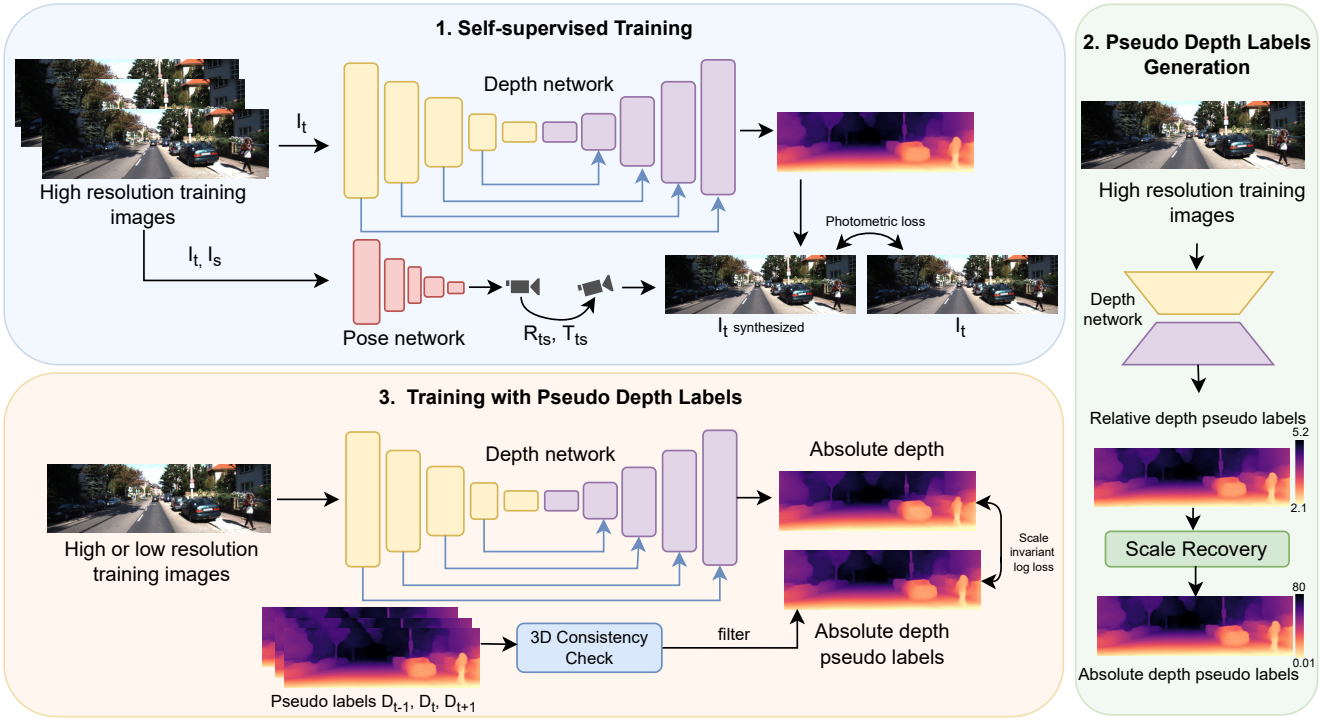

Figure 2. **Our SD-SSMDE training framework.** In the first stage, we train a depth estimation teacher network and a camera pose network in a self-supervised manner. Using the trained depth network, we generate pseudo labels for all the images in the training set. Automatic scale recovery is performed in order to obtain absolute depth values. In the second stage, we train the depth student network from scratch and regress depth maps supervised by the previously generated pseudo labels. In order to remove erroneous depth estimates from pseudo labels, a consistency check is performed, for the same 3D point computed from different views.

frame scale-consistent pseudo labels that have been previously scaled to absolute depth values. With the proposed contributions, we outperform both [39] and [44].

3. SD-SSMDE Method

In this section we describe the self-distillation based two-stage training pipeline and the improved depth network architecture. We also provide details about implementation.

3.1. Self-Distillation based Training Pipeline

We propose a two-stage training pipeline for monocular depth estimation that requires only video frames and no depth ground truth data. In the first stage, the camera pose network [18] and depth teacher network are trained in a self-supervised manner on high-resolution images with the photometric loss. Next, the network is used to infer depth on the entire training set. Since the depth outputs differ from the real-world depth by a scale factor, we employ a scale recovery module [56] to recover the true depth. In the second stage, the camera pose network is fixed and we instantiate a new depth student network having the same or a more lightweight architecture. The student depth network is trained from scratch in a supervised regime with the pseudo labels. During this training phase, a mask is generated on-the-fly, which filters out depth locations with large errors from the loss computation. Starting from the assumption that the same scene captured in three consecutive images should have a high level of 3D consistency in different views, we filter out locations that have high deviations between the 3D points in the camera coordinate system.

3.2. Depth Network Architecture

Our encoder-decoder depth network used by both the teacher and student network follows the design of the panoptic segmentation network Panoptic-DeepLab [6, 45] with several changes. We employ backbone output stride 32 instead of 16. An Atrous Spatial Pyramid Pooling (ASPP) [3] with parallel dilated depthwise separable convolutions [26] extracts context information from the backbone output. The decoder consists of five upsampling stages, where the spatial resolution is gradually increased by a factor of two. Each upsampling stage consists of upsampling, concatenation with low-level features from the backbone and a $[5 \times 5, 256]$ depthwise separable convolution for feature fusion. The low-level features from $1/16$ to $1/2$ are projected to $\{128, 64, 32, 16\}$ channels before concatenation. After

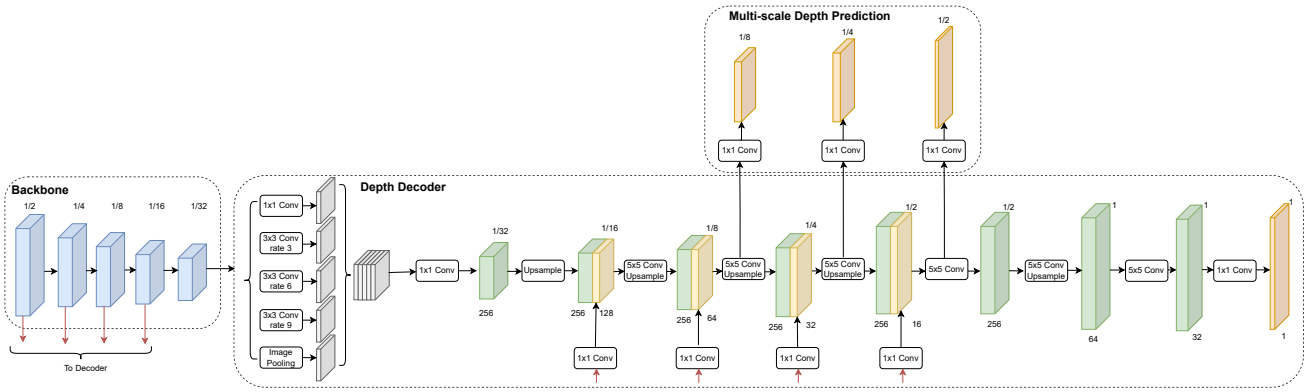

Figure 3. **Our depth network architecture.** We introduce a new depth decoder which provides high quality depth predictions.

the last upsampling stage, a $[5 \times 5, 64]$ depthwise separable convolution and a $2\times$ bilinear upsampling to the original image resolution follow. Finally, two convolutional layers with $[5\times5, 32]$ and $[1\times1, 1]$ yield the final depth map. During training, we adopt multi-scale depth prediction at four scales $\{1/8, 1/4, 1/2, 1\}$ and introduce a $[1\times1, 1]$ convolution after the fusion. Losses are computed using the multi-scale depth predictions [14, 17, 18]. A depiction of our network can be found in Figure 3.

3.3. Self-Supervised Monocular Depth Estimation

The teacher network is trained in a self-supervised manner in the first stage of training. The goal of self-supervised depth estimation from a single image is to predict a depth map aligned to the input image without using any ground truth data during training. The basic mechanism behind the method relies on geometric projections that allow view-synthesis of adjacent frames based on the predicted depth. During inference, the network takes a single image and predicts the depth, but during training three consecutive frames are employed. Two separate networks, a depth estimation and a camera pose estimation network, are jointly trained. The depth estimation network actually learns the disparity, which is the inverse of depth as it was proved to be more robust [14, 18].

Consider a target image I_t and adjacent source images I_s, where $s = \{t-1, t+1\}$ captured by a moving camera. Let $M_{t\to s}$ be the camera pose that defines the 3D translation $T_{t\to s}$ and rotation $R_{t\to s}$ between consecutive 3D scene positions:

$$M_{t\to s} = \begin{bmatrix} R_{t\to s} & T_{t\to s} \\ 0 & 1 \end{bmatrix} \tag{1}$$

Let K be the intrinsic camera matrix, which is known in advance and fixed for the entire dataset [18].

Given a pixel p in the target frame its corresponding position in 3D in homogeneous coordinates x can be computed by backprojection using the predicted target depth:

$$x = \begin{bmatrix} D_t(p)K^{-1}p \\ 1 \end{bmatrix} \tag{2}$$

Assuming camera motion and static scene, x can be reprojected in the source frame I_s after displacing it using the camera pose $M_{t\to s}$:

$$p' = \begin{bmatrix} K|0 \end{bmatrix} M_{t\to s} x \tag{3}$$

The target image is synthesized $I_{s\to t}$ by sampling the source images I_s with bilinear interpolation [17, 18], which we denote with $I_s\langle p' \rangle$. The per-pixel photometric reprojection error $\mathcal{L}_p$ between the target image and the synthesized images is minimized during training. To account for occlusions between views, the minimum reprojection error over all source images is computed as in [18]:

$$\mathcal{L}_p = \min_s pe(I_t, I_{s\to t}) \tag{4}$$

The photometric error pe is the weighted sum between the structural similarity SSIM [52] and L1 error:

$$pe(I_a, I_b) = \alpha \frac{1 - \text{SSIM}(I_a, I_b)}{2} + (1-\alpha) \left\| I_a - I_b \right\|_1 \tag{5}$$

We also adopt an edge-aware smoothness loss [17, 18] that encourages local smoothness in the presence of low image gradient.

3.4. Scale Recovery in Pseudo Labels

The output of the self-supervised depth estimation network is relative depth, *i.e.* depth values for an image are in broad agreement with each other, however they differ from real-world values by a scale factor. We employ the technique from [56] to compute the scale factor. The idea behind the scale recovery module is to find the scale between an estimated and the real camera height. The first step is

to determine the ground points. This is done by computing a surface normal for each 3D point and finding the points that have a normalized normal close to the ideal ground normal $n = (0, 1, 0)^\top$ based on a similarity function. After identifying the ground points, a set of camera heights is estimated for each 3D point. The last step is to compute the depth scale factor as the ratio between the real camera height and the median of the estimated heights. Each pseudo label is scaled with the estimated scale factor, such that we obtain absolute depth values and scale consistent pseudo labels across frames.

3.5. Supervised Monocular Depth Estimation

In the second stage, for training the student network, we formulate the depth estimation task as a supervised regression problem. We adopt the scale-invariant log loss [12,34]:

$$\mathcal{L}_{sp} = \gamma \sqrt{\frac{1}{N}\sum_i d_i^2 - \frac{\lambda}{N^2}\left(\sum_i d_i\right)^2} \tag{6}$$

where $d_i = \log y_i - \log \bar{y_i}$, y_i is the predicted depth and $\bar{y_i}$ is the pseudo ground-truth depth. N represents the number of pixels with valid values and λ is a weighting factor. We scale the range of the loss with γ in order to improve convergence.

During inference, we directly predict the depth values using the logits from the depth regression head, by applying sigmoid and scaling the predicted values by a constant value, which we set to 80 for both KITTI [15] and Cityscapes, where the usual depth range is [0-80m].

3.6. Filtering Errors in Pseudo Labels

Given the high-resolution depth pseudo labels, we check the 3D consistency between consecutive views. This check is valid due to the fact that pseudo labels are inter-frame consistent after scaling them to absolute depth values. We only keep reliable depth estimates for pixels which have similar 3D coordinates in different views. This masking process is done on-the-fly during the second stage training.

Assume we know the intrinsic matrix K and the camera pose M between the target and source coordinate system given by the pose network, which has been trained in the previous stage. Let p be a pixel in the target image. The 3D point x corresponding to p can be obtained by backprojection using equation 2 and its 2D coordinates p' in the source image can be computed using equation 3. Since p' has real valued coordinates, the source depth D_s is sampled with bilinear interpolation $D_s\langle p'\rangle$ and backprojected to the source camera coordinate system x'. Finally, the 3D point x' is displaced using the camera pose $M_{s\to t}$ to the common coordinate system of the target camera. Then, the absolute difference between the two 3D points on the z axis is computed. If the difference is smaller than a predefined threshold T, the point is valid, otherwise it will be filtered out. We adopt a minimum 3D consistency error between the target and adjacent source views in order to account for possible occlusions that may occur in one of the source views. The resulting mask F is computed as the Iverson bracket:

$$F = \left[\min_s \left\| M_{s\to t} D_s\langle p'\rangle K^{-1}p' - D_t(p)K^{-1}p)\right\|_1 < T\right] \tag{7}$$

3.7. Implementation Details

For the depth prediction network, we employ the backbone [24] pretrained on Imagenet [31]. For the self-supervised network, the output of the sigmoid layer is converted to depth with $D = 1/(a\sigma + b)$, where $a = 0.1$ and $b = 100$ represent the scaling interval.

The pose estimation network has a lightweight architecture [18] with a ResNet-18 backbone [24]. The network takes as input pairs of color images, target and source, and predicts the 6DOF camera pose, the translation vector and rotation matrix in terms of three Euler angles. During inference, the pose network is discarded.

On KITTI, the networks are trained in both training stages with a minibatch of 12 images for 66k iterations, using Adam optimizer and a base learning rate of 10^{-4}. In the first stage, we employ step learning rate decay and drop the learning rate by 10 at 50k iterations. In the second stage, we employ polynomial learning rate decay and train for the same number of iterations. On Cityscapes, we train with a minibatch of 12 images for 12k iterations in the first stage and 30k iterations in the second. We apply image augmentation during training, such as random horizontal flipping and random color augmentation with the settings from [18]. In the self-supervised training stage, we weight the smoothness loss by 0.001 and we set α to 0.85 in the photometric loss. In the supervised loss, we set λ to 0.85 and γ to 10 [34]. The threshold T for 3D consistency masking is 1. The self-supervised loss, as well as the supervised loss are computed at four scales.

4. Experiments

In this section, we evaluate our SD-SSMDE teacher and student models on the KITTI and Cityscapes datasets. We perform extensive ablation studies and compare our results with other approaches using standard evaluation metrics.

4.1. Datasets

KITTI [16] is a driving dataset captured in urban, rural and highway areas. We employ the Eigen splits [11] with the pre-processing of Zhou *et al.* [61] where static frames are removed. The training set consists of 39,810 image triplets, while the validation set has 4,424 images. The reported results are evaluated on 697 test images using Garg's

Model	GT Scaling	Auto Scaling	Fixed Scaling	Resolution	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Supervised (reference)			✓	192 × 640	0.097	0.645	4.296	0.180	0.892	0.964	0.983
Self-supervised (teacher)	✓			192 × 640	0.104	0.768	4.513	0.180	0.892	0.964	0.983
Self-supervised (teacher)		✓		192 × 640	0.108	0.795	4.655	0.192	0.878	0.959	0.981
Pseudo-supervised (student)			✓	192 × 640	**0.100**	**0.661**	**4.264**	**0.172**	**0.896**	**0.967**	**0.985**
Supervised (reference)			✓	320 × 1024	0.091	0.567	4.137	0.177	0.902	0.966	0.983
Self-supervised (teacher)	✓			320 × 1024	0.101	0.720	4.339	0.176	0.898	0.967	0.984
Self-supervised (teacher)		✓		320 × 1024	0.104	0.747	4.453	0.185	0.885	0.963	0.983
Pseudo-supervised (student)			✓	320 × 1024	**0.098**	**0.674**	**4.187**	**0.170**	**0.902**	**0.968**	**0.985**

Table 1. **Ablation study for the self-distillation based two-stage self-supervised learning framework.** We perform experiments with ResNet-50 backbone, two image resolutions and three scale recovery methods during inference. In the second stage, we always train on scaled pseudo labels generated from the high-resolution self-supervised model and use 3D consistency check filtering.

Model	Backbone	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25$ ↑
Monodepth2 [18]	ResNet-18	0.115	0.903	4.863	0.193	**0.877**
Ours	ResNet-18	**0.112**	**0.854**	**4.839**	**0.190**	0.876
Monodepth2 [18]	ResNet-50	0.110	0.831	4.642	0.187	0.883
Ours	ResNet-50	**0.104**	**0.768**	**4.513**	**0.180**	**0.892**

Table 2. **Ablation study for the depth decoder of the self-supervised teacher network.** By changing the decoder we obtain significant improvements compared to the Monodepth2 baseline [18]. The network is trained on medium resolution images.

Gt Scaled PS	Auto Scaled PS	Filtering	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25$ ↑
✓			0.102	0.716	4.351	0.177	0.887
	✓		0.103	0.729	4.457	0.181	0.881
	✓	✓	**0.100**	**0.661**	**4.264**	**0.172**	**0.896**

Table 3. **Ablation study for the pseudo supervised training of our student network.** In this stage, we train with pseudo labels (PS) generated with the self-supervised teacher network. The pseudo labels are either scaled with the ground truth or scaled with an automatic scale recovery method [56]. Performing a 3D consistency check to filter out errors from the pseudo labels is beneficial.

crop [14]. We cap the depth values to 80m as in [18]. In Table 9 from the supplementary section B we evaluate our model on the improved KITTI ground truth [51] on the 652 available frames from the Eigen test set.

Cityscapes [9] is an urban driving dataset with high resolution images. The training set has 69,730 image triplets, while the validation set consists of 1,525 images. We employ the cropping and evaluation scheme from [35] and give more details in the supplementary section C. Depth values are capped at 80m as with KITTI.

4.2. Ablation Study on KITTI

Self-distillation based learning framework. In Table 1, we perform ablation experiments related to the self-supervised learning framework. In our experiments, we train our networks on one of the two image resolutions: the medium resolution 192 × 640 and high resolution 320 × 1024. In the first experiment, our depth network is trained in a supervised regime with the improved KITTI ground truth [51]. The depth network with ResNet-50 backbone and the proposed decoder is trained with the scale invariant logarithm loss [34] for depth regression. We train the network under the same conditions as the self-supervised method and with the same hyperparameters. Next, we train the teacher network in the self-supervised regime. Self-supervised methods suffer from scale ambiguity, *i.e.* the output is not scaled to real-world values. We experiment with ground truth median scaling [18], as it is common practice, and an automatic scale recovery method [56] during inference. Adopting the automatic scaling, the error increases, since the scale computed from the predicted depth maps is not as accurate. Using our best model with a ResNet-50 backbone, trained on high-resolution images and using automatic scaling, we generate pseudo labels for the entire training set. In the second stage, the student depth network is supervised by the pseudo labels. A 3D consistency check is applied in order to remove noisy estimates and although the labels will be more sparse, they will also be more accurate. We obtain improved results from the second stage training, for both image resolutions. During inference, a fixed scale factor is employed to map the depth

	Depth estimation error		Pseudo labels error		
Threshold (m)	AbsRel ↓	RMS ↓	% Filtered	AbsRel ↓	RMS ↓
no filtering	0.103	4.457	0	0.082	3.995
1.0	**0.100**	**4.264**	18	**0.069**	**3.245**
1.5	0.101	4.364	12	0.072	3.416

Table 4. **Filtering scheme ablation.** Comparison between student network training with or without pseudo label filtering on the KITTI test set. We also measure the error of pseudo labels on the training set and the amount of 3D points that are filtered.

Model	PS	σ_{scale}	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25$ ↑
Monodepth2 [18] (R18)	-	0.093	0.109	0.623	4.136	0.154	0.873
SD-SSMDE (R50)	unscaled	0.100	0.109	0.494	3.591	0.141	0.888
SD-SSMDE (R18)	scaled	0.061	0.084	0.436	3.550	0.128	0.918
SD-SSMDE (R50)	scaled	**0.040**	**0.076**	**0.377**	**3.304**	**0.117**	**0.933**

Table 5. **Scale variance analysis.** Comparison on KITTI Eigen test split with improved ground truth [51] on 192 × 640 resolution. Our student network learns from unscaled or automatically scaled [56] pseudo labels (PS). During inference, we compute the standard deviation σ_{scale} of individual ground truth median scales. All depth predictions are scaled with a fixed scale factor.

Method	Backbone	Sem	Resolution	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
GeoNet [58]	ResNet-50		192 × 640	0.153	1.328	5.737	0.23	0.802	0.934	0.972
DF-Net [62]	ResNet-50		192× 640	0.146	1.182	5.215	0.213	0.818	0.943	0.978
Guizilini *et al.* [22]	ResNet-50	✓	192 × 640	0.113	0.831	4.663	0.189	0.878	0.971	0.983
SGDepth [29]	ResNet-50	✓	192 × 640	0.112	0.833	4.688	0.190	0.884	0.961	0.981
Monodepth2 [18]	ResNet-50		192 × 640	0.110	0.831	4.642	0.187	0.883	0.962	0.982
FSRE-Depth [28]	ResNet-50	✓	192 × 640	0.102	0.675	4.393	0.178	0.893	0.966	0.984
SD-SSMDE (ours)	ResNet-50		192 × 640	**0.100**	**0.661**	**4.264**	**0.172**	**0.896**	**0.967**	**0.985**
Shu *et al.* [48]	ResNet-50	✓	320 × 1024	0.104	0.729	4.481	0.179	0.893	0.965	0.984
SD-SSMDE (ours)	ResNet-50		320 × 1024	**0.098**	**0.674**	**4.187**	**0.170**	**0.902**	**0.968**	**0.985**
Guizilini *et al.* [22]	ResNet-18	✓	192 × 640	0.117	0.854	4.714	0.191	0.873	0.963	0.981
Monodepth2 [18]	ResNet-18		192 × 640	0.115	0.903	4.863	0.1F93	0.877	0.959	0.981
SGDepth [29]	ResNet-18	✓	192 × 640	0.113	0.835	4.693	0.191	0.879	0.961	0.981
Poggi *et al.* [44]	ResNet-18		192 × 640	0.111	0.863	4.756	0.188	0.881	0.961	0.982
HR-Depth [39]	ResNet-18		192 × 640	0.109	0.792	4.632	0.185	0.884	0.962	0.983
FSRE-Depth [28]	ResNet-18	✓	192 × 640	**0.105**	**0.722**	4.547	0.182	**0.886**	**0.964**	**0.984**
SD-SSMDE (ours)	ResNet-18		192 × 640	0.106	0.751	**4.485**	**0.180**	0.885	**0.964**	**0.984**
Monodepth2 [18]	ResNet-18		320 × 1024	0.115	0.882	4.701	0.190	0.879	0.961	0.982
SGDepth [29]	ResNet-18	✓	384 × 1280	0.107	0.768	4.468	0.186	0.891	0.963	0.982
HR-Depth [39]	ResNet-18		320 × 1024	0.106	0.755	4.472	0.181	0.892	0.966	0.984
FSRE-Depth [28]	ResNet-18	✓	320 × 1024	0.102	**0.687**	4.366	0.178	0.895	**0.967**	0.984
SD-SSMDE (ours)	ResNet-18		320 × 1024	**0.101**	0.700	**4.332**	**0.174**	**0.895**	0.966	**0.985**

Table 6. **Comparison with the state-of-the-art on KITTI Eigen test set.** We report results of methods that use only a single image during inference. *Sem* denotes the use of semantic segmentation. Best results are in bold.

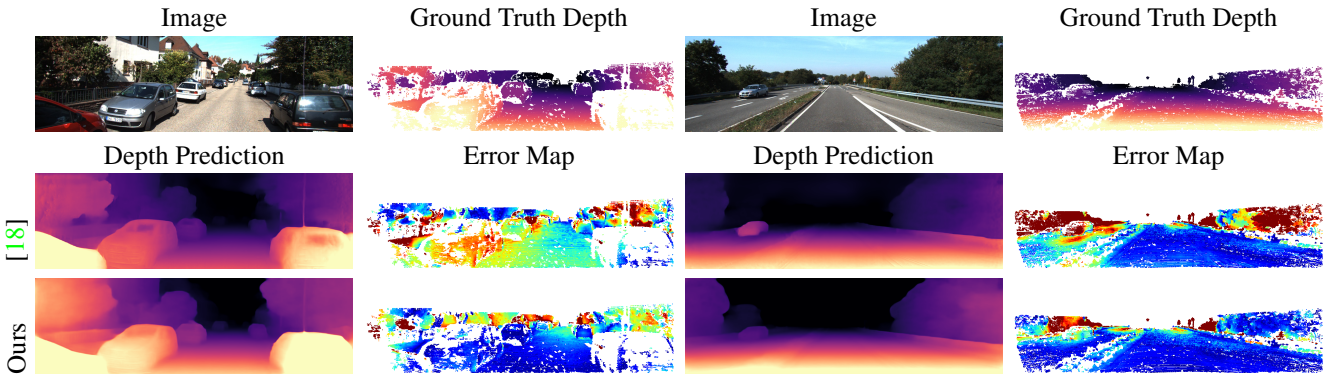

Figure 4. **Qualitative results on the KITTI Eigen test set with improved ground truth.** We compare our SD-SSMDE results with Monodepth2 [18] with ResNet-50 [18] on high resolution images. Our network provides better depth quality with smaller errors.

values in the interval [0, 80m].

Depth decoder. In Table 2, we present the results of our baseline Monodepth2 [18] and our teacher network trained in a self-supervised regime with the photometric loss. We employ the same loss functions as [18], but we propose a different decoder for the depth network. With both the lightweight ResNet-18 and the deeper ResNet-50 [24] backbones, we achieve a lower error than [18]. The proposed decoder is able to better capture context due to the ASPP [6] module and the higher number of channels in each convolutional layer.

Scale recovery and filtering. We perform ablation studies for the second part of our training pipeline in Table 3. We employ medium resolution images for these experiments. First, we would like to see what is the impact of training with pseudo labels scaled with a different scaling factor. In the first experiment, we train our student depth network supervised by the high resolution pseudo labels scaled with the ground truth median scaling. With this setting, in which no error filtering is performed, we improve the results over the self-supervised counterpart. Although trained with noisy labels, the network is able to converge to a better minimum. We believe that this is because in the self-supervised learning stage, the network becomes stuck in a local minimum due to the use of the reprojection loss, while when trained with labels, even noisy ones, the regression loss will guide the network to a global minimum. Our findings agree with other studies that provide depth hints

from stereo [53]. In our second experiment, we train our network with the pseudo labels scaled by an off-the-shelf scale recovery module [56] and no ground truth data is used. Interestingly, we obtain similar results to training with ground-truth scaled depth maps. Removing the dependence on ground truth is a big advantage, therefore in our final experiment we filter the automatically-scaled pseudo labels by using the 3D consistency check in order to get a better training signal. As expected, training with higher quality labels further improves the result.

In Table 4 we perform an ablation study on the threshold used in the filtering scheme and measure the depth error on pseudo labels before and after filtering. By applying the filtering scheme we obtain more accurate but also sparser pseudo labels. A threshold of $T = 1$ achieves the best balance between accuracy and density.

In Table 5 we demonstrate the advantage of using scaled pseudo labels during training. By scaling the pseudo labels with an automatic scale recovery method, not only do we obtain absolute depth values but also inter-frame scale consistent depth labels. Our experiments suggest that training with scaled pseudo labels is required for improved performance. We also perform a scale variance analysis of the depth output of the student network. The scale is equal to the median of all individual ratios of the ground truth depth and the predicted depth maps medians. We report the standard deviation of individual scales σ_{scale} where a lower value indicates increased scale-consistent depth predictions across frames. The best scores and the most scale-consistent predictions are obtained with our SD-SSMDE model with ResNet-50 backbone trained with scaled pseudo labels.

4.3. KITTI Results

In Table 6 we compare our results with the state-of-the-art methods that perform inference on a single image. When trained with medium resolution images, our network with ResNet-50 backbone outperforms all other networks. We also surpass our baseline by a significant margin. With ResNet-18 on medium resolution images, we achieve comparable results with FSRE-Depth [28]. Works such as [22, 28, 29] use semantic segmentation guidance and rely on the existence of pre-trained semantic networks and pixel-level semantic annotations, which may be difficult and expensive to acquire. On the other hand, our network achieves the best results, while being trained only on monocular sequences, without extra data. Another advantage of our method is that we can completely remove the dependence on ground truth data during inference by using a fixed scaling factor with no or minimal loss in accuracy. A disadvantage of our two-stage training framework would be the longer training time, however there is no additional computation cost during inference, which is important from a practical perspective. For high-resolution images, we obtain the best scores overall.

Model	Train	Test	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓
Struct2Depth 2 [2]	C	C	0.145	1.737	7.280	0.205
Monodepth2 [18]	C	C	0.129	1.569	6.876	0.187
Videos in the Wild [20]	C	C	0.127	1.330	6.960	0.195
Li *et al.* [35]	C	C	0.119	1.290	6.980	0.190
Choi *et al.* [8]	C	C	0.115	1.125	6.584	0.195
SD-SSMDE (teacher - GT scaling)	C	C	0.117	1.090	6.468	0.176
SD-SSMDE (student - fixed scaling)	C	C	0.114	1.017	**5.949**	0.169
SD-SSMDE (student - GT scaling)	C	C	**0.110**	**0.988**	5.953	**0.165**
Monodepth2 [18]	K	C	0.153	1.785	8.590	0.234
SD-SSMDE (student - fixed scaling)	K	C	**0.143**	**1.635**	**8.441**	**0.221**

Table 7. **Results on Cityscapes.** Evaluation of models on the Cityscapes dataset, trained on Cityscapes (C) or on KITTI (K). All the competing methods use ground truth median scaling. The input/output resolution for our network is 128×416. Full metrics can be found in the supplementary material in Table 10.

Figure 4 presents a qualitative comparison between Monodepth2 [18] and our results from the student network. We observe that our network yields more accurate depth maps, as seen for example on the ground and on the vehicles. We provide more qualitative results in the supplementary in Figure 6.

4.4. Cityscapes Results

In Table 7 we evaluate our models with the ResNet-50 backbone on the Cityscapes dataset and compare them with the state-of-the-art. We also check the generalization capability of our model trained on KITTI, without any fine-tuning. Compared to Monodepth2 and other competing methods, we achieve better scores across all metrics.

5. Conclusions

We have presented a novel self-distillation based two-stage self-supervised training framework for monocular depth estimation: in the first stage, a self-supervised depth network and camera pose estimation network are trained on monocular sequences, in the second stage, the depth network is trained on high-resolution pseudo labels generated with the first network. We also introduced a new architecture for the depth network which brings significant improvements. We investigated the importance of training with scaled pseudo labels and its effect on depth predictions scale consistency among frames. In the second stage, a filtering scheme based on 3D consistency between consecutive views was proposed for a more accurate supervision signal. Our SD-SSMDE models achieve state-of-the-art results on the KITTI and Cityscapes datasets.

Acknowledgment

This work was supported by the SEPCA grant funded by the Romanian Ministry of Education and Scientific Research, code PN-III-P4-ID-PCCF-2016-0180.

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] 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, 8

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 3

[4] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in Neural Information Processing Systems*, 29:730–738, 2016. 2

[5] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7063–7072, 2019. 2

[6] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020. 3, 7

[7] Hyesong Choi, Hunsang Lee, Sunkyung Kim, Sunok Kim, Seungryong Kim, Kwanghoon Sohn, and Dongbo Min. Adaptive confidence thresholding for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12808–12818, 2021. 2

[8] Hyesong Choi, Hunsang Lee, Sunkyung Kim, Sunok Kim, Seungryong Kim, Kwanghoon Sohn, and Dongbo Min. Adaptive confidence thresholding for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12808–12818, October 2021. 8

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 6

[10] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4738–4747, 2019. 2

[11] 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. 2, 5

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014. 2, 5

[13] 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. 2

[14] 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. 2, 4, 6

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 5

[16] A Geiger, P Lenz, and R Urtasun. Are we ready for autonomous driving? *The kitti vision benchmark suite,in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 5

[17] Clément Godard, Oisin 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. 2, 4

[18] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *The International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[19] 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

[20] 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. 2, 8

[21] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, 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. 2

[22] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *International Conference on Learning Representations*, 2020. 2, 7, 8

[23] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018. 2

[24] 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. 5, 7

[25] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2007.

[26] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[27] 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

[28] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12642–12652, 2021. 2, 7, 8

[29] 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, 7, 8

[30] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–713, 2018. 2

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012. 5

[32] 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

[33] 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. 2

[34] 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, 5, 6

[35] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes. *arXiv preprint arXiv:2010.16404*, 2020. 6, 8

[36] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust prediction. In *Asian Conference on Computer Vision*, pages 663–678. Springer, 2018. 2

[37] 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

[38] 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. 2

[39] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2294–2301, May 2021. 2, 3, 7

[40] 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

[41] 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

[42] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15560–15569, 2021. 2

[43] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019. 2

[44] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020. 2, 3, 7

[45] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, 2021. 3

[46] 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. 2

[47] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 2

[48] 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. 2, 7

[49] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019. 2

[50] 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

[51] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 International Conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 6

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4

[53] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2162–2171, 2019. 2, 8

[54] Jamie Watson, Oisin 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. 2

[55] 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

[56] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2330–2337. IEEE, 2020. 3, 4, 6, 8

[57] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Self-supervised learning of depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7526–7534, 2021. 2

[58] 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. 2, 7

[59] 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

[60] 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

[61] 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. 2, 5

[62] 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. 7