

A Large RGB-D Dataset for Semi-supervised Monocular Depth Estimation

Jaehoon Cho, *Student Member, IEEE*, Dongbo Min, *Senior Member, IEEE*,
 Youngjung Kim, *Member, IEEE*, and Kwanghoon Sohn, *Senior Member, IEEE*

Abstract—The recent advance of monocular depth estimation is largely based on deeply nested convolutional networks, combined with supervised training. However, it still remains arduous to collect large-scale ground truth depth (or disparity) maps for supervising the networks. This paper presents a simple yet effective semi-supervised approach for monocular depth estimation. Inspired by the human visual system, we propose a student-teacher strategy in which a shallow student network is trained with the auxiliary information obtained from a deeper and accurate teacher network. Specifically, we first train the *stereo* teacher network fully utilizing the binocular perception of 3D geometry, and then use depth predictions of the teacher network for supervising the student network for monocular depth inference. This enables us to exploit all available depth data from massive unlabeled stereo pairs that are relatively easier-to-obtain. We further introduce a data ensemble strategy that merges multiple depth predictions of the teacher network to improve the training samples for the student network. Additionally, stereo confidence maps are provided to avoid inaccurate depth estimates being used when supervising the student network. Our new training data, consisting of 1 million outdoor stereo images taken using hand-held stereo cameras, is hosted at the project webpage¹. Lastly, we demonstrate that the monocular depth estimation network provides feature representations that are suitable for some high-level vision tasks such as semantic segmentation and road detection. Extensive experiments demonstrate the effectiveness and flexibility of the proposed method in various outdoor scenarios.

Index Terms—Monocular Depth Estimation, convolutional neural networks, knowledge transfer, semi-supervised learning, stereo dataset, confidence measure.

I. INTRODUCTION

Obtaining 3D depth of a scene is essential to alleviate a number of challenges in computer vision tasks including 3D reconstruction [1], autonomous driving [2], intrinsic image decomposition [3], and scene understanding [4]. The human visual system (HVS) can understand 3D structure by measuring an absolute depth value of the scene through binocular fusion. 3D structure is perceived using a binocular disparity that is inferred from slightly different images of the same scene. Such a mechanism has widely been adopted in the computational stereo approaches that establish correspondence maps across two (or more) images taken for the same scene [5], achieving

J. Cho and K. Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: {rehoon, khsohn}@yonsei.ac.kr).

D. Min is with the Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, South Korea (e-mail: dbmin@ewha.ac.kr).

Y. Kim is with Agency for Defense Development, Daejeon 34186, South Korea (e-mail: read12300@add.re.kr).

¹<https://dimlrgbd.github.io/>

an outstanding performance in recent approaches [6]–[11]. Interestingly, even with a single image, the HVS is capable of interpreting 3D structure thanks to a prior knowledge learned from monocular cues such as shading, motion, texture, and relative size of objects [12]. In order to imitate the 3D perception capability of the HVS, numerous monocular depth estimation approaches have been developed based on the monocular cues, e.g., including object contour [13], segment [14], and shading [15]. However, most methods rely heavily on hand-crafted rules based on one or few monocular cues, and thus they often fail to capture plausible depth from a single image and work only at very restricted environments. It is almost impossible to design very sophisticated rules that take into account all cases in a hand-crafted fashion due to its highly ill-posed characteristics.

Recently, deep neural networks have revolutionized various computer vision tasks. Depth prediction from a single image has also been advanced considerably by making use of convolutional neural networks (CNNs) [16]–[22]. Such a great success stems mainly from the effective representation learning of CNNs. However, supervised learning approaches for monocular depth estimation necessarily have several limitations. Unlike tasks that require an image-level supervision only, e.g., image classification [23], the monocular depth estimation needs (semi-)dense depth maps as a pixel-level supervision for training the deep network, and constructing such a large-scale training data with depth maps is extremely challenging. An active depth sensor, LiDAR, is commonly used to acquire depth maps, but they are usually of low resolution and very sparse, e.g., less than 6% in the KITTI dataset [24]. Due to its sparsity, it can not cover all salient objects in a scene. Additionally, the sensing device is very expensive and often suffers from several internal degradations such as imperfect sensor calibration and photometric distortions. Thus, existing public datasets all provide only a small number of depth maps for rather limited scenes, e.g. mostly consisting of driving scenes obtained from the depth sensor mounted on a vehicle [24], [25].

The problem with insufficient training data is somewhat alleviated by leveraging the data augmentation and the pre-trained model with ImageNet [23], though not perfect. However, the lack of diversity of training data incurs severe domain adaptation issues. A depth accuracy is substantially degraded when the trained model is adapted to different data or novel environment. For instance, performing an inference at the Cityscapes dataset [25] (novel domain) using the monocular depth estimation network trained with the KITTI dataset

[24] (target domain) often leads to substantial performance degradation, though both the KITTI and Cityscapes datasets contain driving scenes (Fig. 1). The domain adaptation issue becomes even worse, when testing at non-driving scenes such as park and trail using the deep network trained with the KITTI dataset. This will be validated in more details in the experiment section. Note that from the best of our knowledge, most approaches for monocular depth estimation [16]–[22] have been trained and tested using the Eigen split [17] of the KITTI dataset.

In this paper, we explore a special regime of *semi-supervised* learning method based on the student-teacher strategy [26] to address the lack of massive ground truth depth maps in learning the monocular depth inference network. It has been known that a deep and wide model tends to be more accurate than a shallow model and provides a large capacity [27], [28]. In this regard, various approaches have been proposed to enhance the shallow network by mimicking the deep and wide model through the student-teacher strategy [26]. Training is accomplished by using state-of-the-art deep model as a teacher network with a rich information [27]. With this strategy, the shallow network can be as accurate as the deep teacher network, providing much higher performance than learning directly with ground truth data. In the proposed method, depth maps computed from the existing stereo matching network [29], which acts as a deep teacher network, are used to train the student network for monocular depth inference.

Considering that acquiring stereo images is much easier than sensing depth maps in the outdoor environment. The proposed semi-supervised learning approach uses massive unlabeled stereo images as inputs. To guarantee a scene diversity, we built up a new dataset, called DIML/CVL dataset [30], by capturing stereo image pairs in various scenes including park, brook, apartment, and so on (Fig. 3). We then generate pseudo ground truth depth maps with off-the-shelf stereo matching method fully trained with a large capacity through deep and wide CNN architecture. We are equipped with the accurate and robust stereo matching model that generates fewer errors than the monocular depth estimation (see Fig. 1 (b), (f), and (j)). Additionally, inspired by data distillation [31], we fuse multiple depth maps estimated on various scales when generating pseudo ground truth depth maps. This ensemble approach improves the accuracy of the pseudo ground truth depth maps by collecting a non-trivial knowledge beyond a single prediction. To compensate for inaccurate depth estimation due to occlusions, specular surfaces, and poor illumination, we also utilize stereo confidence maps [32] as auxiliary data. They are used to avoid inaccurate stereo depth values being utilized when training the monocular depth estimation networks. Experimental results demonstrate that the proposed semi-supervised approach outperforms state-of-the-arts for monocular depth inference and the DIML/CVL dataset [30] is complementary to existing outdoor scenes [24], [25]. More details about the dataset will be given in Section III.

The stereo teacher network is trained on the KITTI 2015 [24] with ground truth depth maps, and thus our method is a semi-supervised learning approach. To address the lack of large-scale ground truth depth maps in the monocular

depth estimation, various unsupervised learning approaches by using stereo images as inputs have been proposed [20], [33], [34]. Image alignment or reconstruction losses [20], [33] are proposed to train monocular depth estimation networks in an unsupervised fashion. They, however, often fail to handle occluded regions and obtain sharp depth boundaries. We will show that the proposed semi-supervised learning approach outperforms these unsupervised approaches in terms of both subjective and objective evaluations. The work of [21] proposed a semi-supervised learning approach that uses both the supervised loss using ground truth depth maps as well as the unsupervised reconstruction loss. However, it still faces a performance degradation when the trained model feeds an image from a novel domain as shown in Fig. 1. Contrarily, our approach does not suffer from such a domain adaptation issue on the novel domain by utilizing the pseudo ground truth depth maps intelligently through the student-teacher strategy.

Our pre-trained model for monocular depth prediction can also be used as a powerful proxy task for scene understanding tasks such as semantic segmentation and road detection. It achieves an outstanding performance over a scratch training, and is comparable to the pre-trained model with ImageNet [23], a massive manually labeled dataset.

Our main contributions are highlighted as follows.

- We propose a novel semi-supervised learning for monocular image depth estimation based on student-teacher strategy [26] with data distillation [31] and stereo confidence measure [32].
- We introduce a new RGB-D dataset, called DIML/CVL dataset, which is complementary to the existing datasets [24], [25], and validate its effectiveness through various experimental results.
- We demonstrate that the feature representation of our monocular depth estimation provides a rich knowledge for scene understanding tasks.

The remainder of this paper is organized as follows. Section II describes related works. The proposed method and our DIML/CVL dataset are presented in Section III. Extensive performance validation is then provided in Section IV, including ablation study, comparison with state-of-the-arts, and transfer learning to semantic segmentation and road detection tasks. Section IV concludes this paper.

II. RELATED WORKS

In this section, we briefly review and discuss three lines of works that are most relevant to our work.

A. Stereo Matching

The objective of stereo matching is to find a set of corresponding points between two (or more) images. The correspondence map is converted into a depth map using stereo calibration parameters. Early studies based on CNN attempted to measure the similarity between patches of two images. Han *et al.* [6] proposed a siamese network that extracts features from patches followed by the similarity measure. Zbontar and LeCun [7] computed the matching cost through CNNs and applied it to classical stereo matching pipelines consisting

of cost aggregation, depth optimization, and post-processing. Luo *et al.* [8] proposed to compute the matching cost by learning a probability distribution over all depth values and then performing an inner product between two feature maps. Note that these approaches focused on computing the matching cost through CNNs and remaining procedures for the stereo matching still rely on conventional hand-crafted approaches.

Recent approaches have attempted to predict a depth map in an end-to-end fashion, achieving a substantial performance gain. Mayer *et al.* [9] proposed a new method, named DispNet, that uses a series of convolutional layers for cost aggregation and then regresses a depth map. Pang *et al.* [29] introduced a two-stage network, called cascade residual learning (CRL), by extending the DispNet [9]. The first and second stages calculate depth maps and their multi-scale residuals, and then the outputs of both stages are combined to form a final depth map. Kendall *et al.* [10] introduced a new end-to-end approach that performs the cost aggregation using a series of 3-D convolutions. Chang *et al.* [11] incorporated a contextual information through 3-D convolutions using stacked multiple hourglass networks over cost volume. The above-mentioned stereo matching approaches can be utilized as the teacher network in our framework.

B. Monocular Depth Estimation

The performance of the monocular depth estimation has been advanced dramatically through the supervised learning approach that uses depth maps acquired from active sensors as ground truth for input images. Eigen *et al.* [17] designed a multi-scale deep network that predicts a coarse depth map and then progressively refines the depth map. Liu *et al.* [19] casted the monocular depth estimation into a continuous conditional random field (CRF) learning problem that jointly learns the unary and pairwise potentials of the CRF in a unified deep CNN framework. Luo *et al.* [22] formulated the monocular depth estimation with two sub-networks, view synthesis network and stereo matching network. They first synthesize stereo pairs from an input image, and then apply the stereo matching network to produce the depth map. Kuznetsov *et al.* [21] proposed to use both supervised and unsupervised losses with ground truth depth maps and stereo image pairs. This semi-supervised approach boosts the performance of the monocular depth estimation, but it faces a performance degradation by the domain adaptation issue, as shown in Fig. 1.

To address the lack of massive ground truth depth maps, several approaches have attempted to learn the monocular depth inference in an unsupervised manner. Garg *et al.* [33] proposed an encoder-decoder architecture to learn the network using an image alignment loss with a pairs of source and target images only. The image alignment loss is computed by warping the source image into the target image with a predicted depth map. Xie *et al.* [34] proposed a novel training scheme that synthesizes the right view from the left view. The network estimates a probabilistic map for different depth levels instead of directly regressing depth values. Using a left-right consistency constraint, Godard *et al.* [20] proposed an improved architecture for training the monocular depth estimation from stereo images.

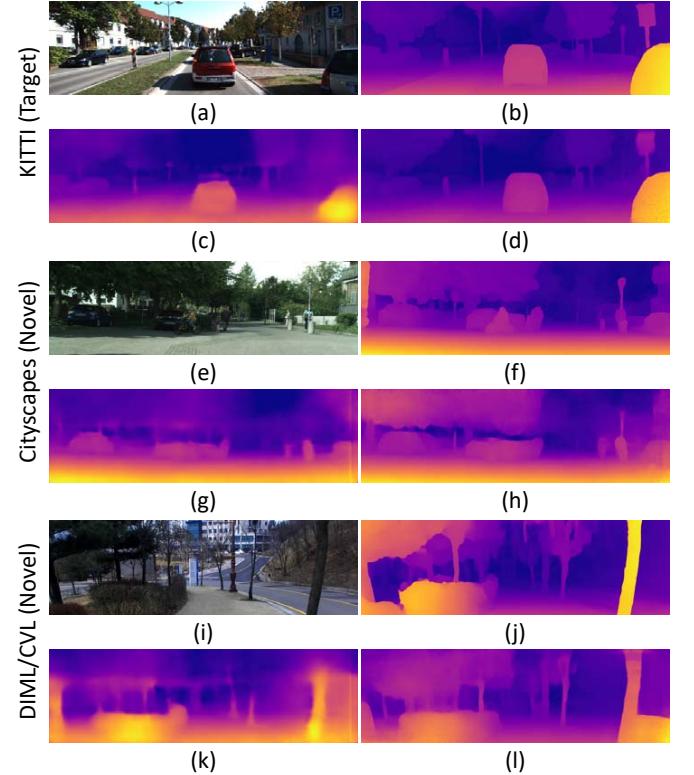


Figure 1. Sample images collected from various dataset and estimated depth maps: (a), (e), (i) input images, (b), (f), (j) depth maps predicted by deep stereo matching network [29], (c), (g), (k) depth maps estimated from state-of-the-art monocular depth estimation network [21], and (d), (h), (l) depth maps estimated from the proposed monocular depth estimation network. Note that the stereo matching method, which serves as a teacher network in our method, is less sensitive to the domain adaptation issue. The method of [21] results in blurry artifact when performing an inference at novel domain, while our semi-supervised approach achieves consistent results for both target and novel domains thanks to the student-teacher strategy.

C. Feature Learning via Pretext Task

Several approaches have attempted to leverage a pretext task as an alternative form of supervision in some applications where it is difficult to construct massive ground truth data. Doersch *et al.* [35] proposed to train the deep network by predicting a relative position between two patches randomly extracted from unlabeled images. They utilized the resultant feature representation as a proxy task for object detection and visual data mining. Noroozi and Favaro [36] proposed to solve Jigsaw puzzles for representing object parts and their spatial arrangements, and then applied it to object classification and detection tasks. Pathak *et al.* [37] proposed an unsupervised image inpainting approach that generates contents of an arbitrary image region conditioned on its surroundings. The learned encoder features are applied to object classification/detection and semantic segmentation tasks. Larsson *et al.* [38] investigated an image colorization as the proxy task in replacement of ImageNet pre-training.

The features learned with these works have been successfully transferred to high-level tasks such as classification, detection and segmentation. In our work, we demonstrate that the network pre-trained for monocular depth prediction is a powerful proxy task for learning feature representations in

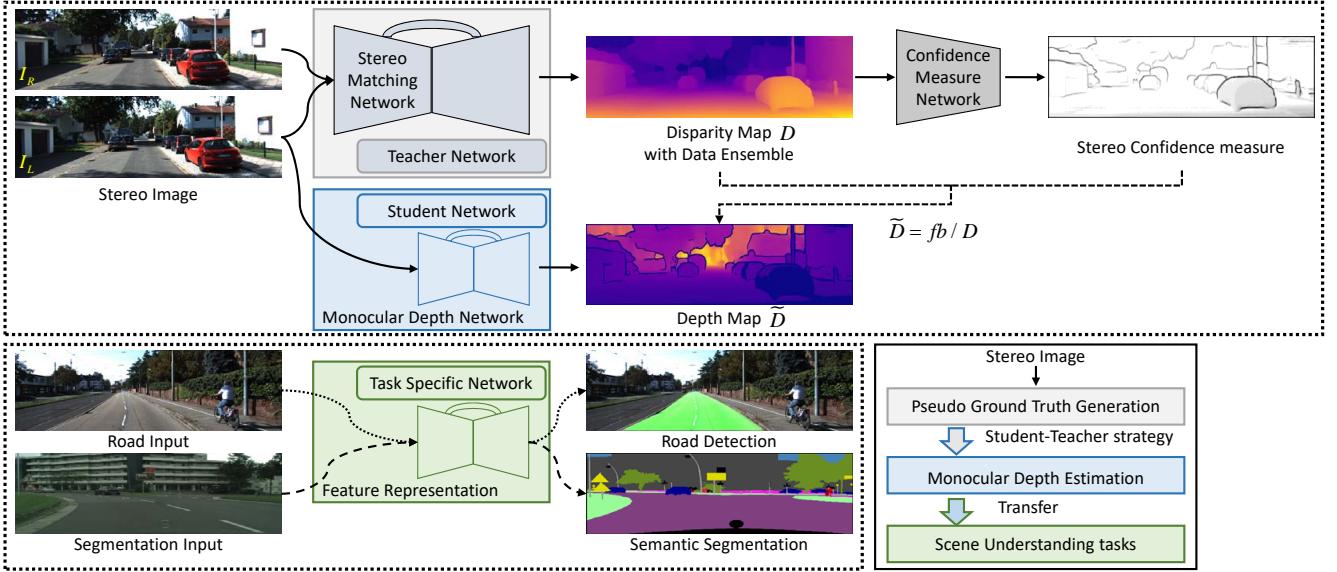


Figure 2. The overall framework of the proposed method. We first generate depth maps from stereo image pairs using the deep teacher network [29] trained with ground truth depth data. Here, we adopt the data ensemble [31] to fuse depth maps estimated on multiple scales. The stereo confidence map [32] is estimated to identify inaccurate stereo depth values. Then, the monocular student network is trained with the pseudo ground truth depth maps and stereo confidence maps. Additionally, we transfer our network parameters to scene understanding tasks such as road detection and semantic segmentation to show that our model trained for monocular depth estimation can be used as a powerful proxy task for the high-level vision tasks.

scene understanding tasks such as semantic segmentation and road detection.

III. PROPOSED METHOD

A. Motivation & Overview

Due to the lack of scene diversity, deep networks for the monocular depth estimation often face the severe domain adaptation issue. For instance, feeding a single (monocular) image from a target domain, in which the deep networks are trained, yields satisfactory results (Fig. 1(c)). Contrarily, when we test an image from a novel domain, an output depth map produces blurry edges, making it hard to distinguish objects. The monocular depth estimation network trained with the KITTI dataset [24] does not work well on the Cityscapes dataset [25], though both include driving scenes. The result for non-driving scenes of our new dataset becomes even worse. In contrast, the stereo matching network using [29] produces fine-grained depth maps on both target and novel domains as shown in Fig. 1(b), (f), and (j), even though it is trained with the KITTI dataset only.

The stereo matching aims to find similar patches from a number of candidates extracted from two images. Thus, it is enough to train the network with similar patches (positive samples) and dissimilar patches (negative samples) [7]. Though some methods propose to train the stereo matching network using two images at once in order to additionally leverage a global context on the stereo matching [11], the underlying principle is to locally explore the patch-level similarity for two-view matching. Contrarily, the monocular depth estimation, which infers a depth value from a single image by making use of monocular cues, is highly ill-posed and more challenging than the stereo matching. Thus, the global context is crucial to predict an overall 3D structure of scenes.

In this regard, the monocular depth estimation network is trained using the image and depth map, not a pair of patches extracted from them, to consider the global context. It makes the monocular depth estimation network more sensitive to the domain adaptation issue than the stereo matching network. Thus, a great variety of scenes is needed to train the monocular depth estimation network, while the stereo matching network is relatively free from such constraints.

We propose a simple yet effective semi-supervised approach for monocular depth estimation by leveraging on the student-teacher strategy. The shallow student network learns from more informative deep teacher network. Our method involves following steps. Given a number of stereo images, we generate depth maps using the deep stereo matching network trained with ground truth data, e.g., consisting of 3D LiDAR points, as a teacher network. When generating depth maps, we fuse depth maps estimated on multiple scales to provide non-trivial knowledge [31] from multiple predictions. Stereo confidence map is then estimated as auxiliary data to avoid inaccurate stereo depth values being utilized when training the monocular depth estimation network. The depth maps and stereo confidence maps are used as ‘pseudo ground truth’ for supervising the monocular student network. Experimental results will show that the pseudo ground truth data generalizes well on the novel domain. Furthermore, the monocular depth estimation induces the feature representation that improves scene understanding tasks such as semantic segmentation and road detection. The overall framework is illustrated in Fig. 2.

B. Large-scale Outdoor Stereo Dataset

Our new outdoor stereo dataset, called DIML/CVL RGB-D dataset [30], is complementary to existing RGB-D dataset such as the KITTI dataset and Cityscapes dataset. To ensure

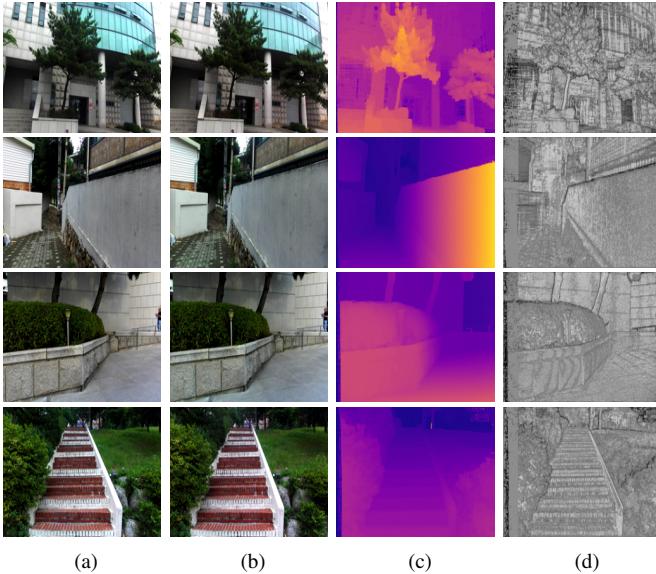


Figure 3. Samples of RGB-D pairs with stereo confidence map from DIML/CVL dataset [30]. (a) left image, (b) right image, (c) depth map using [7], and (d) stereo confidence map using [39].

the diversity of training data, we attempted to obtain non-driving scenes (e.g., park, building, apartment, trail, and street) using hand-held stereo cameras, unlike the existing dataset consisting of mostly driving scenes (e.g., road and traffic scenes) obtained from the depth sensor mounted on a vehicle. This dataset was taken from fall 2015 to summer 2017 in four different cities (Seoul, Daejeon, Cheonan, and Sejong) of South Korea.

Two types of stereo cameras, ZED stereo camera [40] and built-in stereo camera, were used with different camera configurations of baseline and focal length. The commercial ZED camera has a small baseline (12cm), so its sensing range is rather limited (up to 20m). We designed the built-in stereo system with mvBlueFox3 sensors [41] whose baseline is 40cm, increasing the maximum sensing range up to 80m. The stereo image was captured with 1920×1080 or 1280×720 resolutions. More details of our dataset are described in our technical report [42]. All scenes were taken steadily with a tripod and slider in a hand-held fashion.

It is comprised of 1 million stereo images, depth maps computed from stereo matching algorithm [7], and stereo confidence map [39], [43]. The depth maps in the DIML/CVL RGB-D dataset [30] were generated by MC-CNN [7] which was the state-of-the-art stereo matching algorithm at the time of stereo image acquisition. In our experiments, we used CRL [29], which is a more advanced stereo matching network. Of 1 million stereo image pairs, we selected 22,000 stereo images from various scenes as training data. Note that any kind of stereo matching network can be adopted to obtain depth maps. The stereo confidence map [39], [43] is used to mitigate side effects of incorrectly estimated depth values. Fig. 3 shows sample RGB-D pairs of our dataset. We believe that our outdoor dataset can promote depth related applications based on deep networks. Our dataset differs from existing ones

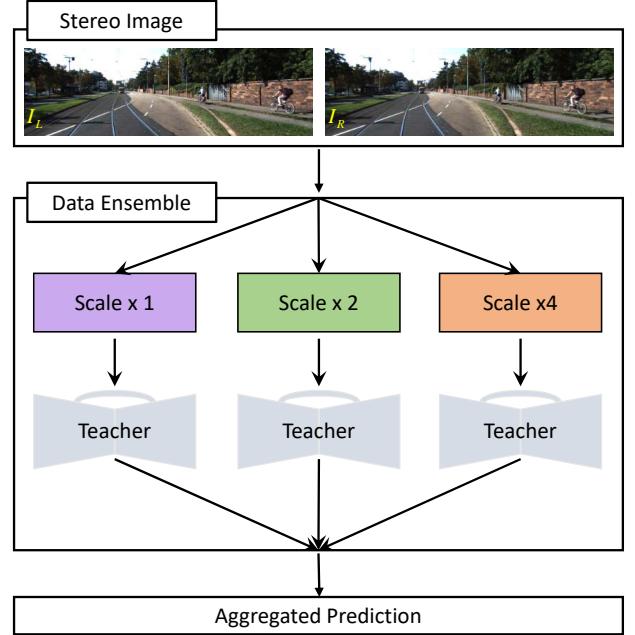


Figure 4. Data ensemble approach. We generate depth maps of stereo images in different scales via the deep teacher network, and then ensemble them on the smallest scale. The data ensemble can provide an auxiliary information from multiple predictions beyond a single prediction.

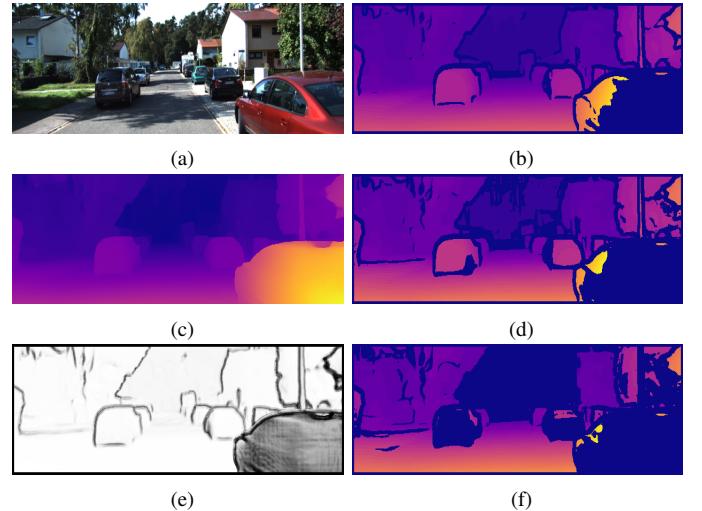


Figure 5. Pseudo ground truth data samples. (a) input image, (c) estimated depth map with data ensemble, and (e) predicted stereo confidence map. (b), (d), (f): pseudo ground truth depth maps thresholded by the stereo confidence map (e) with = 0.3, 0.55, and 0.75, respectively. Black pixels indicate unreliable pixels detected by the stereo confidence map.

in the following aspects:

- 1) It is comprised of 1 million RGB-D data for outdoor scenes.
- 2) Unlike existing outdoor datasets for driving scenes, ours was taken using hand-held stereo cameras for non-driving scenes.
- 3) Stereo confidence maps are provided together to quantify the accuracy of depth maps computed from stereo images.

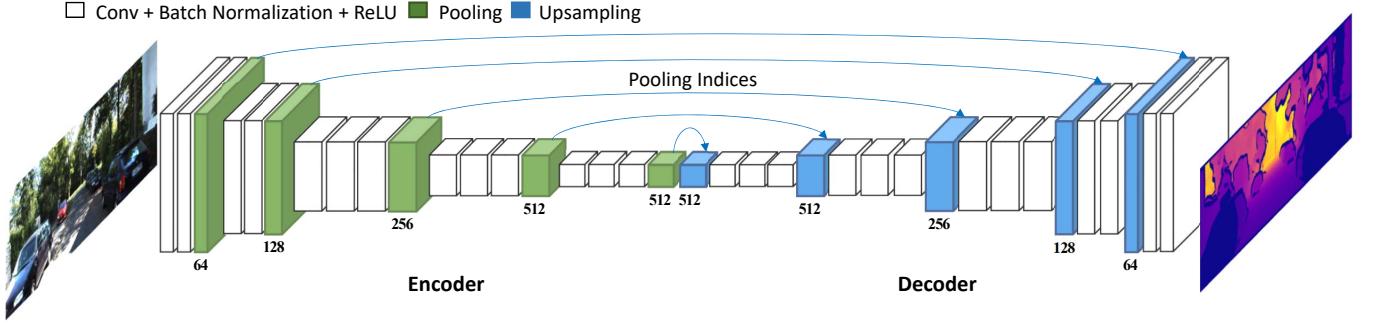


Figure 6. Monocular depth estimation network. We design a variant of U-Net architecture [44] as a baseline network. The convolution layers consist of 3×3 convolutions with batch normalization [45] and rectified linear unit (ReLU). The max-pooling with a 2×2 window and stride 2 is performed at the encoder. The indices of max locations are computed and stored during pooling. The decoder upsamples the feature maps through unpooling with the stored pooled indices and a sequence of convolution layers.

C. Pseudo Ground Truth Generation

It is difficult to generalize the monocular depth estimation to a novel scenario due to its inherent ill-posed geometric ambiguity [22], while the stereo matching network generates accurate depth maps in both target and novel domains. The stereo matching network takes stereo images (I_l, I_r) as input and outputs a depth map D aligned to the left image I_l . We adopt the cascade residual learning (CRL) [29] as the stereo teacher network. To further improve the depth map of the teacher network, we adopt an ensemble prediction method that merges output depth maps on various scales, as explained in Fig. 4. It has been shown that the data generation can be improved by applying the same model to multiple transformations (e.g., scale, rotation, and flipping) of the input and then aggregating the results [17], [23], [46], [47]. We generate depth maps on 3 different scales and we average them on the smallest scale. This improves the depth accuracy by a good margin. The RMSE(in) of single and ensemble predictions is 3.578 and 3.475 on the KITTI dataset based on Eigen split [17], respectively.

In order to alleviate the estimation error from deep stereo teacher network, we estimate the stereo confidence map, called confidence estimated with convolutional neural network (CCNN) [32]. We use the ensemble depth map to determine the degree of uncertainty. The CCNN differs from existing stereo confidence measures that need additional cues, hand-designed features [39], [48] or cost volume [43]. The CCNN extracts a square patch from the depth map and forward it to a CNN to infer a normalized confidence value $C(p) \in [0, 1]$. We denote a confidence threshold as hyper-parameter τ . The depth value of each pixel is set to be reliable when $C(p) \geq \tau$, and vice versa. By adjusting τ , we can control the sparseness and reliability of the depth map. As τ grows, unreliable areas such as textureless regions and occlusions are removed effectively, but the depth map becomes sparse. Sample images of depth maps and stereo confidence map are shown in Fig. 5. Note that since CCNN uses squares patch extracted from a depth map without using padding or stride, it assigns zero to the boundary of the confidence map. The pseudo ground truth data generated from the teacher network and confidence measure are used for supervising the student network.

D. Semi-supervised Monocular Depth Learning

We design monocular depth estimation networks based on a variant of U-Net architecture [44]. As shown in Fig 6, the encoder network consists of the first 13 convolutional layers in the VGG [49] network, similar to [50]. We discard the fully connected layers in favor of maintaining a spatial information. Each convolution layer at the encoder has the corresponding convolution layer at the decoder, and thus the decoder network also has 13 layers. The decoder upsamples the decoder feature map using the memorized max-pooling indices [50] from the corresponding encoder feature map. This sparse feature map is then densified by convolving it with a trainable decoder filter bank. Using such pooling indices boosts the performance and enables more efficient training. A final decoder output is fed to a regression loss. Specifically, given a monocular input image and pseudo ground truth depth map $\hat{D}(p)$, we use the *stereo confidence guided regression loss* \mathcal{L}_c :

$$\mathcal{L}_c = \frac{1}{\sum_p M_p} \sum_p M_p \cdot \left| \hat{D}(p) - \tilde{D}(p) \right|_1, \quad (1)$$

$$M_p = \begin{cases} 1, & \text{if } C(p) \geq \tau \\ 0, & \text{if } C(p) < \tau \end{cases}. \quad (2)$$

where $\hat{D}(p)$ denotes the depth map predicted by the monocular depth estimation network. In the experiment section, we will validate the effect of stereo confidence measure by adjusting the hyper-parameter τ .

Experimental results validate that such a simple architecture outperforms state-of-the-arts for monocular depth estimation thanks to the semi-supervised learning strategy. It is expected that more sophisticated network will further improve the depth accuracy, but we reserve this as future work since our objective is to investigate the effectiveness of the semi-supervised learning approach in the monocular depth estimation.

E. Transfer of Feature Representation

We also study the effectiveness of our pre-trained monocular depth estimation network by transferring its feature representations as a pretext task for training other similar tasks such as road detection and semantic segmentation. Experimental

results will demonstrate that our pre-trained model is comparable to the ImageNet pre-trained model that often serves as the pretext task for various vision applications [21], [50], [51]. It should be noted that extending our pre-trained model into a new domain is very easy in that only stereo image pairs are needed for supervision, different from the ImageNet where the manual supervision for an object class should be provided.

We fine-tune our pre-trained model using the KITTI road benchmark [24] for road detection and the Cityscapes [25] for semantic segmentation, respectively. Both datasets include a small amount of manually annotated training data. We transfer both encoder and decoder weights of the pre-trained model into the road detection and semantic segmentation. The softmax loss is used to fine-tune the network.

F. Implementation Details

We obtained the results of the stereo matching network (CRL) [29] by using the author-provided pre-trained model², and re-implemented the stereo confidence estimation networks (CCNN) [32] based on the author-provided code. We implemented the monocular depth estimation network of the encoder-decoder architecture using VLFeat MatConvNet library [52].

1) *Confidence measure network*: The CCNN was trained using 50 image pairs consisting of ground truth depth maps and stereo image pairs provided in the KITTI 2012 dataset. The stereo confidence estimation network is typically trained with a set of patches that are paired with the depth map and ground truth confidence value [32]. Thus, a small amount of training data is enough to train the stereo confidence measure network. The ground truth stereo confidence map is obtained by comparing an absolute difference between the predicted depth map and ground truth depth map (KITTI LiDAR points). Following the literature [24], we set the confidence value to 1 when the absolute difference is smaller than 3 pixels, and 0 otherwise.

During the training phase, we use a binary cross entropy loss after applying a sigmoid function to the output of the network. We carried out 100 training epochs with an initial learning rate of 0.001, decreased by a factor 10 every 10 epochs, and a momentum of 0.9.

2) *Monocular depth network*: For training the monocular depth estimation network, we collect 22,600, 21,283, and 22,000 images from KITTI, Cityscapes and DIML/CVL dataset, respectively. Then, the pseudo ground truth training data is generated using the stereo matching network and stereo confidence map. The monocular student network was trained for 30 epochs with a batch size 4. The Adam solver [53] was adopted for an efficient stochastic optimization with a fixed learning rate of 0.001 and momentum of 0.9. We select the model that works best on the validation dataset of the Eigen split [17].

IV. EXPERIMENTAL RESULTS

In this section, we validate the effectiveness of our semi-supervised monocular depth estimation through quantitative

and qualitative comparisons with the state-of-the-art methods in outdoor scenes. For the quantitative comparisons, we employ several metrics which have been used in prior works [17], [20]–[22]:

- Threshold: % s.t. $\max\left(\frac{d_i}{u_i}, \frac{u_i}{d_i}\right) = \delta < thr$
- Abs rel: $\frac{1}{N} \sum_i |d_i - u_i| / d_i$
- Sqr rel: $\frac{1}{N} \sum_i \|d_i - u_i\|^2 / d_i$
- RMSE(lin): $\sqrt{\frac{1}{N} \sum_i \|d_i - u_i\|^2}$
- RMSE(log): $\sqrt{\frac{1}{N} \sum_i \|\log d_i - \log u_i\|^2}$

where u_i denotes a predicted depth at pixel i , and N is a total number of pixels.

A. Dataset

We generated the pseudo ground truth depth maps using stereo images provided in the KITTI [24], Cityscapes [25], and DIML/CVL dataset. All images are resized to 620×188 for training and testing.

1) *KITTI*: This dataset consists of outdoor driving scenes with sparse depth maps captured by the Velodyne LiDAR [54]. The depth map is very sparse (less than 6% of density) and depth values are available only at the bottom parts of a color image. The dataset contains 42,382 rectified stereo pairs from 61 scenes, with a typical image being 1242×375 pixels in size. Following the Eigen split [17], we split stereo image pairs into 22,600 images for training, 888 images for validation, and 697 images for test.

2) *Cityscapes*: It was originally constructed for semantic segmentation and provides manually annotated segmentation maps for 19 semantic classes, consisting of 2,975 images for training, 500 images for validation, and 1,525 images for test. Additionally, they provide 22,973 stereo image pairs with spatial dimensions of 2048×1024 . We split stereo image pairs into 21,283 for training, 500 for validation, and 3,215 for test. We cropped the stereo images by discarding bottom parts (the car hood) of 20% and then resized them.

3) *DIML/CVL*: Our outdoor dataset consists of 1 million stereo image pairs, depth maps, and stereo confidence maps. The original spatial resolution of this dataset is 1920×1080 or 1280×720 . Of 1 million image pairs, we selected 23,500 image pairs similar to KITTI and Cityscapes. We split them into 22,000 images for training, 800 images for validation, and 700 images for test. Input images were cropped and resized by discarding bottom parts containing mostly grounds.

B. Ablation Study

1) *Impact of label quality*: We first investigated the performance gain of the data ensemble and the trade-off between accuracy and density that determine the performance of the pseudo ground truth depth maps. The KITTI dataset using the Eigen split [17] was used for experiments. The density and accuracy of pseudo ground truth depth maps are controlled by the confidence threshold τ . As shown in Fig. 7, the higher τ , the better the accuracy of pseudo ground truth depth maps. However, this reduces the density of the depth maps. For instance, when $\tau = 0.75$, only half of depth pixels is chosen as reliable.

²<https://github.com/Artifineuro/crl>

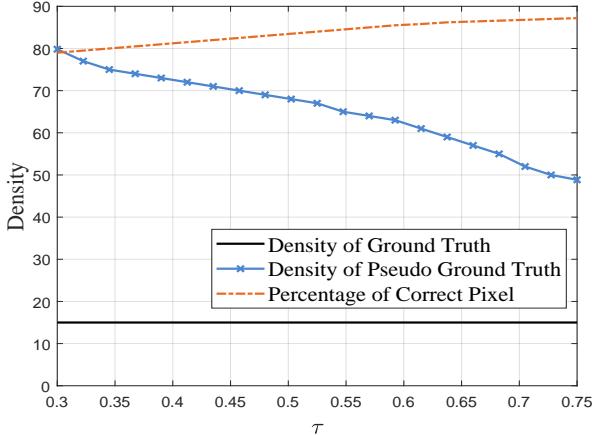


Figure 7. Quality analysis of pseudo ground truth depth maps controlled by the confidence threshold τ . The KITTI dataset [24] is used for experiments. We achieve more accurate pseudo ground truth depth map with a higher. ‘Ground truth’ indicates the sparse depth maps provided in the KITTI dataset.

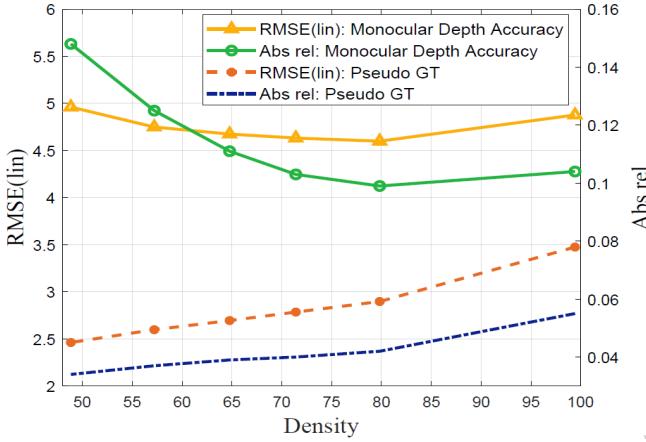


Figure 8. Trade-off analysis between the accuracy and density of pseudo ground truth depth maps. There is inevitably an estimation error in the pseudo ground truth depth maps computed from the deep stereo matching network. Thus, we use the stereo confidence maps to identify inaccurate depth values (Fig. 7) and avoid them being used, when training the monocular depth estimation network. Here, we investigate an inference accuracy of the monocular depth estimation according to the density of the pseudo ground truth depth maps. Training with more accurate pseudo ground truth depth maps (yet with a lower density) does not necessarily yield a higher accuracy in the monocular depth estimation. We achieve the best monocular depth accuracy, when the density of pseudo ground truth depth maps is about 80% with the confidence threshold = 0.3. Refer to Fig. 7 for the relationship between the density and the stereo confidence threshold τ .

To study the trade-off between the accuracy and density of pseudo ground truth depth maps, we measured the accuracy of the monocular depth network according to the confidence threshold τ . Fig. 8 shows the RMSE(lin) and Abs rel of pseudo ground truth depth maps and monocular depth maps. As mentioned above, the pseudo ground truth depth map with a lower density tends to be more accurate. However, using more accurate pseudo ground truth depth maps does not necessarily lead to a higher accuracy in the monocular depth network. Since semi-dense pseudo ground truth depth maps often have no valid depth values around object boundaries or thin objects, the monocular depth networks trained with such semi-dense depth maps may fail to recover reliable depth values around these regions.

The monocular depth accuracy of density 80% is better than that of 100%. However, when the density of the pseudo ground truth depth maps is about 72%, the monocular depth accuracy becomes worse even though the pseudo ground truth depth maps are more accurate. This indicates that there is the trade-off between the density and accuracy of the pseudo ground truth depth maps. In our experiment, the monocular depth network achieves the best accuracy when the density is about 80% ($\tau = 0.3$). We generated pseudo ground truth training data using $\tau = 0.3$ for all KITTI, Cityscapes and DIML/CVL datasets.

We studied the effectiveness of the data ensemble when generating the pseudo ground truth depth maps. The monocular depth accuracy was measured with the RMSE(lin) and Abs rel. Without the data ensemble, the RMSE(lin) and absolute relative error (Abs rel) are 4.995 and 0.109, respectively. When we applied the data ensemble, the RMSE(lin) and the absolute relative error (Abs rel) are 4.877 and 0.104, respectively. The data ensemble achieves a meaningful gain in both metrics.

2) *Impact of scene diversity:* We conducted the ablation study to demonstrate the effectiveness of the DIML/CVL dataset [30] in our semi-supervised approach. Fig. 9 shows qualitative results of training with four different combinations (5th to 8th rows) of pseudo ground truth training data in the proposed method. Similar results are obtained in both the target and novel domains when using the DIML/CVL dataset for training. This indicates that the proposed method addresses the generalization issue well in a novel domain. To be specific, using the DIML/CVL dataset for training leads to a performance gain in the novel domain, e.g. when comparing the results of the proposed method trained with KITTI (5th row) and KITTI+DIML/CVL (6th row). The proposed method consistently generates smooth depth maps with sharp edges and recovers an overall scene layout well. We also included results (2nd to 4th rows) of state-of-the-art approaches [20]–[22] for a qualitative comparison. When testing at the Cityscapes and DIML/CVL dataset (novel domain), they face a significant performance drop. Note that except for [20], it is not possible to train [22] and [21] with Cityscapes and DIML/CVL dataset providing stereo image pairs as the two methods require ground truth depth maps for training. The unsupervised approach [20] can be learned with training data from both target and novel domains, but the unsupervised loss used incurs blurry depth boundaries and inaccurate estimates in the occlusion, as shown in the results of the 2nd row in Fig. 9.

Table I reports a depth accuracy using the Eigen split [17] in the KITTI dataset. Since ground truth depth maps are provided in the KITTI dataset only, the objective evaluation was done in the KITTI dataset. Nevertheless, it is found that the proposed method based on the simple encoder-decoder architecture still outperforms state-of-the-arts with complicate network architectures even when only KITTI dataset is used for training. The depth accuracy is further improved by training the network with DIML/CVL dataset, demonstrating its complementary property. We describe more analysis about the quantitative results in the next section.

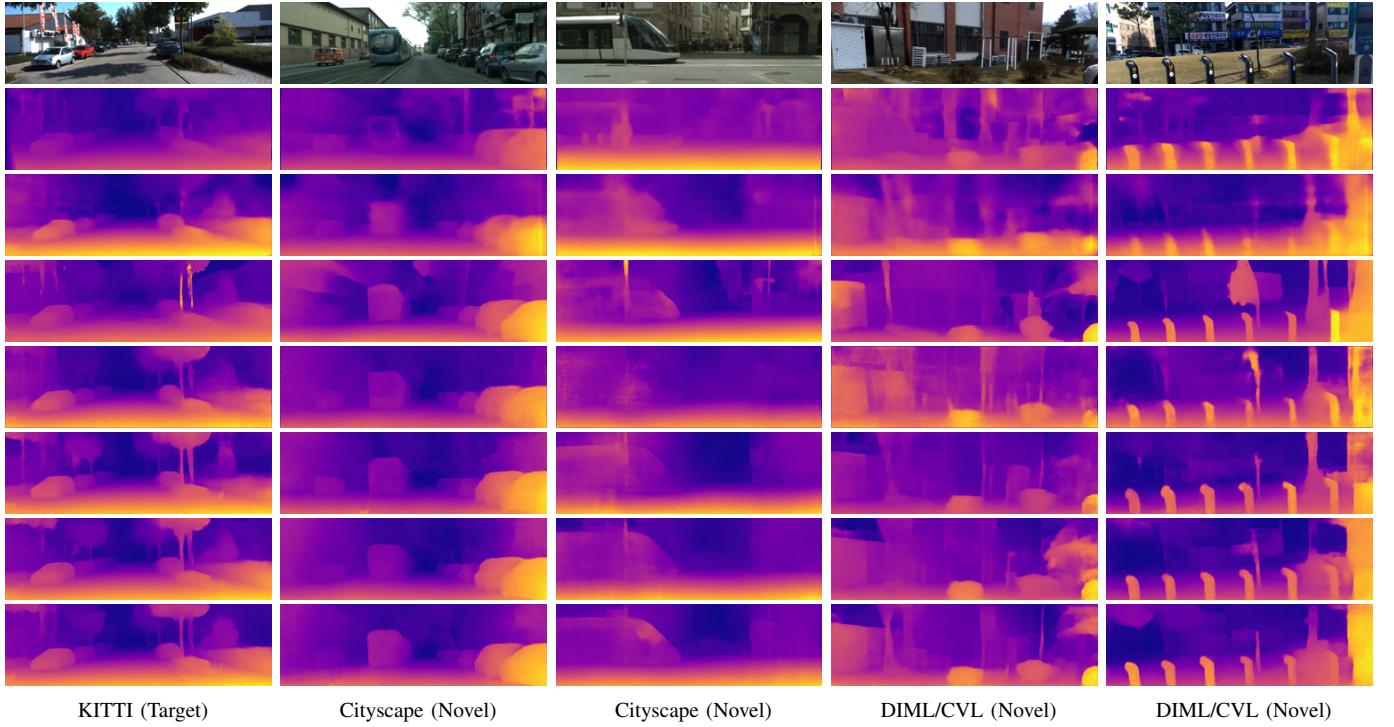


Figure 9. Impact of scene diversity (from top to bottom): input image, Godard *et al.* [20] trained with stereo image pairs of the KITTI + Cityscapes, Kuznetsov *et al.* [21] trained with stereo image pairs and ground truth depth map of KITTI, Luo *et al.* [22] trained with left image and ground truth depth map of Flying Things synthetic dataset [9], and the proposed method trained with KITTI, KITTI + DIML/CVL, KITTI + Cityscapes, and KITTI + Cityscapes + DIML/CVL, respectively. Specifically, we show qualitative results (5th to 8th rows) of training with four different combination of pseudo ground truth depth maps in the proposed method. We also included results (2nd to 4th rows) of state-of-the-art approaches [20]–[22] for a qualitative comparison. Refer to Table I to check what training data is used for the state-of-the-arts approaches.

C. Comparison with state-of-the-arts

We compare existing monocular depth estimation approaches including supervised [17], [21], unsupervised [20], and semi-supervised methods [21], [22] with the proposed approach. In Table I, our method consistently outperforms recent approaches except for $\delta < 1.25^3$. Following the literatures [20]–[22], the depth value was truncated at 80m or 50m. The proposed method was trained through various combination of KITTI, Cityscapes, and our dataset. Eigen *et al.* [17] was trained using ground truth depth maps augmented from the KITTI 2015 dataset. Godard *et al.* [20] proposed an unsupervised approach where stereo images of the KITTI and/or Cityscapes dataset were used. Though this method requires no ground truth depth maps during training, it is difficult to handle an occlusion and obtain a sharp depth boundary due to the limitation of the image reconstruction loss. Kuznetsov *et al.* [21] employed both supervised regression loss and unsupervised reconstruction loss [20], achieving a performance gain over existing supervised and unsupervised approaches [17], [20]. However, it still requires ground truth depth maps as supervision. Luo *et al.* [22] proposed to use view synthesis network and stereo matching network in an unified framework. However, their model is a supervised approach and was trained with synthetic FlyingThings3D data [9], and thus incurs domain adaptation gaps between synthetic and realistic images and can not generalizes well on real data.

Qualitative results were also provided in Fig. 10. Our

monocular depth network achieves more accurate and edge-preserved depth maps than state-of-the-arts [20]–[22]. Note that our method was trained from scratch, while Kuznetsov *et al.* [21] and Luo *et al.* [22] adopted the pre-trained model from ImageNet. It is also expected that our semi-supervised approach may produce more accurate depth maps when adopting more sophisticated networks.

D. Transfer to high-level tasks

To investigate the applicability of our model trained for monocular depth prediction, we transfer the network parameters to scene understanding tasks such as semantic segmentation and road detection.

1) *Semantic Segmentation:* We adopted the Cityscapes [25] dataset for training and evaluation. We validate methods with the mean intersection-over-union (IoU) that computes a mean value over all classes including background. Experiments were conducted with images of half-resolution for a fast computation. Following literatures [23], [47], we augmented training data with random scaling, random cropping and horizontal flipping. The network was trained for 300 epochs with a batch size of 4. We used Adam solver [53] for training with the weight decay of 0.0005 and initial learning rate of 0.0001, which decreases by a factor of 10 every 10 epoches.

Table II and Fig. 11 report quantitative and qualitative evaluation results including the scratch learning, ImageNet pre-trained model [49], and ours. All results were obtained

Table I

QUANTITATIVE EVALUATION OF MONOCULAR DEPTH ESTIMATION ON THE EIGEN SPLIT [17] OF KITTI [24] DATASET. *Sup.*, *Unsup.*, AND *Semi-sup.* DENOTE SUPERVISED, UNSUPERVISED, AND SEMI-SUPERVISED LEARNING APPROACHES, RESPECTIVELY. GT REPRESENTS GROUND TRUTH DEPTH MAPS. *pp* DENOTES A POST PROCESSING APPLIED TO THE NETWORK OUTPUT [20]. FOR DATASET, K = KITTI, CS = CITYSCAPES, AND OURS = DIML/CVL.

Method	Training data	Approach	Dataset	RMSE(lin)	RMSE(log)	Abs rel	Sqr rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Lower is better										
cap 80m										
Eigen <i>et al.</i> [17]	Left + LiDAR	<i>Sup.</i>	K	7.156	0.270	0.215	1.515	0.692	0.899	0.967
Godard <i>et al.</i> [20]	Stereo	<i>Unsup.</i>	K	5.927	0.247	0.148	1.344	0.803	0.922	0.964
Godard <i>et al.</i> + <i>pp</i> [20]	Stereo	<i>UnSup.</i>	K + CS	4.935	0.206	0.114	0.898	0.861	0.949	0.976
Kuznetsov <i>et al.</i> [21]	Left + LiDAR	<i>Sup.</i>	K	4.815	0.194	0.122	0.763	0.845	0.957	0.987
Kuznetsov <i>et al.</i> [21]	Stereo + LiDAR	<i>Semi-sup</i>	K	4.621	0.189	0.113	0.741	0.862	0.960	0.986
Luo <i>et al.</i> [22]	(Synthetic) Stereo + GT	<i>Sup.</i>	K	4.681	0.200	0.102	0.700	0.872	0.954	0.978
cap 50m										
Garg <i>et al.</i> [33]	Stereo	<i>Unsup.</i>	K	5.104	0.273	0.169	1.080	0.740	0.904	0.962
Godard <i>et al.</i> [20]	Stereo	<i>Unsup.</i>	K	4.471	0.232	0.140	0.976	0.818	0.931	0.969
Godard <i>et al.</i> + <i>pp</i> [20]	Stereo	<i>Unsup.</i>	K + CS	3.729	0.194	0.108	0.657	0.873	0.954	0.979
Kuznetsov <i>et al.</i> [21]	Stereo + LiDAR	<i>Semi-sup</i>	K	3.518	0.179	0.108	0.595	0.875	0.964	0.988
Luo <i>et al.</i> [22]	(Synthetic) Stereo + GT	<i>Sup.</i>	K	3.503	0.187	0.097	0.539	0.885	0.960	0.981
Our Method	Left + Pseudo GT	<i>Semi-sup</i>	K + CS + Ours	3.162	0.175	0.095	0.613	0.884	0.964	0.986

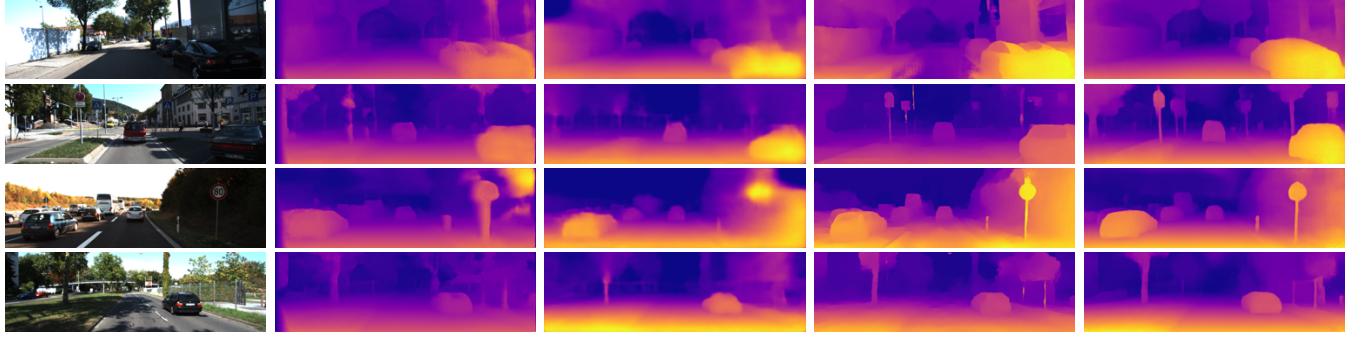


Figure 10. Qualitative results on the Eigen split [17] of KITTI dataset [24]: (a) input image, (b) Godard *et al.* [20] trained with stereo image pairs of the KITTI + Cityscapes, (c) Kuznetsov *et al.* [21] trained with stereo image pairs and ground truth depth map of KITTI, (d) Luo *et al.* [22] trained with left image and ground truth depth map of Flying Things synthetic dataset [9], and (e) the proposed method trained with KITTI + Cityscapes + DIML/CVL dataset. Our approach produces depth maps better aligned with input images and recovers complicated objects such as trees, poles, and traffic sign very well. Refer to Table II to check what training data is used for existing approaches.

with the same encoder-decoder architecture used in the monocular depth estimation. Our pre-trained model significantly outperforms the model learned from scratch, and is comparable to the pre-trained model with ImageNet, which is a massive manually labeled dataset. It is also shown in Table II that the more accurate the monocular depth network, the higher IoU in the semantic segmentation.

Note that the experiments intend to show that our monocular depth network based on the simple encoder-decoder architecture is a powerful proxy task for the semantic segmentation. Though our network is inferior to state-of-the-arts in the semantic segmentation benchmark [55], it is expected that the accuracy can be improved by using more sophisticated deep architectures.

2) *Road Detection*: We investigate the effectiveness of our pre-trained model for road detection in the KITTI road benchmark [24] that provides 289 training images with annotated

ground truth data and 290 test images. It is divided into three categories: single lane road with markings (UM), single-lane road without markings (UU), and multi-lane road with markings (UMM). Following literatures [56], the training data was augmented through various transformations. We randomly scale the image by a factor between 0.7 and 1.4 and perform color augmentation by adding value -0.1 and 0.1 to the hue channel of the HSV space. For an efficient stochastic optimization, we use the Adam optimizer [53], fixed learning rate of 0.0001 and weight decay of 0.0005. The road detection network was trained for 40 epochs with a batch size of 4.

For quantitative comparison, we measure both maximum F1-measurement (Fmax) and average precision (AP). Table III shows that our pre-trained model consistently outperforms both the model learned from scratch and ImageNet pre-trained model [49]. Similar to the semantic segmentation, the accuracy of road detection is correlated with that of the monocular

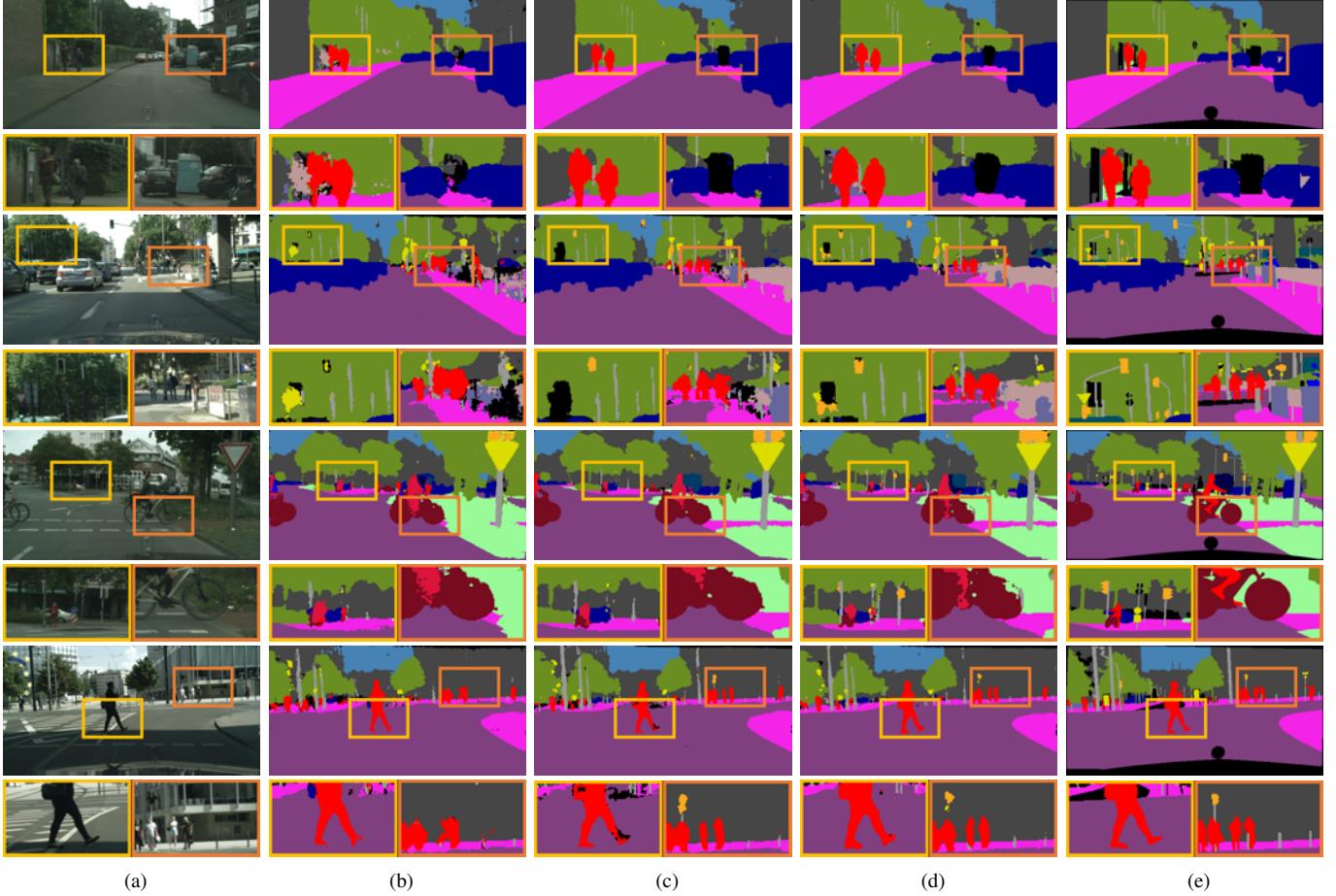


Figure 11. Semantic segmentation results on the Cityscapes dataset: (a) input images, (b) ~ (d) results of fine-tuning with different initialization methods. (b) scratch, (c) ImageNet pre-trained model [49], (d) our pre-trained model, and (e) ground truth annotations. The results show a clear benefit of our pre-trained model for semantic segmentation.

Table II

QUANTITATIVE COMPARISON OF THE PRE-TRAINED MODEL USING THE SAME NETWORK ARCHITECTURE FOR THE CITYSCAPES BENCHMARK. THE HIGHER THE MEAN IOU, THE BETTER.

Semantic Segmentation		
Initialization	Pretext	mean IoU
Scratch	-	52.27
ImageNet pre-trained model [49]	Classification	66.27
K	Depth	62.82
K + Ours	Depth	64.54
K + CS	Depth	65.02
K + CS + Ours	Depth	65.47

depth estimation network used as the pretext. Interestingly, the proposed method based on the simple encoder-decoder architecture outperforms state-of-the-art methods using complicate network architecture including Oliveria *et al.* [56] ($F_{max} = 93.83$, $AP = 90.47$) and Teichmann *et al.* [51] ($F_{max} = 94.88$, $AP = 93.71$). Fig. 12 shows that our method distinguishes between roads and sidewalks very well.

V. CONCLUSION

In this paper, we propose a novel and effective semi-supervised approach for monocular depth estimation. We adopt the student-teacher strategy where a shallow student network is trained by leveraging deep and accurate teacher network. With massive stereo image pairs consisting of diverse outdoor scenes provided in the DIML/CVL dataset, we use the deep stereo matching network as a teacher network to generate pseudo ground truth depth maps. To improve the depth map of the teacher network, we apply the data ensemble and stereo confidence measure. As a student network, the monocular depth network is trained with the pseudo ground truth depth maps and stereo confidence measures. We verified through extensive experiments that the proposed semi-supervised approach is free from the domain adaptation issue and achieves state-of-the-art performance. Additionally, we show that our model trained for the monocular depth estimation provides semantically meaningful feature representations for scene understanding tasks. It is easy to extend our semi-supervised approach into a new domain in that only stereo image pairs are needed as a supervision. We expect that the proposed method serves as a key component in addressing the domain adaptation issue in various vision tasks.

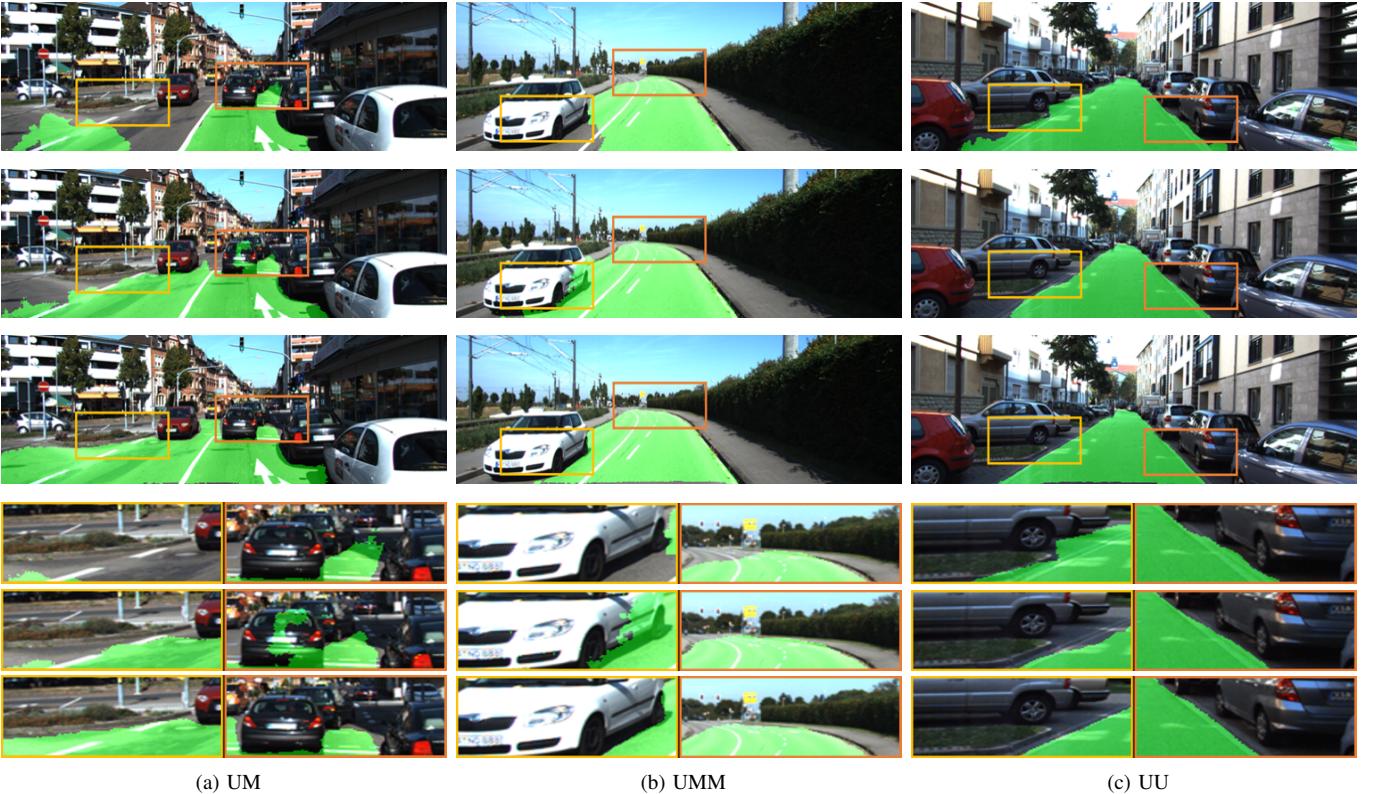


Figure 12. Road detection results on the KITTI dataset for different scene categories: (from top to bottom) results learned from scratch, ImageNet pre-trained model [49], and our pre-trained model. Corresponding enlarged parts of boxes are shown together. The road detection results using our pre-trained model show the effectiveness in distinguishing road and sidewalk.

Table III

QUANTITATIVE COMPARISON OF THE PRE-TRAINED MODEL USING THE SAME NETWORK ARCHITECTURE FOR THE KITTI ROAD BENCHMARK. THE HIGHER THE FMAX AND AP, THE BETTER.

Road Detection			
Initialization	Pretext	Fmax	AP
Scratch	-	93.82	90.87
ImageNet pre-trained model [49]	Classification	94.28	92.25
K	Depth	94.41	92.04
K + Ours	Depth	94.92	92.28
K + CS	Depth	95.12	93.09
K + CS + Ours	Depth	95.65	94.46

REFERENCES

- [1] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen, “Dense 3d object reconstruction from a single depth view,” in *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.
- [2] J. Cho, Y. Kim, H. Jung, C. Oh, J. Youn, and K. Sohn, “Multi-task self-supervised visual representation learning for monocular road segmentation,” in *IEEE Int. Conf. on Multi. and Expo.*, 2018.
- [3] S. Kim, K. Park, K. Sohn, and S. Lin, “Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields,” in *Eur. Conf. on Comput. Vis.*, 2016, pp. 143–159.
- [4] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, “Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation,” in *IEEE Int. J. Comput. Vis.*, vol. 112, no. 2, 2015, pp. 133–149.
- [5] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” in *IEEE Int. J. Comput. Vis.*, vol. 47, 2002, pp. 7–42.
- [6] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, “Matchnet: Unifying feature and metric learning for patch-based matching,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2015, pp. 3279–3286.
- [7] J. Zbontar and Y. LeCun, “Computing the stereo matching cost with a convolutional neural network,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2015, pp. 1592–1599.
- [8] W. Luo, A. G. Schwing, and R. Urtasun, “Efficient deep learning for stereo matching,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2016, pp. 5695–5703.
- [9] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, “A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2016, pp. 4040–4048.
- [10] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, “End-to-end learning of geometry and context for deep stereo regression,” in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 65–75.
- [11] J.-R. Chang and Y.-S. Chen, “Pyramid stereo matching network,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2016, pp. 5410–5418.
- [12] W. Zhuo, M. Salzmann, X. He, and M. Liu, “Indoor scene structure analysis for single image depth estimation,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2015, pp. 614–622.
- [13] D.-C. Lee, M. Hebert, and T. Kanade, “Geometric reasoning for single image structure recovery,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2009.
- [14] D. Hoiem, A. A. Efros, and M. Hebert, “Geometric context from a single image,” in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 65–75.
- [15] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, “Shape-from-shading: A survey,” in *IEEE Trans. Pattern Anal. Mach. Intell.*, 1999.
- [16] Y. Kim, H. Jung, D. Min, and K. Sohn, “Deep monocular depth estimation via integration of global and local predictions,” in *IEEE Trans. Image Process.*, vol. 27, no. 8, 2018, pp. 4131–4144.
- [17] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Proc. Advances in Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.
- [18] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation

- and support inference from rgbd images,” in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.
- [19] F. Liu, C. Shen, and G. Lin, “Deep convolutional neural fields for depth estimation from a single image,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2015.
- [20] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2017.
- [21] Y. Kuznetsov, J. Stuckler, and B. Leibe, “Semi-supervised deep learning for monocular depth map prediction,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2017.
- [22] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, “Single view stereo matching,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2018, pp. 155–163.
- [23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” in *IEEE Int. J. Comput. Vis.*, vol. 115, 2015, pp. 211–252.
- [24] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2012.
- [25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2016.
- [26] Z. Xu, Y.-C. Hsu, and J. Huang, “Training student networks for acceleration with conditional adversarial networks,” in *Proc. Brit. Mach. Vis. Conf.*, 2018.
- [27] L.-J. Bi and R. Caruana, “Do deep nets really need to be deep,” in *Proc. Advances in Neural Inf. Process. Syst.*, 2014, pp. 2654–2662.
- [28] G. Hinton, O. Vinyal, and J. Dean, “Distilling the knowledge in a neural network,” in *arXiv*, 2015.
- [29] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, “Cascade residual learning: A two-stage convolutional neural network for stereo matching,” in *Proc. Int. Conf. Comput. Vis. Work. on Geometry Meets Deep Learning*, 2017.
- [30] “<https://dimlrgbd.github.io/>”
- [31] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, “Data distillation: Towards omni-supervised learning,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2018.
- [32] M. Poggi and S. Mattoccia, “Learning from scratch a confidence measure,” in *Proc. Brit. Mach. Vis. Conf.*, 2016.
- [33] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, “Unsupervised cnn for single view depth estimation: Geometry to the rescue,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 740–756.
- [34] J. Xie, R. Girshick, and A. Farhadi, “Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 842–857.
- [35] C. Doersch, A. Gupta, and A.-A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 1422–1430.
- [36] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.
- [37] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, “Context encoders: Feature learning by inpainting,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2016.
- [38] G. Larsson, M. Maire, and G. Shakhnarovich, “Colorization as a proxy task for visual understanding,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2017.
- [39] M. Park and K. Yoon, “Leveraging stereo matching with learning-based confidence measures,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2015.
- [40] “Stereolabs. zed camera. url <https://www.stereolabs.com/>”
- [41] “<https://www.matrix-vision.com/usb3-vision-camera-mvbluefox3.html>”
- [42] “https://dimlrgbd.github.io/downloads/technical_report.pdf”
- [43] S. Kim, D. Min, B. Ham, S. Kim, and K. Sohn, “Deep stereo confidence prediction for depth estimation,” in *Proc. IEEE Conf. Image. Process.*, 2017.
- [44] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Int. Conf. on Medical Image Computing and computer-assisted intervention.*, 2015.
- [45] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 448–456.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2016.
- [47] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, 2015.
- [48] S. Kim, D. Min, S. Kim, and K. Sohn, “Feature augmentation for learning confidence measure in stereo matching,” in *IEEE Trans. Image Process.*, vol. 26, no. 12, 2017, pp. 2019–6033.
- [49] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition?” in *arXiv*, 2014.
- [50] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, 2017, pp. 2481–2495.
- [51] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, “Multinet: Real-time joint semantic reasoning for autonomous driving,” *arXiv preprint arXiv:1612.07695*, 2016.
- [52] A. Vedaldi, K. Lenc, and A. Gupta, “Matconvnet: Convolutional neural networks for matlab,” in *Proc. ACM Int. Conf. Multimedia*, pp. 689–692, 2015.
- [53] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv.
- [54] “Velodyne. accessed: Feb. 15, 2017. [online]. available: <http://velodynelidar.com/>”
- [55] “<https://www.cityscapes-dataset.com/benchmarks/#scene-labeling-task>.”
- [56] G. L. Oliveira, W. Burgard, and T. Brox, “Efficient deep methods for monocular road segmentation,” in *IEEE/RSJ Int. Conf. on Intell. Robot. and Syst.*, 2016.