

Towards Scene Understanding: Unsupervised Monocular Depth Estimation with Semantic-aware Representation

Po-Yi Chen^{1,3}Alexander H. Liu^{1,3}Yen-Cheng Liu²Yu-Chiang Frank Wang^{1,3}¹National Taiwan University²Georgia Institute of Technology³MOST Joint Research Center for AI Technology and All Vista Healthcare

pychen0@ntu.edu.tw, r07922013@ntu.edu.tw, ycliu@gatech.edu, ycwang@ntu.edu.tw

Abstract

Monocular depth estimation is a challenging task in scene understanding, with the goal to acquire the geometric properties of 3D space from 2D images. Due to the lack of RGB-depth image pairs, unsupervised learning methods aim at deriving depth information with alternative supervision such as stereo pairs. However, most existing works fail to model the geometric structure of objects, which generally results from considering pixel-level objective functions during training. In this paper, we propose SceneNet to overcome this limitation with the aid of semantic understanding from segmentation. Moreover, our proposed model is able to perform region-aware depth estimation by enforcing semantics consistency between stereo pairs. In our experiments, we qualitatively and quantitatively verify the effectiveness and robustness of our model, which produces favorable results against the state-of-the-art approaches do.

Figure 1: Integrating depth estimation and semantic segmentation towards scene understanding. With image representation jointly learned from the above objectives preserving geometric/semantic information, unsupervised depth estimation can be realized.

1. Introduction

With the development of robotics and autonomous driving, scene understanding has become a crucial yet challenging problem. One goal of scene understanding is to recognize and analyze 3D geometric information from a 2D scene image. Toward this end, several methods [5, 14, 12] attempt to estimate depth information from a monocular image by learning a supervised regression model with a great amount of 2D-3D image pairs or multiple observations from different viewpoints. However, as most supervised learning methods, collecting ground truth data is costly and time-consuming. Thus, recent works attempted to learn unsupervised depth estimation models based on either stereo image pairs [8] or video sequences [27].

Most unsupervised depth estimation methods derive

depth information by reconstructing the geometric structure of a scene, while in addition to the geometric cue, we human estimate depth information according to semantic information of a scene. For example, we know that pixels labeled as “sky” must accompany with large values of depth. Furthermore, the depth values of the pixels within a segmentation mask (i.e., an object) should be close and relative, and significant changes of depth between adjacent pixels implicitly indicate the boundary of an object. Based on these properties, several works [13, 20, 4, 18] have explored to mutually positive transfer between semantic segmentation and depth estimation, while the requirement of pairwise depth and semantic labels limits the applicability of these models.

In this paper, we first point out the current state-of-the-arts like [8] predict the disparity maps for stereo views only based on one monocular image. This results in unawareness of structural information from the other view in the inference stage and further affects the performance of disparity prediction. With the proposed *SceneNet*, the mismatching problem can be significantly alleviated by our training

Indicates equal contribution.

strategy. We will verify our design is more reasonable by comparing the performances with the state-of-the-art unsupervised depth estimation.

More importantly, our model further achieves improved depth estimation by leveraging semantic understanding. Fig. 1 illustrates the idea of SceneNet to learn semantic-aware scene representation to advance our depth estimation. SceneNet is an encoder-decoder based network that takes scene images and encodes them into representations. The decoder acts as a multi-task yet shared classifier that transforms scene representation into the prediction of depth or segmentation. This is accomplished by a unique *task identity* mechanism, which allows the shared decoder to switch the outputs between semantic segmentation and depth estimation. Based on the conditioned task identity information, SceneNet thus can be viewed as a cross-modal network model, bonding depth and segmentation modalities together. To further strengthen the bonding between geometric and semantic understanding, we introduce *left-right semantic consistency* and *semantics-guided disparity smoothness*, two self-supervised objective functions that refine depth estimation with semantic prediction.

In our experiments, we demonstrate that SceneNet not only produces satisfactory results on depth estimation, its integration of geometric and semantic information also realizes general scene understanding. With a small amount of data with annotated semantic ground truth labels, our model gains significant improvement over depth estimation.

We highlight the contributions of our work as follow:

- We point out possible mismatch problems in recent unsupervised monocular depth estimation methods utilizing left-right consistency.
- Our proposed SceneNet work towards scene understanding by integrating both geometric and semantic information, with our proposed modules preserving task identity, left-right semantic consistency and semantics-guided disparity smoothness.
- The end-to-end learning procedure allows our model to learn from disjoint cross-modal datasets of stereo images and semantically labeled images.
- In our experiments, we qualitatively and quantitatively verify the effectiveness and robustness of our model over state-of-the-art methods on benchmark datasets.

2. Related Work

Depth Estimation

Generally, depth information can be represented in an absolute depth value or a disparity value (the former is inversely proportional to the latter). Traditional methods re-

lied on additional observations such as multi-view from several cameras [21] and motion cue from video frames [9] to derive the corresponding depth of a scene. With only a single monocular image during the inference stage, Liu *et al.* [14] used a deep convolution neural network and continuous condition random field as patch-wise depth predictor to estimate the depth information. Eigen *et al.* [5] incorporated the coarse and fine cues to predict the depth map. With sparse ground-truth depth map, Kuznietsov *et al.* [12] learned to predict the dense depth map in a semi-supervised manner. Although promising results were reported, their requirement of a large amount of pixel-level annotation and lack of ability in handling noisy depth sensory data would be concerns.

On the other hand, unsupervised depth estimation methods rely on the supervision from either stereo image pairs [6, 8, 25] or video sequences [27, 19, 16, 23, 25]. With the stereo images in the training stage, Garg *et al.* [6] applied the inverse warping loss to learn a monocular depth estimation CNN. Godard *et al.* [8] inferred the disparities by warping the left-viewpoint image to match the right-viewpoint one (and vice versa) with a left-right consistency objective function. As noted previously, the derived disparities map could later be converted into the depth map. On the other hand, some works [27, 19] explored image sequences and proposed the temporal photometric warp loss between the adjacent frames to derive the depth information. Mahjourian *et al.* [16] similarly used temporal consistency and further imposed more 3D geometric constraints. Yin *et al.* [23] learned depth information together with optical flow and camera pose by taking advantage of the nature of 3D scene geometry. Zhan *et al.* [25] further proposed spatial and temporal warp objective function for learning the depth map using both temporal and stereo views.

Leveraging Semantic Segmentation

Since monocular depth estimation methods rely heavily on the property of perspective geometry or annotated ground truth, seeking assistance from semantic segmentation of image has been an inevitable direction of research. Prior works [13, 20, 4, 18] explored the possibility to combine supervised depth estimation and semantic segmentation with multi-task learning. Either through a hierarchical network, multi-stage training or sharing latent feature, they all found the two tasks are indeed strongly correlated and mutually beneficial. Jiao *et al.* [10] studied the long tail property of the distribution of depth and improved supervised depth estimation with attention and semantic segmentation. Zhang *et al.* [26] proposed a joint task-recursive learning framework to recursively refine the results of both semantic segmentation and supervised depth estimation through serialized task-level interactions. Chen *et al.* [2] proposed a self-supervised proxy task predicting

Figure 2: Architecture of our proposed SceneNet. SceneNet takes an image I as input and encodes it into a scene representation z . This representation can be decoded into the output \tilde{Y} along with the introduced task identity layer t . Based on the conditioned t , \tilde{Y} can later be transformed into pixel-wise prediction of semantic segmentation output s or depth estimation output d , while these two outputs would be properly aligned based on the corresponding semantic information.

relative depth for urban scenes, which can then be adapted to semantic segmentation or depth estimation through fine-tuning model (with ground truth provided).

While these prior works were closely related to ours in terms of pursuing a more general scene understanding for cross depth estimation and segmentation, we state the difference between our work and the previous works as following: First, unlike the aforementioned works, we choose to build a unified model to jointly exploit both tasks. Second, our method does not require paired training data to learn shared scene representation for depth estimation and semantic segmentation (i.e., training data of these two tasks can be completely disjoint). Third, depth estimation remains unsupervised with our proposed model, we do not use any given disparity map or sparse ground truth. Last, while learning shared representations for different downstream tasks, our approach remains end-to-end trainable. Neither pre-training nor fine-tuning model is required.

3. Proposed Method

The goal of our proposed model, *SceneNet*, is to predict a dense depth map directly from a monocular image. During training, our model is trained on stereo pairs and RGB-segmentation pairs. Unlike existing multi-task learning models like [13, 20, 4, 18], our model does not require the stereo images and semantic-annotated images to be paired.

As illustrated in Fig. 2, the encoder of our model first converts a scene image I into a scene representation z . Our decoder further takes both the scene representation z and

a task identity t (detailed in Sect. 3.1) as input, and outputs the cross-modal prediction \tilde{Y} . To train *SceneNet*, we apply the objective functions for unsupervised depth prediction and supervised semantic segmentation in Sect. 3.2. Later in Sect. 3.3, we refine the cross-modal prediction by introducing two self-supervised objective functions – left-right semantic consistency and semantic grounded disparity smoothness. In Sect. 3.4, we summarize the learning objective and detail the inference procedure of SceneNet.

3.1. Task Identity for Cross-modal Prediction

Most existing works that jointly learn disparity estimation and semantic segmentation use task-specific classification/regression sub-networks to obtain disparity maps and segmentation masks. However, hyper-parameters such as the number of sharing/non-sharing layers across different branches are required to be tuned and decided according to the task shift. This restricts the practicality of the model, especially when adapting to different datasets.

To address the limitation, we merge cross-modal predictions by utilizing a unified decoder conditioned on a task identity t (as shown in Fig. 2). In practice, we set the task of disparity estimation as $t = 1$ and task of semantic segmentation as $t = 0$. Our decoder further generates the cross-modal prediction \tilde{Y} from the scene representation z and the task identity t :

$$\tilde{Y} = D((z, t)), \quad (1)$$

where \parallel is a operation of concatenation and D is our cross-modal decoder with no activation function in last layer.

Specifically, the semantic segmentation prediction s (red

lines in Fig. 2) is computed as:

$$s = \text{softmax}(\tilde{Y}_s), \quad (2)$$

where $\tilde{Y}_s = D(z, t = 1)$ and softmax is a softmax function. The disparity map prediction d (green lines in Fig 2) is derived as:

$$d = \text{sigmoid}(f_\mu(\tilde{Y}_d)), \quad (3)$$

where $\tilde{Y}_d = D(z, t = 0)$, f_μ refers to pixel-wise average pooling and sigmoid is the sigmoid function.

Note that since \tilde{Y} is conditioned on the task identity t , our model is able to arbitrarily switch the output between different tasks by assigning a different value to t . We note that the use of a unified decoder allows sharing geometric and semantic information across different modalities and contributes positive transfer for both tasks. We would later verify the effectiveness of this unified decoder in our experiments.

3.2. Depth Estimation & Semantic Segmentation

Unsupervised Depth Estimation Inspired by existing unsupervised models on depth estimation [6, 8], we utilize the stereo image pairs I^l, I^r as supervision during training in order to derive a disparity map from a monocular image in inference stage.

Given an RGB monocular image, our model predicts a pixel-wise disparity map, which is used to warp an image from one viewpoint to another. To be more specific, we input left-view image I^l and predicts its corresponding disparity map d^l , which is applied to warp the right-view image I^r and reconstruct the left-view image $I^{r \rightarrow l}$.

To learn our disparity prediction model, we compute the image reconstruction loss L_{re} with element-wise L1 loss:

$$L_{re} = \|I^l - I^{r \rightarrow l}\|_1 + \|I^r - I^{l \rightarrow r}\|_1. \quad (4)$$

where $I^{r \rightarrow l}$ is obtained from warping the right image I^r based on the left-view disparity d^l .

To further match the consistency between right and left disparities and maintain the smoothness of predicted disparity maps, we apply the left-right disparity consistency loss and disparity smoothness loss introduced by Godard *et al.* [8]. Thus, our entire objective function for learning depth estimation can be defined as:

$$L_{depth} = L_{re} + \lambda_r \|d^l - d^{r \rightarrow l}\|_1 + \lambda_d \|d^r - d^{l \rightarrow r}\|_1 + \lambda_{ds} \sum_x d^l e^{-\alpha d^l} + \lambda_{ys} \sum_y d^r e^{-\alpha d^r}, \quad (5)$$

where λ_r and λ_{ds} are the weights for the associated terms. Note that $d^{r \rightarrow l}$ can be obtained by warping right-view disparity d^r according to left-view disparity d^l (similar remarks can be applied to $d^{l \rightarrow r}$).

Figure 3: Model design differences between [8] and ours. Note that [8] predicts both disparity maps d^l and d^r given only the input left-view image I^l , causing d^r to align with I^l instead of I^r , the mismatching problem therefore arises. We predict a disparity map given the input image, and advance the same warping techniques to preserve left-right prediction consistency via image flipping. This not only avoids possible mismatch but also simplifies the learning process.

The Mismatching Problem It is worth noting that Godard *et al.* [8] predicts both disparity maps d^l and d^r from one input image I^l as shown in Fig. 3. We show that this might not properly maintain the structural alignment between the right-view RGB image I^r and the right-view disparity map d^r . This is because that, without the structural and textural information of right-view image I^r , it would be difficult to accurately estimate the right-view disparity d^r from a single left-view image I^l .

Instead of predicting both disparity maps from a single view, we choose to output only one disparity map which aligns with the input image. To obtain the right disparity map d^r , we horizontally flip the right-view image I^r .

Supervised Semantic Segmentation Existing depth estimation methods generally focus on pixel-wise disparity estimation [6, 8, 25] and regard all pixels within an image as spatial homogeneity, which would lead to unfavorable disparity estimation along object boundaries. To overcome the limitation, we perform disparity estimation by leveraging semantics information from segmentation-image pairs. We thus define the semantic segmentation loss L_{seg} as:

$$L_{seg} = H(s_{gt}, s), \quad (6)$$

where H indicates the cross-entropy loss and s_{gt} denotes the ground truth labels from additional disjoint dataset.

3.3. Self-supervised Learning of SceneNet

To reinforce the semantic awareness when estimating disparity, we further introduce two self-supervised regularization losses, *left-right semantic consistency* and *semantics-guided disparity smoothness*.

Left-Right Semantic Consistency In Sect. 3.2, we consider the left-right consistency loss between RGB stereo image pairs. However, such consistency over the color value of each pixel is likely to be affected by optical changes between left-right views. For instance, specular reflection on a glass would vary across different viewpoints. To mitigate the problem, we further observe such left-right consistency at the semantic level, since semantic segmentation is less sensitive to optical changes.

By replacing stereo image I^r and I^l in (4) into their semantic segmentation S^r and S^l , the left-right semantic consistency can be defined as:

$$L_{lrsc} = \|S^l - S^{r \cdot l}\| + \|S^r - S^{l \cdot r}\|, \quad (7)$$

where $S^{r \cdot l}$ can be obtained by warping S^r according to d^l and we follow the same rule to obtain $S^{l \cdot r}$.

Semantics-Guided Disparity Smoothness In addition to left-right semantic consistency, we also regularize the smoothness of disparity values within each segmentation mask. This semantics-guided disparity smoothness is defined as:

$$L_{smooth} = \|d - f(d) \cdot (1 - (S - f(S)))\|, \quad (8)$$

where \cdot is the operation which sets the maximize value along each channel as 1 and sets the remaining values as 0, \odot denotes element-wise multiplication, and f is the operation of shifting input one pixel along the horizontal axis. The second term is similar to applying an edge detector to identify edges of segmentation masks. Note that the smoothness loss is also calculated along the vertical axis, but here we omit it in (8) for simplicity.

3.4. Learning of SceneNet

During training, SceneNet takes either single view image with semantic label or stereo view image as input. The full objective of SceneNet can be defined as

$$L = L_{depth} + \omega_{seg} L_{seg} + \omega_{lrsc} L_{lrsc} + \omega_{smooth} L_{smooth}, \quad (9)$$

where ω_{seg} , ω_{lrsc} and ω_{smooth} are the weights for each loss. During the inference stage, SceneNet takes a monocular image to produce both semantic segmentation and disparity map (which can then be transformed into depth as specified in [8]) by manipulating the task identity.

4. Experiments

In order to quantitatively and qualitatively evaluate our model and to fairly compare with recent works, we train our SceneNet on the stereo image pairs from the KITTI dataset [7]. As for learning semantic segmentation ability, we use the fully annotated images of the Cityscapes dataset [3]. Note that we do not require any images to have both stereo image pairs and the ground truth semantic segmentation map. The detail of datasets used in our experiments are given as follow:

Eigen Split Eigen et al. [5] selected 697 images from the KITTI dataset [7] as test set for single view depth estimation. To fairly compare with the prior works, we followed their setting to use 22,600 images for training and the rest for evaluation.

KITTI Split To further recognize the scene understanding ability of SceneNet, we also evaluate our method on the KITTI split of KITTI dataset following the work of Godard et al. [8]. The training set of KITTI split contains 29,000 image pairs from various scenes and 200 images for the test set. Moreover, the test set not only provides ground truth disparities, but also comes along with ground truth semantic segmentation labels, which are consistent with the annotations used in the Cityscapes Dataset. Although no semantic annotation from KITTI split is utilized during training, it allows us to evaluate both depth prediction and semantic segmentation abilities of our model on the test set.

Cityscapes Dataset The Cityscapes Dataset [3] provides images of urban street scenes that is paired with pixel-wise segmentation masks. This dataset is used as our only segmentation data for training SceneNet. The provided training set contains 2,975 images and the corresponding ground truth semantic labels. Note that the amount of training data we used to train SceneNet for semantic segmentation is about 10 times less than the amount used for depth. As for evaluation, the testing set contains 500 annotated images. To understand the scene as much as possible, SceneNet uses up to 19 semantic classes, which are commonly shared among segmentation works.

4.1. Implementation Details

Network Architecture Our proposed SceneNet is composed of a pair of encoder and decoder modified from DispNet [17]. As dilated residual networks (DRNs) [24] has shown promising results, our encoder utilizes dilated Resnet layers to obtain better scene understanding features. With the encoder extracts scene representations from the input image, the task identity t will be appended to these features. Depending on which task is performed, the values of the layer are either all 1s or 0s. These features will later be decoded by our decoder, which is inspired by Godard et al. [8]. Our decoder uses four skip connections [15] from the encoder to enhance the resolution of our predictions.

Table 1: Quantitative results of depth estimation on the Eigen split of KITTI dataset. Following previous works, we conduct experiments capped at 80/50 meters in depth.

Method	cap	(Lower is better)				(Higher is better)		
		Abs Rel	Sq Rel	RMSE	RMSE log	< 1.25	< 1.25 ²	< 1.25 ³
Zhou <i>et al.</i> [27]	80m	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yang <i>et al.</i> [22]		0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian <i>et al.</i> [16]		0.163	1.240	6.220	0.250	0.762	0.916	0.968
Yin <i>et al.</i> [23]		0.155	1.296	5.857	0.233	0.793	0.931	0.973
Garg <i>et al.</i> [6]		0.152	1.226	5.849	0.246	0.784	0.921	0.967
Zou <i>et al.</i> [28]		0.150	1.124	5.507	0.223	0.806	0.933	0.973
Godard <i>et al.</i> [8]		0.141	1.186	5.677	0.238	0.809	0.928	0.969
Zhan <i>et al.</i> [25]		0.135	1.132	5.585	0.229	0.820	0.933	0.971
Ours (w/o seg)		0.128	0.996	5.444	0.226	0.820	0.936	0.972
Ours		0.118	0.905	5.096	0.211	0.839	0.945	0.977
Zhou <i>et al.</i> [27]	50m	0.201	1.391	5.181	0.264	0.696	0.900	0.966
Garg <i>et al.</i> [6]		0.169	1.080	5.104	0.273	0.740	0.904	0.962
Yin <i>et al.</i> [23]		0.147	0.936	4.348	0.218	0.810	0.941	0.977
Godard <i>et al.</i> [8]		0.134	0.872	4.305	0.224	0.824	0.937	0.973
Zhan <i>et al.</i> [25]		0.128	0.815	4.204	0.216	0.835	0.941	0.975
Ours (w/o seg)		0.122	0.742	4.103	0.212	0.835	0.944	0.977
Ours		0.112	0.673	3.871	0.198	0.852	0.951	0.980

Also, the outputs of our model are in four different scales ($1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$) of the input image size. The outputs with lower resolution are only used for loss calculation. We also adopt the exponential linear units (ELU) interlaced with the convolutional layers within our model except for the prediction layers. At last, the predicted outputs will be sent through either pixel-wise average pooling or softmax depending on the task identity t . As a reference, the proposed SceneNet contains about 15 million trainable parameters. A more detailed description of our model is available in the supplementary material.

Training Details We implement the proposed model using the TensorFlow framework [1]. During training, we resize the input images to a resolution of 256×512 . Data augmentation is also performed to avoid overfitting. To be more specific, we perform the augmentation (with a fifty percent chance) by sampling three numbers from uniform distributions in ranges of $[0.8, 1.2]$, $[0.5, 2.0]$ and $[0.8, 1.2]$ respectively. The sampled numbers will be used to shift gammas, brightness and three channels of RGB colors respectively. Our SceneNet is optimized by Adam [11], with the initial learning rate $\eta = 1e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-5$. The weights for different terms in the objective function are set as $w_{lr} = 0.2$, $w_{ds} = 0.02$, $w_{seg} = 0.1$, $w_{smooth} = 0.2$ and $w_{smooth} = 2.0$. Since our self-supervising losses rely on the quality of both depth estimation and semantic segmentation, we only apply them after both L_{depth} and L_{seg} start converging. The training procedure requires 32

hours on a single GTX 1080 GPU to train on a total of 22 thousand paired images and 2,975 annotated images for 20 epochs with batch size set as 4. At the inference stage, we input image I and obtain both d and s by changing the values in the task identity t . We also input the horizontally flipped I and obtain the flipped outputs d', s' . By flipping back the outputs we obtain d, s that aligns with the original predictions d, s . For disparity maps, we follow similar post-processing technique of Godard *et al.* [8]. As for segmentation maps, we simply take the average of s and s' as our final results.

4.2. Quantitative Results

We first evaluate our model on the testing set of Eigen split as shown in Table 1. Notice that even without semantic segmentation data, our SceneNet surpasses the state-of-the-art unsupervised depth estimation. Later in our ablation study, we verify that this is achieved by addressing the mismatching problem as noted in Sect. 3.2.

With the auxiliary semantic annotated data (2,975 images from Cityscapes), a significant improvement made by SceneNet can be observed. For more detailed studies on varying the volume of training data and evaluation over different semantic classes, please refer to the appendix.

4.3. Ablation Study

To verify the impact of each idea we proposed and each decision we made, we perform ablation studies and list the results in Table 2, with examples shown in Figures 4 and 5.

Table 2: Ablation study of our model on the KITTI dataset. The baseline model is SceneNet with separate decoders for each task. Note that K denotes stereo images from KITTI, and CS denotes semantically annotated images from Cityscapes. We have HF indicate the use of the proposed horizontally flipping technique to address the mismatching problem, U for the use of unified classifier, t as task identity, L_{sc} for left-right semantic consistency, and L_{ss} for semantics-guided disparity smoothness. In addition to depth estimation, semantic segmentation results in terms of mean Intersection-Over-Union (mIOU) on both KITTI and Cityscapes are presented.

Method	Data		Improvement					Depth		(Higher is better) < 1.25 / 1.25 ² / 1.25 ³	Segmentation (Higher is better) mIOU	
	K	CS	HF	U	t	L_{sc}	L_{ss}	(Lower is better)	(Higher is better)		K	CS
								Rel	RMSE			
								Abs / Sq	raw / log			
Godard <i>et al.</i> ^y [8]								0.117 / 1.177 0.114 / 1.086	5.804 / 0.206 5.776 / 0.204	0.848 / 0.943 / 0.977 0.849 / 0.944 / 0.977	-	-
Baseline								0.116 / 1.145 0.112 / 1.111	5.762 / 0.208 5.812 / 0.204	0.843 / 0.941 / 0.977 0.848 / 0.941 / 0.977	-	-
								-	-	-	33.83%	41.36%
								0.112 / 0.999	5.564 / 0.197	0.854 / 0.944 / 0.979	5.45%	47.44%
Ours								0.111 / 1.216 0.104 / 0.913	5.585 / 0.197 5.286 / 0.185	0.855 / 0.945 / 0.979 0.862 / 0.953 / 0.983	14.93%	46.81%
								0.104 / 0.940 0.104 / 0.913	5.340 / 0.187 5.276 / 0.187	0.863 / 0.952 / 0.982 0.861 / 0.953 / 0.983	39.13% 38.49%	48.39% 47.81%
								0.102 / 0.890	5.203 / 0.183	0.863 / 0.955 / 0.984	37.69%	47.87%

^y Results are better than those reported in the cited paper since we applied the post-processing method [8] for the sake of fairness.

(a) Input Image

(b) Baseline Disparity Map

(a) Input Image

(b) Baseline Semantic Map

(c) SceneNet Semantic Map

(d) SceneNet Disparity Map

(c) SceneNet Disparity Map

(d) SceneNet Semantic Map

Figure 4: Ablation study on depth estimation, the baseline model is SceneNet trained w/o semantic segmentation. For the same input image, we can observe that SceneNet is able produces better depth map ((b).v.s.(d)) with the aid of its semantic understanding (as demonstrated in (c)), especially for the traffic light in the figure.

Figure 5: Ablation study on semantic segmentation, the baseline model is SceneNet trained w/o depth estimation. Even though SceneNet targeted on depth estimation, improvement over semantic segmentation can still be observed ((b).v.s.(d)) with the aid of geometric understanding, especially for the vehicles in the figure.

The study is performed on the KITTI split, this allows us to evaluate the performance of SceneNet on semantic segmentation. The baseline model shares the encoder architecture with SceneNet but uses a separate decoder for each task.

We first evaluate the contribution and effectiveness of addressing the mismatching problem with the horizontally flipping (HF) technique (as noted in Sect. 3.2 and Fig. 3).

More specifically, we apply the exact network architecture of Godard *et al.* [8] with and without using our proposed HF technique; we also additionally evaluate our architecture without HF (and without segmentation either). We see that HF successfully addressed the mismatching problem with satisfactory performances, making our method state-of-the-art on monocular depth estimation.

Figure 6: Example results of SceneNet on KITTI split. Leveraging semantic understanding, our model is able to provide clear and disparity map on smaller objects such as traffic signs and trunk. As the results showed, SceneNet successfully derived a robust scene representation for depth estimation and semantic segmentation.

Next, it is clear that employing the unified classifier with task identity not only improves the performance of depth estimation, but also enables our model to predict satisfying semantic segmentation on the stereo dataset. This verified our assumption that using separate classifier limits the capacity to jointly learn from both tasks. Note that with the separately trained classifier for each task, the semantic classifier failed to produce acceptable results on images from stereo datasets, indicating that it may be overfitting the segmentation dataset without learning robust scene representation. It is also worth noting that, although our goal is to advance unsupervised depth estimation with semantic segmentation, results show the fact that segmentation performance also benefits from depth estimation by sharing information through SceneNet. Finally, with each component of loss functions being gradually added to our architecture, the full version of our SceneNet is obtained and compared to others in Table 1.

4.4. Visualization

Examples of the depth and semantic prediction are provided in Fig. 6, along with the corresponding input and ground truth from KITTI Split test set. It is apparent that our SceneNet not only performs favorable results of disparity prediction, but also provides satisfying quality of semantic

masks in complicated scenes. At last, it is worth mentioning that we do not require any ground truths of KITTI dataset. Due to the space limit, please refer to the appendix for more qualitative results and comparison against prior works.

5. Conclusion

In this paper, we propose SceneNet to address the mismatching problem of existing unsupervised depth estimation models. Our model advances depth estimation by leveraging semantic segmentation, with our proposed task identity enables SceneNet to perform both semantic segmentation and depth estimation with a unified structure. In addition, our self-supervised regularization (left-right semantic consistency and semantics-guided disparity smoothness) further allows performance improvements via semantic understanding. Our SceneNet can be trained in an end-to-end manner without ground truth depths map and any pre-trained models; moreover, it can be learned from disjoint stereo pairs and segmentation datasets without the requirement of paired training instances. In our experiments, our model performed favorably against state-of-the-art methods on unsupervised depth estimation.

Acknowledgement. This work is supported in part by the Ministry of Science and Technology of Taiwan under grant MOST 108-2634-F-002-018.

References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, 2016. **6**
- [2] W. Chen and J. Deng. Learning single-image depth from videos using quality assessment networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. **2**
- [3] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. **5**
- [4] D. Eigen and R. 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 (ICCV)*, 2015. **1, 2, 3**
- [5] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. **1, 2, 5**
- [6] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. **2, 4, 6**
- [7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. **5**
- [8] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. **1, 2, 4, 5, 6, 7**
- [9] H. Ha, S. Im, J. Park, H.-G. Jeon, and I. So Kweon. High-quality depth from uncalibrated small motion clip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. **2**
- [10] J. Jiao, Y. Cao, Y. Song, R. W. H. Lau, and R. Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. **2**
- [11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. **6**
- [12] Y. Kuznetsov, J. Stuckler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. **1, 2**
- [13] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. **1, 2, 3**
- [14] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. **1, 2**
- [15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. **5**
- [16] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. **2, 6**
- [17] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. **5**
- [18] A. Mousavian, H. Pirsaviash, and J. Kořecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *International Conference on 3D Vision (3DV)*, 2016. **1, 2, 3**
- [19] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. **2**
- [20] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. **1, 2, 3**
- [21] T.-C. Wang, M. Srikanth, and R. Ramamoorthi. Depth from semi-calibrated stereo and defocus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. **2**
- [22] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. *arXiv preprint arXiv:1711.03665*, 2017. **6**
- [23] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. **2, 6**
- [24] F. Yu, V. Koltun, and T. A. Funkhouser. Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. **5**
- [25] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. **2, 4, 6**
- [26] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. **2**
- [27] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. **1, 2, 6**
- [28] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. **6**