Pertanika J. Sci. & Technol. 28 (1): 163 - 177 (2020)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

DFRNets: Unsupervised Monocular Depth Estimation Using a Siamese Architecture for Disparity Refinement

John Paul Tan Yusiong[1,2]* and Prospero Clara Naval, Jr.[1]

[1]Computer Vision and Machine Intelligence Group, Department of Computer Science, College of Engineering, University of the Philippines, Diliman, Quezon City, Philippines
[2]Division of Natural Sciences and Mathematics, University of the Philippines Visayas Tacloban College, Tacloban City, Leyte, Philippines

ABSTRACT

Monocular depth estimation is gaining much interest in the computer vision community because it has broad applications in autonomous driving systems, robotics, and scene understanding. Significant progress has been made in solving the monocular depth estimation problem using deep learning techniques. Unsupervised learning methods are particularly appealing since the problem can be treated as an image reconstruction task, thereby forgoing the need for ground-truth depths. This paper presents an unsupervised approach to training convolutional neural networks for monocular depth estimation by introducing a novel architecture called DFRNets. DFRNets shares weight parameters between the image reconstruction sub-network and the disparity refinement sub-network and adopts a multi-scale structure for disparity predictions. The proposed method computes dense disparity maps directly from monocular images and refines them in an end-to-end fashion to reduce visual artifacts and blurred boundaries, thereby improving the method's overall performance. Experiment results using the KITTI test set showed that the proposed method outperformed many state-of-the-art methods, since it achieved the best performance on the two distance ranges: 0–80 meters and 1–50 meters. Moreover, the

ARTICLE INFO

Article history:
Received: 27 July 2019
Accepted: 15 November 2019
Published: 13 January 2020

E-mail addresses:
jtyusiong@up.edu.ph (John Paul Tan Yusiong)
pcnaval@dcs.upd.edu.ph (Prospero Clara Naval, Jr.)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

qualitative results revealed that the method generated more detailed and accurate depth maps of the scenes, with no border artifacts around the image boundary.

Keywords: Disparity refinement, monocular depth estimation, siamese architecture, unsupervised learning methods

INTRODUCTION

Depth estimation is a fundamental problem in computer vision, with various applications in autonomous driving systems, robotics, and scene understanding. The problem of estimating depth from a single image is ill-posed and inherently ambiguous, and as a result, a variety of methods to solve it have been proposed (Cadena et al., 2016; Liu et al., 2014; Saxena et al., 2005; Saxena et al., 2008). With the rapid development of deep learning methods and the availability of large training datasets, the performance of depth estimation models has improved significantly. Recently, there has been a growing interest in solving the monocular depth estimation problem using deep learning methods (Eigen et al., 2014; Garg et al., 2016; Godard et al., 2017; Yusiong & Naval, 2019; Zhou et al., 2017) since these methods combine local and global contexts to automatically infer a depth map from a single image.

Existing monocular depth estimation methods can be divided into two categories: supervised and unsupervised. Supervised learning methods require many training data with ground-truth depths since models must be trained using these ground-truth depths (Eigen et al., 2014). However, such training data may not always be available since it is quite challenging and expensive to collect numerous and diverse training data with ground-truth depths from different real-world scenarios; these ground-truth depths must also be carefully aligned and calibrated. Unsupervised learning methods overcome this limitation by training models to infer depth; this is accomplished by minimizing the photometric loss using a warping-based view synthesis procedure, thereby forgoing the need for ground-truth depths. The unsupervised methods can be further sub-divided into two groups based on the training data used: methods that employ monocular video sequences (Zhou et al., 2017) and methods that use only rectified stereo images (Garg et al., 2016; Godard et al., 2017; Yusiong & Naval, 2019).

Unsupervised learning methods also have certain limitations. As shown in the work of Zhou et al. (2017), training a model from monocular video sequences lowers the quality of depth predictions at test time. Also, in addition to estimating depth, the model requires a separate pose network to determine the ego-motion between temporal image pairs. It also requires the intrinsic camera parameters and the video frames as inputs during training. Moreover, models trained on monocular video sequences must address a scene's motion or depth–speed ambiguity and occluded regions. For the latter, occlusion masks must be

integrated into the loss function to indicate the valid pixel coordinates when computing the training loss (Mahjourian et al., 2018; Zhou et al., 2017). In contrast to training on monocular video sequences, with rectified stereo images (Garg et al., 2016; Godard et al., 2017; Yusiong & Naval, 2019), the model requires only the images as inputs during training, and it can achieve promising results even though the predicted depth maps have visual artifacts and blurred boundaries. These visual artifacts and blurred boundaries are due to occlusions, since some parts of the scene are not visible given a fixed camera baseline. To resolve these issues and improve the model's overall accuracy, mechanisms to handle occlusions are necessary. One such mechanism is the introduction of a post-processing step to refine the predicted disparity maps, but this decouples the final disparity maps from the training (Godard et al., 2017).

This research is another step toward solving the monocular depth estimation problem using the unsupervised learning method. First, this paper addresses the issue of decoupling depth estimation from disparity refinement by presenting a deep network that is trained using only rectified stereo images but can predict and refine a disparity map from a single image simultaneously and in an end-to-end manner. The proposed approach transforms the idea of a post-processing step into a trainable component of the model presented here so that it can perform depth estimation and refinement simultaneously, unlike in previous models (Garg et al., 2016; Godard et al., 2017; Yusiong & Naval, 2019), which can only perform depth estimation. The proposed method employs a novel Siamese architecture called DFRNets that has two autoencoders. These autoencoders generate high-quality disparity maps by sharing weight parameters between the image reconstruction sub-network and the disparity refinement sub-network, and they adopt a multi-scale structure for disparity predictions. In essence, training a model with this method requires performing a forward pass using the proposed Siamese architecture and inputting the original images to the image reconstruction sub-network and the horizontally flipped images to the disparity refinement sub-network. The predicted disparity maps are fused with a pixel-wise mean operation, while image boundaries are handled in a manner similar to that used by Godard et al. (2017). Next, the DFRNets is trained to jointly perform learning and refinement of depth maps in an end-to-end manner by reformulating an existing training loss function that was initially designed for depth estimation only. This paper presents a comprehensive evaluation of the proposed method using the challenging KITTI 2015 driving dataset, and experiment results show that, with the proposed Siamese architecture, the model achieves state-of-the-art results in an unsupervised setting, both quantitatively and qualitatively. The proposed unsupervised framework generates better disparity maps than other frameworks by converting the post-processing step into a trainable component of the model. It does this by training the model to simultaneously perform depth estimation and disparity refinement in an end-to-end manner using rectified stereo images only. This work is the first of its kind

to use a Siamese network consisting of two autoencoders that share weight parameters to handle the unsupervised monocular depth estimation problem by training the model to learn two related tasks jointly. The sample predictions in Figure 1 reveal that the proposed method effectively recovers scene structures such as street symbols. The main contributions of this work are the following:

1. It introduces an unsupervised learning framework that can jointly perform learning and refinement of depth maps in an end-to-end manner, thereby transforming the idea of a post-processing step into a trainable component of the model. This framework deviates from the usual approach, which involves training a model for depth estimation only.
2. It employs a novel Siamese architecture called DFRNets to simultaneously perform depth estimation and refinement of depth maps. This architecture consists of two autoencoders and shares weight parameters between the image reconstruction sub-network and the disparity refinement sub-network.
3. It reformulates an existing training loss function for joint learning and refinement of depth maps even though it was originally designed for depth estimation only.
4. It demonstrates the effectiveness of the proposed method using the KITTI 2015 driving dataset and compares the results against existing state-of-the-art methods in unsupervised monocular depth estimation, both quantitatively and qualitatively.

METHODOLOGY

This section describes the proposed method in detail. Essentially, the proposed method involves learning to simultaneously predict and refine disparity maps in an unsupervised manner, that is, in an end-to-end manner with only rectified stereo images as inputs. Figure 2 provides an overview of the framework and its components. At the core of this method is a Siamese network architecture called DFRNets, which consists of two autoencoders

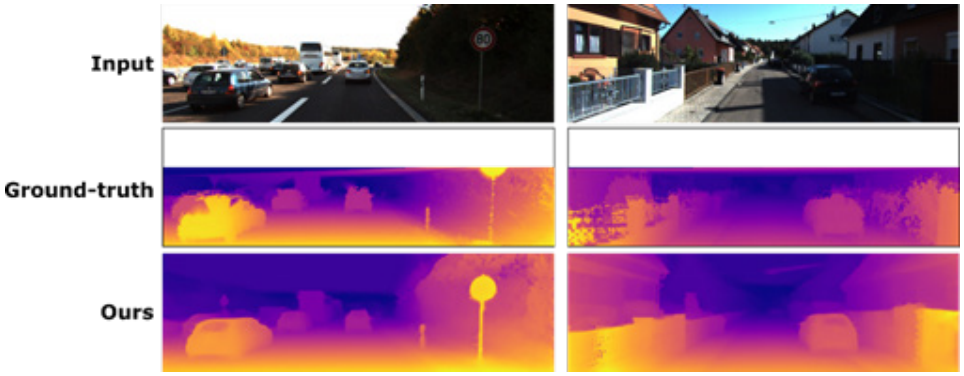

Figure 1. Example predictions generated by the proposed method using the KITTI 2015 test set. Top to bottom: input left image, ground-truth depth map, and the proposed method's prediction.

that share weight parameters. More precisely, each sub-network of the DFRNets is an autoencoder; one handles the image reconstruction task while the other handles the visual artifacts and blurred boundaries to refine the predicted disparity maps.

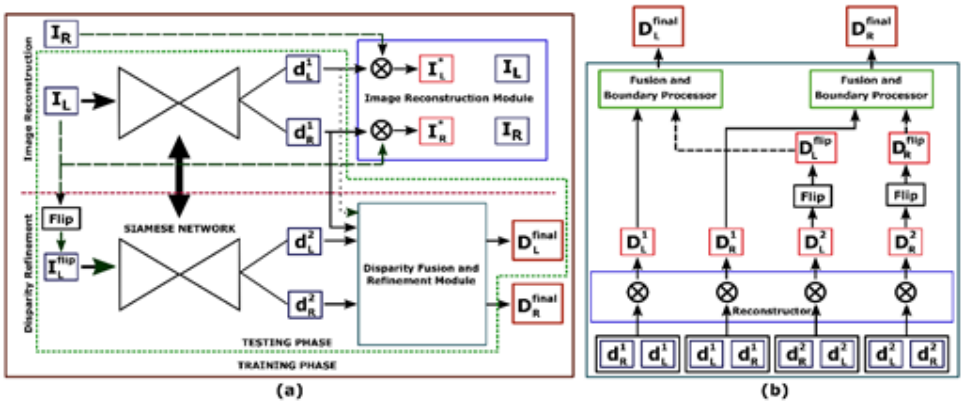

Figure 2. Network architecture: (a) DFRNets, (b) **d**isparity **f**usion and **r**efinement module

Network Architecture

The proposed framework adopts a Siamese network using the AsiANet model (Yusiong & Naval, 2019) as the autoencoder, but the autoencoder can be any architecture that can produce a disparity map. The Siamese architecture consists of two sub-networks: the image reconstruction sub-network and the disparity refinement sub-network. These two sub-networks share weight parameters to perform two different tasks simultaneously: predicting disparity maps and refining the predicted disparity maps. The image reconstruction sub-network receives the original left images I_L as inputs, while the disparity refinement sub-network receives the horizontally flipped left input images I_{flipL}. The sub-networks produce two pairs of disparity maps (d_{L1}, d_{R1}) and (d_{L2}, d_{R2}) for each scale, respectively.

Image Reconstruction Module. The main objective of training the Siamese network is to minimize image reconstruction errors between the input image I and the reconstructed image I^*; therefore, the image reconstruction sub-network contains a module that transforms the disparity maps d_{L1} and d_{R1} and the images I_R and I_L to reconstruct I^*_L and I^*_R at each scale using the sampler from the spatial transformer network (Jaderberg et al., 2015) that performs bilinear interpolation. Essentially, the module accepts two pairs of inputs (I_L, d_{R1}) and (I_R, d_{L1}) to reconstruct I^*_R and I^*_L at each scale, respectively.

Disparity Fusion and Refinement Module. As shown in Figure 2(a), the disparity refinement sub-network contains a disparity fusion and refinement module that processes the disparity maps (d_{L1}, d_{R1}) and (d_{L2}, d_{R2}) and outputs a pair of refined disparity maps (D_{finalL},

D_{finalR}) at each scale. This module is one of the main features of the DFRNets; it processes the inputted pairs of disparity maps by performing fusion and refinement to generate a pair of refined disparity maps, (D_{finalL}, D_{finalR}). Although DFRNets generates (D_{finalL}, D_{finalR}) from the left image I_L at four different scales, only the left disparity map D_{finalL} with scale equal to 1 is relevant at test time. This module has three key components: the left–right disparity reconstructor, the disparity flip operator, and the disparity fusion and boundary processor. The left–right disparity reconstructor, which is based on the left–right consistency term from Godard et al. (2017), enforces coherence when generating two pairs of refined disparity maps (D_{L1}, D_{R1}) and (D_{L2}, D_{R2}), which are given in Equation 1, 2, 3 and 4.

$$D_{L1}(x,y) = d_{R1}(x - d_{L1}(x,y), y), \quad [1]$$
$$D_{R1}(x,y) = d_{L1}(x + d_{R1}(x,y), y), \quad [2]$$
$$D_{L2}(x,y) = d_{R2}(x - d_{L2}(x,y), y), \quad [3]$$
$$D_{R2}(x,y) = d_{L2}(x + d_{R2}(x,y), y). \quad [4]$$

In designing this module, the authors expanded the left–right consistency principle of Godard et al. (2017) instead of merely using it as a term in the training loss function. Specifically, extending this principle required creating a left–right disparity reconstructor that generates two pairs of refined disparity maps from the Siamese network. Conversely, the disparity flip operator performs the horizontal flip operation on the disparity maps (D_{L2}, D_{R2}) to produce (D_{flipL}, D_{flipR}). To generate the final left disparity map D_{finalL}, the disparity fusion and boundary processor fuses (D_{L1}, D_{flipL}) by performing a pixel-wise mean operation and then removing the disparity ramps on the boundary pixels using the same technique as described in Godard et al. (2017). Essentially, removing the disparity ramps on the boundary pixels of the final left disparity map entails assigning the first 5% of D_{flipL} to the left of the final left disparity map and the last 5% of D_{L1} to the right of D_{finalL}. A similar step is taken to fuse (D_{R1}, D_{flipR}) and produce the final right disparity map D_{finalR}. Integrating this module as a trainable component of the model improves the model's performance significantly because it enables it to more effectively address the visual artifacts and blurred boundaries while performing depth estimation.

Loss Function

The model is designed to adopt an existing training loss function by reformulating it for joint depth estimation and refinement using a Siamese architecture, even though the original function was designed for depth estimation only and did not consider disparity refinement as a trainable component. As shown in Equation (5), the training loss at each scale s is a combination of three terms – appearance dissimilarity, disparity smoothness,

and left–right consistency – and is aggregated through four different scales for a total loss of $\mathcal{L} = \sum_{s=1}^{4} \mathcal{L}_s$, given in Equation 5, 6, 7 and 8.

$$\mathcal{L}_s = \alpha\mathcal{L}_{app} + \beta\mathcal{L}_{smooth} + \gamma\mathcal{L}_{lr}, \quad [5]$$

$$\mathcal{L}_{app} = \mathcal{L}_{app}^{left} + \mathcal{L}_{app}^{right}, \quad [6]$$

$$\mathcal{L}_{smooth} = \mathcal{L}_{smooth}^{left} + \mathcal{L}_{smooth}^{right}, \quad [7]$$

$$\mathcal{L}_{lr} = \mathcal{L}_{lr}^{left} + \mathcal{L}_{lr}^{right}, \quad [8]$$

where $\mathcal{L}_{app}$ measures the quality of the reconstructed images, $\mathcal{L}_{smooth}$ encourages the predicted disparities to be locally smooth, $\mathcal{L}_{lr}$ enforces consistency between the left and right disparities, and α, β, γ are the loss weightings for each term. This section provides details only for the left components $\mathcal{L}^{left}$ of the loss function since the right components $\mathcal{L}^{right}$ are defined symmetrically.

Appearance Dissimilarity Term. The appearance dissimilarity term measures the quality of the reconstructed image and usually involves minimizing the dissimilarity of pixel-wise correspondence between a target image and a reconstructed image. This term is a linear combination of the single-scale structural similarity (SSIM) term (Wang et al., 2004) and the L_1 photometric term, as defined in Equation (9). It is used in several studies to evaluate the quality of a reconstructed image (Godard et al., 2017; Li et al., 2018; Mahjourian et al., 2018; Wang et al., 2018; Yin & Shi, 2018; Yusiong & Naval, 2019). This term is given in Equation 9.

$$\mathcal{L}_{app}^{left} = \frac{1}{N}\sum_{x,y} \omega \frac{1-SSIM\left(I_L(x,y),\, I^*{}_L(x,y)\right)}{2} + (1-\omega)\|I_L(x,y) - I^*{}_L(x,y)\| \quad [9]$$

with a 3-by-3 box filter for the SSIM term, and ω is set to 0.85.

Disparity Smoothness Term. The disparity smoothness term is used to regularize the predicted disparities in textureless, low-gradient, and occluded regions to enforce the assumption that the predicted disparities must be locally smooth. As shown in Equation (10) and described in Godard et al. (2017) and Mahjourian et al. (2018), this term considers the gradient of the corresponding input image to allow for sharp changes in depth at pixel locations where there are sharp changes in the image. However, to train the DFRNets, this term is modified to include the final left disparity map D_{finalL} in the training loss. This term is given in Equation 10.

$$\mathcal{L}_{smooth}^{left} = \frac{1}{N}\sum_{x,y} \left| \partial_x D_{finalL}(x,y) e^{-|\partial_x I_L(x,y)|} + \partial_y D_{finalL}(x,y) e^{-|\partial_y I_L(x,y)|} \right|. \quad [10]$$

Left–Right Consistency Term. As described in Godard et al. (2017), Li et al. (2018), and

Yusiong and Naval (2019), the left–right consistency term enforces consistency between the left and right disparities and is crucial when generating the refined disparity maps. This modified term considers the final left disparity map D_{finalL}. This term is given in Equation 11.

$$\mathcal{L}_{lr}^{left} = \frac{1}{N}\sum_{x,y}\left|D_{finalL}(x,y) - D_{L1}(x,y)\right|. \quad [11]$$

Enabling the model to handle visual artifacts and blurred boundaries requires a simple modification to the disparity smoothness term and the left–right consistency term to incorporate the refined disparity maps, D_{finalL} and D_{finalR}, in the training loss computation. Reformulating the existing loss function is necessary to allowing the training algorithm to optimize all the outputs of the network by minimizing training loss, which enables the model to generate depth maps of the scenes with no border artifacts around the image boundary.

Implementation Details

This research utilized TensorFlow (Abadi et al., 2016) to implement DFRNets and trained the model from scratch using a single GTX 1080 Ti (11GB) GPU. The Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.999,$ and $\varepsilon = 10^{-8}$ trained the model for 50 epochs with a batch size of 2 by optimizing the training loss. The learning rate was initially set to $\lambda = 10^{-4}$ for the first 30 epochs and halved every 10 epochs afterward until the training was completed. The weightings of the different terms of the loss function were $\alpha = 1.0, \beta = \frac{0.1}{2^s}$, and $\gamma = 1.0$, where s is the output scale. The dimensions of the stereo image pairs were reduced to 256 by 512 for training, but at testing time, the network could predict disparity maps for single images of varying dimensions. The weight parameters were initialized randomly using the Xavier initialization procedure (Glorot & Bengio, 2010). To prevent overfitting, L_2 regularization was applied to all the weight parameters by adding a small constant, 0.00001. Furthermore, DFRNets was trained using the same train/test split as used in Eigen et al. (2014); this split is often referred to as the Eigen split. The dataset consisted of 22,600 stereo image pairs for training and 697 for testing. Training involved data augmentation, as in Godard et al. (2017).

RESULTS AND DISCUSSION

This section presents the results of the experiments conducted to evaluate the proposed framework, DFRNets, for monocular depth estimation in an unsupervised manner. The model was evaluated using the publicly available KITTI 2015 driving dataset (Gieger et al., 2012). Training the model entailed the use of rectified stereo image pairs, while testing required the left image to generate a depth map; the corresponding Velodyne data served as the ground-truth depth for benchmarking. Furthermore, the proposed model's performance was compared both quantitatively and qualitatively with that of existing state-of-the-art methods. An ablation study was also conducted to show the versatility of the proposed

framework and the advantages of integrating a disparity refinement component into the depth estimation model.

Specifically, the performance of DFRNets in monocular depth estimation was evaluated using the Velodyne ground-truth data of the test images. The experiment results were compared with various state-of-the-art methods by directly using the results reported in the original papers. As in the previous studies, an experiment was performed that involved pre-training the network on the Cityscapes dataset and then fine-tuning it on KITTI. Table 1 and Table 2 show the quantitative comparisons between the proposed model and other state-of-the-art methods in unsupervised monocular depth estimation using the depth evaluation metrics introduced in Eigen et al. (2014). For the training dataset, *K* means trained on the KITTI dataset, and *CS* + *K* means pre-trained on the Cityscapes dataset and fine-tuned on the KITTI dataset. For the training protocol, *depth* means the methods used ground-truth depths at training time, *mono* means the methods used monocular sequences for training, and *stereo* means the methods used rectified stereo images for training. The evaluation results using the KITTI test set reveal that the proposed model achieved the best performance on the two distance ranges: 0–80 meters and 1–50 meters, since it obtained the lowest errors and achieved the highest accuracy compared to the previous methods.

In addition to the quantitative results, qualitative comparisons to certain related methods using the KITTI test set, as shown in Figure 3, reveal that the proposed method generated depth maps that are visually more accurate than those produced by other methods, since these predicted depth maps have no border artifacts around the image boundary. Also, these results show that the proposed method significantly reduced the ghosting and shadow artifacts around the boundaries of the objects, thereby enabling the model to capture the underlying geometry of distant objects and objects in areas with thin structures and homogeneous regions. Moreover, the model can successfully reconstruct various objects that are difficult to recover, such as poles, tree trunks, and street symbols, and recover scene structures with more explicit object boundaries. These results demonstrate that simultaneously generating depth maps from the monocular images and refining the predicted depth maps in an end-to-end manner lead to better performance.

Architectural Analysis

The ablation study introduced three more variants by using ResNet50 (He et al., 2016), the modified DispNet with skip connections (Godard et al., 2017), and U-Net (Ronneberger et al., 2015) to better illustrate the effectiveness of jointly performing depth estimation and refinement instead of using a post-processing heuristic that is decoupled from the training process. Some modifications to the different network architectures were performed to incorporate a multi-scale structure for disparity predictions into the decoder section of the network. Experiments involved using the KITTI 2015 driving dataset to train the different

Table 1

Error metrics. Monocular depth estimation results using the KITTI test set and the Eigen split. The bold values indicate the best results

Method	**Training Dataset**	**Train**	**Error Metric (Lower Is Better)**			
			ARD	**SRD**	**RMSE (Linear)**	**RMSE (Log)**
Depth range: 0–80 meters						
Eigen et al. (2014) Coarse	K	Depth	0.194	1.531	7.216	0.273
Eigen et al. (2014) Coarse + Fine	K	Depth	0.190	1.515	7.156	0.270
DDVO (Wang et al., 2018)	K	Mono	0.151	1.257	5.583	0.228
	CS + K	Mono	0.148	1.187	5.496	0.226
GeoNet (Yin & Shi, 2018)	K	Mono	0.155	1.296	5.857	0.233
	CS + K	Mono	0.153	1.328	5.737	0.232
Mahjourian et al. (2018)	K	Mono	0.163	1.240	6.220	0.250
	CS + K	Mono	0.159	1.231	5.912	0.243
Zhou et al. (2017)	K	Mono	0.208	1.768	6.856	0.283
	CS + K	Mono	0.198	1.836	6.565	0.275
Godard et al. (2017)	K	Stereo	0.148	1.344	5.927	0.247
	CS + K	Stereo	0.124	1.076	5.311	0.219
AsiANet (Yusiong & Naval, 2019)	K	Stereo	0.145	1.349	5.909	0.230
	CS + K	Stereo	0.128	1.161	5.470	0.213
Ours (DFRNets)	K	Stereo	0.133	1.137	5.332	0.212
	CS + K	Stereo	0.114	0.927	4.885	**0.194**
Depth range: 1–50 meters						
GeoNet (Yin & Shi, 2018)	K	Mono	0.147	0.936	4.348	0.218
Mahjourian et al. (2018)	K	Mono	0.155	0.927	4.549	0.231
	CS + K	Mono	0.151	0.949	4.383	0.227
Zhou et al. (2017)	K	Mono	0.201	1.391	5.181	0.264
	CS + K	Mono	0.190	1.436	4.975	0.258
Garg et al. (2016) L12 Aug. 8x	K	Stereo	0.169	1.080	5.104	0.273
Godard et al. (2017)	K	Stereo	0.140	0.976	4.471	0.232
	CS + K	Stereo	0.117	0.762	3.972	0.206
AsiANet (Yusiong & Naval, 2019)	K	Stereo	0.122	0.786	4.014	0.198
	CS + K	Stereo	0.107	0.663	3.717	0.184
Ours (DFRNets)	K	Stereo	0.111	0.679	3.675	0.183
	CS + K	Stereo	0.096	0.539	3.325	**0.168**

Table 2

Accuracy metrics. Monocular depth estimation results using the KITTI test set and the Eigen split. The bold values indicate the best results

Method	**Training Dataset**	**Train**	**Accuracy Metric (Higher Is Better)**		
			$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Depth range: 0–80 meters					
Eigen et al. (2014) Coarse	K	Depth	0.679	0.897	0.967
Eigen et al. (2014) Coarse + Fine	K	Depth	0.692	0.899	0.967
DDVO (Wang et al., 2018)	K	Mono	0.810	0.936	0.974
	CS + K	Mono	0.812	0.938	0.975
GeoNet (Yin & Shi, 2018)	K	Mono	0.793	0.931	0.973
	CS + K	Mono	0.802	0.934	0.972
Mahjourian et al. (2018)	K	Mono	0.762	0.916	0.968
	CS + K	Mono	0.784	0.923	0.970
Zhou et al. (2017)	K	Mono	0.678	0.885	0.957
	CS + K	Mono	0.718	0.901	0.960
Godard et al. (2017)	K	Stereo	0.803	0.922	0.964
	CS + K	Stereo	0.847	0.942	0.973
AsiANet (Yusiong & Naval, 2019)	K	Stereo	0.824	0.936	0.970
	CS + K	Stereo	0.858	0.947	0.974
Ours (DFRNets)	K	Stereo	0.848	0.947	0.976
	CS + K	Stereo	0.878	0.958	**0.979**
Depth range: 1–50 meters					
GeoNet (Yin & Shi, 2018)	K	Mono	0.810	0.941	0.977
Mahjourian et al. (2018)	K	Mono	0.781	0.931	0.975
	CS + K	Mono	0.802	0.935	0.974
Zhou et al. (2017)	K	Mono	0.696	0.900	0.966
	CS + K	Mono	0.735	0.915	0.968
Garg et al. (2016) L12 Aug. 8x	K	Stereo	0.740	0.904	0.962
Godard et al. (2017)	K	Stereo	0.818	0.931	0.969
	CS + K	Stereo	0.860	0.948	0.976
AsiANet (Yusiong & Naval, 2019)	K	Stereo	0.864	0.953	0.978
	CS + K	Stereo	0.893	0.960	0.981
Ours (DFRNets)	K	Stereo	0.885	0.962	0.982
	CS + K	Stereo	0.909	0.969	**0.985**

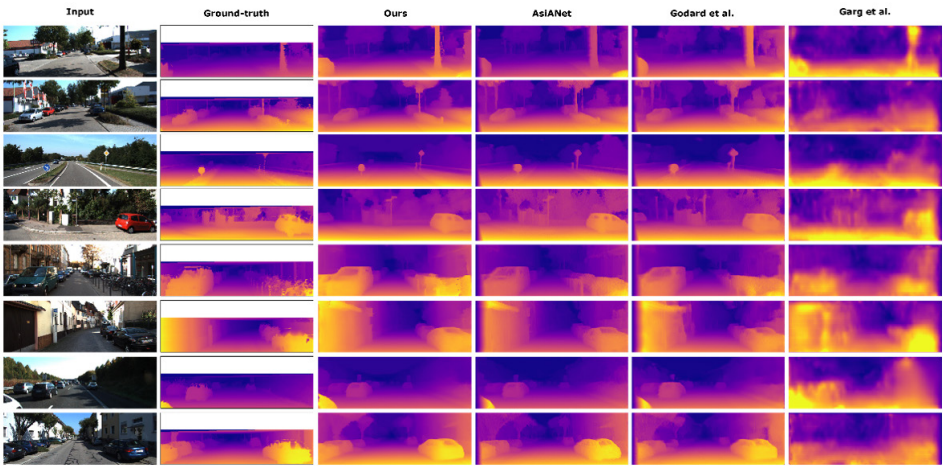

Figure 3. Qualitative results using the KITTI test set. A visual comparison of the results generated by the proposed method and with the results of Garg et al. (2016), Godard et al. (2017), and AsiANet (Yusiong & Naval, 2019). The ground-truth depth maps are interpolated for visualization purposes only. Best viewed in color.

non-Siamese networks and then applying the post-processing step as in Godard et al. (2017). Table 3 shows that the models based on the proposed framework perform much better than the other models. In Table 3, *K* means the network was trained on the KITTI dataset without the post-processing step, similar to Godard et al. (2017); *pp* means a post-processing step was performed on the output of the model, as in Godard et al. (2017); and *Ours* means DFRNets was implemented with the specified network architecture as the autoencoder. The results also demonstrate the versatility of the proposed framework, since it may use any network architecture that can generate disparity maps. Most importantly, the results clearly show the advantages of integrating a disparity refinement component into the depth estimation model.

Table 3

Architectural analysis. Results using the KITTI test set with a depth range of 0–80 meters. The bold values indicate the best results

Architecture	Method	Error Metric (Lower Is Better)			
		ARD	SRD	RMSE (Linear)	RMSE (Log)
DispNet	K	0.163	1.620	6.265	0.247
	pp	0.153	1.360	5.884	0.235
	Ours	**0.149**	**1.316**	**5.788**	**0.229**

Table 3 *(Continued)*

Architecture	Method	Error Metric (Lower Is Better)			
		ARD	SRD	RMSE (Linear)	RMSE (Log)
ResNet	K	0.148	1.344	5.839	0.233
	pp	0.140	1.181	5.557	0.223
	Ours	0.137	1.149	5.449	**0.217**
U-Net	K	0.151	1.466	5.980	0.237
	pp	0.141	1.223	5.585	0.224
	Ours	0.138	1.210	5.543	**0.219**
AsiaNet	K	0.145	1.349	5.909	0.230
	pp	0.135	1.132	5.475	0.217
	Ours	0.133	1.137	5.332	**0.212**

CONCLUSIONS

This work has presented an unsupervised learning framework, DFRNets, for jointly performing depth estimation and depth refinement using rectified stereo images during training. In essence, the model can be trained to simultaneously predict and refine disparity maps using a Siamese network architecture consisting of two autoencoders and a novel DFRM that performs disparity refinement as a trainable component of the model. The DFRM enables the model to more effectively handle visual artifacts and blurred boundaries, resulting in better performance. Moreover, this paper has shown that an existing training loss function can be reformulated for the joint learning and refinement of depth maps even though the original purpose was for depth estimation only. Experiment results using the KITTI 2015 driving dataset reveal that the proposed method achieved superior quantitative and qualitative performance compared to previous unsupervised state-of-the-art methods. In addition, the ablation study confirmed that the proposed framework is versatile, since it can use any encoder–decoder network architecture. Also, the results have revealed the advantages of performing these two tasks simultaneously and in an end-to-end manner rather than introducing a post-processing heuristic as a separate component of the model.

ACKNOWLEDGEMENT

This research work has been supported by the Department of Science and Technology-Engineering Research and Development for Technology Program. We also wish to express our appreciation to the editor and anonymous reviewers for their valuable comments and suggestions.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin., M., … Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. In *Computer science – Distributed, parallel and cluster computing* (pp. 1-19). Ithaca, New York: Cornell University.

Cadena, C., Latif, Y., & Reid, I. D. (2016, October 9-14). Measuring the performance of single image depth estimation methods. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 4150-4157). Daejeon, South Korea.

Eigen, D., Puhrsch, C., & Fergus, R. (2014, December 8-13). Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* (pp. 2366-2374). Palais des Congrès de Montréal, Montréal, Canada.

Garg, R., Kumar, V., Carneiro, G., & Reid, I. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. In B. Leibe, J. Matas, N. Sebe & M. Welling (Eds.), *European Conference on Computer Vision* (pp. 740-756). Cham, Switzerland: Springer.

Geiger, A., Lenz, P., & Urtasun, R. (2012, June 16-21). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3354-3361). Providence, RI, USA.

Glorot, X., & Bengio, Y. (2010, May 13-15). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Artificial Intelligence and Statistics Conference* (pp. 249-256). Sardinia, Italy.

Godard, C., Aodha, O. M., & Brostow, G. J. (2017, July 21-26). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 270-279). Honolulu, Hawaii.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June 27-30). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778). Las Vegas, NV, USA.

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015, December 7-12). Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* (pp. 2017-2025). Montreal, Canada.

Kingma, D., & Ba, J. (2015, May 7-9). Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations* (pp. 1-15). San Diego, CA, USA.

Li, R., Wang, S., Long, Z., & Gu, D. (2018, May 21-25). UnDeepVO: Monocular visual odometry through unsupervised deep learning. In *Proceedings of the IEEE International Conference on Robotics and Automation* (pp. 7286-7291). Brisbane, QLD, Australia.

Liu, M., Salzmann, M., & He, X. (2014, June 23-28). Discrete-continuous depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 716-723). Columbus, OH, USA.

Mahjourian, R., Wicke, M., & Angelova, A. (2018, June 18-22). 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* (pp. 5667-5675). Salt Lake City, Utah.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October 5-9). U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241). Munich, Germany.

Saxena, A., Chung, S. H., & Ng, A. Y. (2005, December 5-8). Learning depth from single monocular images. In *Proceedings of the 18th International Conference on Neural Information Processing Systems* (pp. 1161-1168). British Columbia, Canada.

Saxena, A., Chung, S. H., & Ng, A. Y. (2008). 3-d depth reconstruction from a single still image. *International Journal of Computer Vision, 76*(1), 53-69.

Wang, C., Buenaposada, J., Zhu, R., & Lucey, S. (2018, June 18-22). Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2022-2030). Salt Lake City, Utah.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error measurement to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600-612.

Yin, Z., & Shi, J. (2018, June 18-22). GeoNet: Unsupervised learning of dense, optical flow and camera pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1983-1992). Salt Lake City, Utah.

Yusiong, J. P. T., & Naval, P. C. Jr. (2019, January 7-11). AsiANet: Autoencoders in autoencoder for unsupervised monocular depth estimation. In *Proceedings of the IEEE Winter Conference of Applications on Computer Vision* (pp. 443-451). Waikoloa Village, USA.

Zhou, T., Brown, M., Snavely, N., & Lowe, G. (2017, July 21-26). Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6612-6619). Honolulu, Hawaii.