Unsupervised Monocular Depth Estimation with Multi-Baseline Stereo

Saad Imran[1]
saadimran@kaist.ac.kr
Muhammad Umar Karim Khan[2]
umar@kaist.ac.kr
Sikander Bin Mukarram[1]
sikander_ssalab@kaist.ac.kr
Chong-Min Kyung[2]
kyung@kaist.ac.kr

[1] Korea Advanced Institute of Science and Technology, Daejeon, South Korea
[2] Center of Integrated Smart Sensors, Daejeon, South Korea

Abstract

Unsupervised deep learning methods have shown promising performance for single-image depth estimation. Since most of these methods use binocular stereo pairs for self-supervision, the depth range is generally limited. Small-baseline stereo pairs provide small depth range but handle occlusions well. On the other hand, stereo images acquired with a wide-baseline rig cause occlusions-related errors in the near range but estimate depth well in the far range. In this work, we propose to integrate the advantages of the small and wide baselines. By training the network using three horizontally aligned views, we obtain accurate depth predictions for both close and far ranges. Our strategy allows to infer multi-baseline depth from a single image. This is unlike previous multi-baseline systems which employ more than two cameras. The qualitative and quantitative results show the superior performance of multi-baseline approach over previous stereo-based monocular methods. For 0.1 to 80 meters depth range, our approach decreases the absolute relative error of depth by 24% compared to Monodepth2. Our approach provides 21 frames per second on a single Nvidia1080 GPU, making it useful for practical applications. The code and dataset are publicly available at https://github.com/saadi297/MultiBaselineDepth

1 Introduction

Depth estimation is a commonly studied problem in computer vision due to the large number of applications. Accurate depth information is important for tasks such as 3D reconstruction and autonomous navigation. In the past, active techniques such as time-of-flight [46] and structured light [32], and passive systems based on stereo matching [31], and structure from motion (SFM) [36] have been used for depth estimation.

Recently, with the advancement of deep learning, many researchers have used Convolutional Neural Networks (CNNs) with self-supervision for single image depth estimation. Self-supervised methods for monocular depth estimation have shown promising performance in recent years. These methods treat depth estimation as an image reconstruction problem.

More specifically, a CNN is trained to generate disparity maps that are used to reconstruct the target images from the reference images. Self-supervised methods are preferred over supervised deep learning approaches as the former do not require ground truth depth data, which is expensive and hard to gather.

Existing monocular depth estimation models can be trained by either using monocular video or rectified stereo pairs. These approaches have some challenges. Besides training a depth estimation network, monocular video-based methods also require relative pose information between the adjacent frames in a sequence. On the other hand, stereo approaches do not require training a pose estimation network and are more effective than video-based methods. Despite this advantage, existing stereo techniques use two cameras (left and right images) with a fixed baseline for training, which can cause occlusion and limited depth range.

To deal with these problems, researchers have proposed multi-camera systems [15, 25] with multiple baselines. The advantage of a multi-baseline setup is that it can provide good depth accuracy both in near and far ranges compared to standard stereo, which only works well in a certain range. However, the problems with such systems are that they are quite expensive and have high computational load due to multiple cameras; hence, they are not commonly used. In this paper, we aim to improve disparity estimation with a monocular camera by leveraging multiple baselines at training time.

We present an unsupervised learning approach that uses two different baselines during training and a single image at test time. Our approach makes use of the advantages of multi-baseline stereo without increasing the computational complexity at inference. In contrast to two-camera stereo-based monocular methods, our method gives improved disparity maps in both near and far ranges. To our knowledge, we are the first to employ the principle of multi-baseline to unsupervised disparity estimation. Experimental results show that our method yields much improved results compared to stereo-based self-supervised monocular depth estimation.

2 Related Work

Although various methods have been proposed in the past to extract depth from images, we discuss the literature related to our work only.

Classical Stereo Depth Estimation. Stereo matching that involves two horizontally displaced cameras to observe a given scene is one of the most popular approaches. The shift between the corresponding pixels in observed left and right images gives the disparity, which is inversely proportional to the depth at the pixels. Traditional stereo matching algorithms usually include all or some of the four tasks: computing the matching cost, aggregating the cost, computing the disparity, and refining the disparity. A detailed description of these tasks is given in [31]. Among different stereo matching techniques, semi-global matching (SGM) [14] is one of the most frequently used approaches because of its efficiency. Single baseline stereo setup poses some problems. For example, using wider baseline increases the chance of false matches due to large disparity search range. On the other hand, short baseline reduces the risk of false matching but suffers from poor accuracy in the far range. It is well known that using more views can solve these problems. Ito and Ishii [17] proposed using three views in a triangular configuration to improve the matching and handle the occlusion. Okutomi and Kanade [25] generated multiple baselines by laterally displacing the camera. They showed that matching across different baseline stereo images circumvents the problem of incorrect

matching and results in more accurate disparity maps. Gallup *et al*. [7] proposed a multi-baseline, multi-resolution technique and achieved a constant depth accuracy by changing the baseline and resolution accordingly. Honegger *et al*. [15] used four cameras in a parallel arrangement to propose a multi-baseline stereo system that works in real time.

Supervised Monocular and Stereo Depth Estimation. Many supervised learning based schemes have been proposed for depth extraction. Make3D [30] modified a Markov Random Field (MRF) to predict the 3D structure of a scene from a single image. Eigen *et al*. [5] used two neural networks to infer monocular depth by combining global and local structure of the input image. Lie *et al*. [21] showed the advantage of jointly training a CNN and a conditional random field (CRF) for monocular depth estimation. In [6], authors treat monocular depth estimation as an ordinal regression problem, while in [20], authors use relative depth maps to achieve state-of-the-art results. Many researchers have used deep learning for stereo depth estimation. Zbontar and LeCun [42] proposed two network architectures to compare image patches for computing the stereo matching cost followed by traditional post-processing steps. Luo *et al*. [22] obtained better results by treating stereo matching problem as a multi-class classification task. Kendall *et al*. [19] presented an end-to-end learning technique for disparity estimation, which does not require additional post-processing. In [41], authors proposed a multi-scale approach to predict depth from unrectified stereo images. Tosi *et al*. [37] combined local and global features to obtain accurate confidence scores. Tulyakov *et al*. [39] proposed a stereo matching technique to deal with large memory and dynamic disparity range requirements. Zhang *et al*. [43] introduced two additional layers for efficient and accurate cost aggregation. In [4], authors proposed a faster method of stereo matching by discarding most of the disparities during cost aggregation.

Unsupervised Monocular and Stereo Depth Estimation. Currently, unsupervised methods for depth estimation are becoming more popular as they do not require expensive ground truth data. Garg *et al*. [8] were the first to propose a fully unsupervised approach for monocular depth estimation. They used rectified stereo images for training and performed Taylor series expansion to linearize the image warping process. Godard *et al*. [10] used the bilinear sampler [18] for image warping and introduced left-right consistency loss to obtain accurate depth results. In [29], authors imposed trinocular stereo assumptions to yield enhanced depth results. Poggi *et al*. [28] deployed a pyramidal architecture to enable monocular depth estimation on embedded systems, while the authors in [27] used adversarial learning for stereo depth estimation. Tosi *et al*. [38] added proxy supervision by obtaining disparity maps through traditional stereo matching technique. SuperDepth [26] incorporated Single-Image Super-Resolution (SISR) [33] technique to obtain high resolution disparity maps. Recently, many researchers have used monocular videos for self-supervision to predict depth from a single image [11, 12, 45]. Our method uses multi-baseline stereo images as self-supervision for monocular depth estimation.

3 Unsupervised Multi-Baseline approach

Although supervised learning can be used for multi-baseline stereo, it has its problems. We can extend the work of [39, 42] to match more than two views, however, using such approaches has two major drawbacks. First, they require pixel-wise labelling to generate ground truth depths for training. Second, we also need to use more than two cameras at test time to perform matching, which is highly undesirable. Therefore, we train the model in a self-supervised fashion and use a single image for inference.

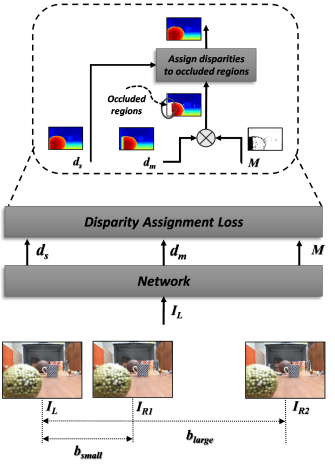

Figure 1: The concept behind our training approach. b_{small} and b_{large} refer to small and wide baselines, respectively.

Figure 1 shows the basic idea of our approach. For training, we use three aligned views to get two different baselines. The images I_L and I_{R1} act as a small-baseline stereo system, and hence have fewer occlusions [2] but provide accurate near depth. On the other hand, the images I_L and I_{R2} act as a wide-baseline stereo system, and thereby have more occlusions but provide accurate depth at far range. The network outputs occlusion mask M, small-baseline disparity d_s, and disparity d_m. The purpose of occlusion mask is to find the occluded pixels in the disparity d_m. The disparities at these occluded pixels are replaced by the corresponding small-baseline disparities using disparity assignment loss. This is based on the assumption that occlusion is not a serious problem for small-baseline stereo systems. In the remaining text, d_m will indicate multi-baseline disparity.

3.1 Proposed Network

Given the three rectified images during training, our network learns to infer depth from a single image at test time. As depicted in Figure 2, the left image I_L is fed into the network. We use a shared encoder and three decoders to train the model. Each of the decoders has its own purpose. Decoder 1 utilizes small-baseline image pair I_L and I_{R1} to generate small-baseline disparity map d_s, whereas Decoder 2 uses wide-baseline stereo images I_L and I_{R2} for self-supervision to generate left disparity d_l and right disparity d_r, respectively. Multi-baseline disparity d_m is generated by Decoder 3, which makes use of the image I_L, small-baseline disparity d_s, and the occlusion mask M for supervision.

Similar to [13], we compute the occlusion mask M using the left-right consistency check between the output disparities of Decoder 2,

$$M = |d_l' - d_l| > 1, \tag{1}$$

where d_l' is obtained by warping d_r to d_l. In occluded areas, disparities will have different values [9]; therefore, we set the threshold to greater than 1 pixel. We assign 1 and 0 to occluded and non-occluded pixels, respectively.

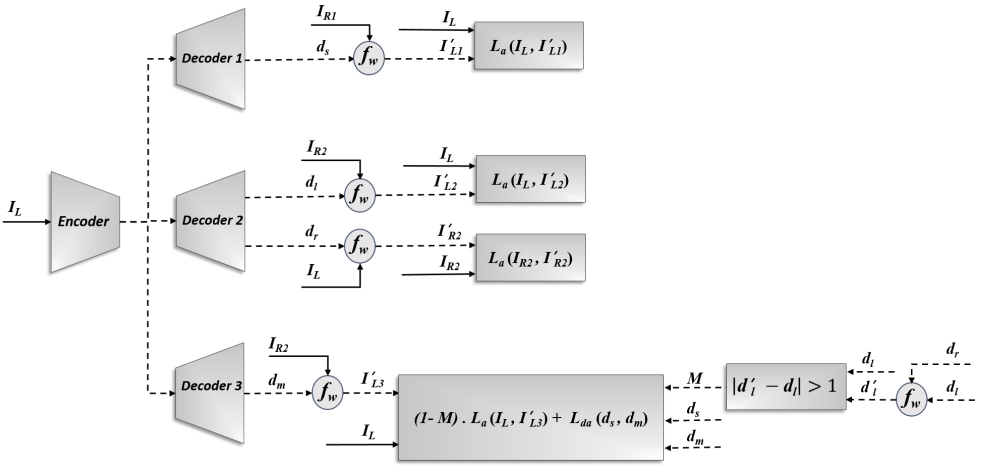

Figure 2: Illustration of our multi-baseline approach. We use three decoders for training. At test time, we only use Decoder 3 to output disparity d_m. Occlusion mask M is computed using left-right consistency check. The function f_w performs warping operation. Dotted lines represent internal signals.

Our network architecture is similar to the encoder-decoder architecture of [10]. The encoder is based on VGG16 [35] and all the decoders have same architecture, except Decoder 2 has two output channels for the disparity maps. Each decoder outputs disparity maps at four scales: one-eighth, quarter, half and full resolutions. Note that the three decoders are only needed for training. Only Decoder 3 is used at inference.

3.2 Training Losses

We train our network with multiple losses. In addition to image reconstruction loss L_a and disparity smoothness loss L_s, we also use disparity assignment loss L_{da}. All the losses are minimized at four scales.

Image Reconstruction Loss Following [10], we use the weighted sum of SSIM [40] and L1 loss to minimize the photometric error between the reconstructed and original images,

$$L_a(I,I') = \frac{1}{N}\sum_{i,j}\left(\alpha\frac{1-SSIM(I_{ij},I'_{ij})}{2} + (1-\alpha)\left|I_{ij}-I'_{ij}\right|\right), \tag{2}$$

where N is the number of pixels, and I and I' are the original and reconstructed images. To compute SSIM, we use a block filter of size 3x3 instead of a Gaussian. SSIM loss is based on three measurements: contrast, luminance and structure. Hence, it is also effective in case of high illumination differences between left and right stereo images. α is set to 0.85 based on results in [44].

Disparity Smoothness Loss Similar to [10], we define an edge-aware smoothness loss to deal with disparity discontinuities.

$$L_s(d,I) = \frac{1}{N}\sum_{i,j}|\partial_x d_{ij}|e^{-|\partial_x I_{ij}|} + |\partial_y d_{ij}|e^{-|\partial_y I_{ij}|}. \tag{3}$$

Here d is the disparity, I is the corresponding image, and ∂_x, ∂_y are the horizontal and vertical gradients, respectively. Note that this loss discourages disparity smoothness in absence of small image gradients.

Disparity Assignment Loss To replace the occluded pixels of the disparity map d_m (Fig. 2) by the small-baseline disparity map d_s, we again employ the combination of L1 and SSIM losses as

$$L_{da}(d_s, d_m) = M \cdot \frac{1}{N} \sum_{i,j} \left(\beta \frac{1 - SSIM(r \cdot d_s, d_m)}{2} + (1-\beta) |r \cdot d_s - d_m| \right), \tag{4}$$

where β is the weighting factor set to 0.85 based on experiments (see supplementary material), and r is the ratio of wide baseline to small baseline. The occlusion mask M ensures that only occluded pixels are considered. The factor r is used to scale the disparity d_s to match the disparity range of d_m. To ensure that the disparity d_m follows the disparity d_s in occluded regions, the gradients for this loss function are not computed with respect to d_s. In other words, only the weights of Decoder 3 will change to minimize L_{da}. The total loss is the combination of image reconstruction losses L_{recon}, smoothness losses L_{smooth} and Decoder 3 loss L_{dec3}.

$$L_{total} = L_{recon} + \lambda (L_{smooth}) + L_{dec3}, \tag{5}$$

where λ is the weighting factor set to 0.1. The losses L_{recon}, L_{smooth} and L_{dec3} are defined as follows:

$$L_{recon} = L_a(I_L, I'_{L1}) + L_a(I_L, I'_{L2}) + L_a(I_{R2}, I'_{R2}) \tag{6}$$

$$L_{smooth} = L_s(d_s, I_L) + L_s(d_l, I_L) + L_s(d_r, I_{R2}) \tag{7}$$

$$L_{dec3} = (1-M) \cdot L_a(I_L, I'_{L3}) + L_{da}(d_s, d_m) + \lambda \cdot L_s(d_m, I_L). \tag{8}$$

In Eq. 8, the term $(1-M)$ ensures that the occluded pixels do not contribute to the image reconstruction loss $L_a(I_L, I'_{L3})$. The occluded pixels are filled using L_{da} loss only.

4 Experimental Results

Due to the unavailability of multi-baseline stereo datasets, researchers have not investigated the advantages of using more than one baseline during training. Although some researchers [23, 34] have acquired trinocular stereo datasets from Bumblebee XB3 camera, they are more focused on localization and mapping. Moreover, these datasets do not provide the calibration parameters to horizontally align the three views. As the existing stereo-based unsupervised methods require rectified image pairs for training, such datasets are not feasible to use for the task of multi-baseline depth estimation. We develop our own dataset to show the importance of using multiple baselines during training.

We evaluate the effectiveness of our approach both qualitatively and quantitatively. We compare the performance of our approach with the stereo-based methods proposed in the past. We compare our results against Monodepth [10], monoResMatch [32], Monodepth2 [11], and 3Net [29]. For fair comparison, we train monoResMatch without the proxy-supervised loss. The results validate that the proposed approach yields more accurate disparity predictions.

Method	Baseline	Lower the better				Higher the better		
		AbsRel	*SqRel*	*RMSE*	*RMSELog*	δ <1.25	δ <1.25^2	δ <1.25^3
Depth Range = 0.1-10 *m*								
Monodepth [ ]+pp	10*cm*	0.2378	14.9776	9.321	0.311	0.962	0.973	0.976
Monodepth[ ]+pp	54*cm*	0.3667	24.9676	10.709	0.387	0.953	0.966	0.971
3Net [ ]+pp	10*cm*	0.2076	13.3266	8.868	0.298	0.970	0.977	0.979
3Net [ ]+pp	54*cm*	0.9180	65.9368	14.974	0.612	0.920	0.935	0.942
monoResMatch [ ]+pp	10*cm*	0.5110	35.2223	9.324	0.330	0.958	0.966	0.969
monoResMatch [ ]+pp	54*cm*	0.1784	11.8352	7.499	0.250	0.974	0.981	0.984
Monodepth2 [ ]	10*cm*	0.1182	7.0760	6.892	0.221	0.980	0.986	0.988
Monodepth2 [ ]	54*cm*	0.1863	12.5571	6.949	0.238	0.975	0.982	0.984
Ours	10*cm*,54*cm*	0.1232	7.6787	6.273	0.207	0.979	0.986	0.988
Depth Range = 10-80 *m*								
Monodepth [ ]+pp	10*cm*	0.2027	3.8602	9.929	0.311	0.727	0.883	0.940
Monodepth[ ]+pp	54*cm*	0.1772	3.0741	9.834	0.321	0.753	0.882	0.934
3Net [ ]+pp	10*cm*	0.2105	5.2326	10.102	0.301	0.771	0.898	0.942
3Net [ ]+pp	54*cm*	0.1391	2.3609	8.347	0.273	0.829	0.918	0.953
monoResMatch [ ]+pp	10*cm*	0.3239	10.6855	13.157	0.377	0.667	0.843	0.908
monoResMatch [ ]+pp	54*cm*	0.1677	4.7352	9.406	0.287	0.838	0.912	0.945
Monodepth2 [ ]	10*cm*	0.1932	4.7333	9.930	0.284	0.784	0.906	0.947
Monodepth2 [ ]	54*cm*	0.1223	2.3100	7.896	0.240	0.864	0.937	0.963
Ours	10*cm*,54*cm*	0.1276	2.0940	7.967	0.255	0.843	0.928	0.960
Depth Range = 0.1-80 *m*								
Monodepth[ ]	10*cm*	0.1520	4.4573	7.028	0.259	0.890	0.949	0.969
Monodepth [ ]+pp	10*cm*	0.1023	1.5862	5.804	0.202	0.891	0.954	0.976
Monodepth [ ]	54*cm*	0.3192	12.3581	8.279	0.397	0.870	0.921	0.944
Monodepth[ ]+pp	54*cm*	0.1364	3.5626	6.150	0.250	0.892	0.947	0.969
3Net [ ]	10*cm*	0.1047	2.2487	6.036	0.203	0.909	0.959	0.977
3Net [ ]+pp	10*cm*	0.1028	2.1953	5.950	0.201	0.911	0.960	0.977
3Net [ ]	54*cm*	0.2794	11.4045	7.255	0.361	0.897	0.938	0.956
3Net [ ]+pp	54*cm*	0.2779	11.3585	7.214	0.360	0.898	0.938	0.956
monoResMatch [ ]	10*cm*	0.4658	25.1243	10.801	0.377	0.824	0.904	0.939
monoResMatch [ ]+pp	10*cm*	0.3711	20.1240	8.948	0.299	0.864	0.930	0.954
monoResMatch [ ]	54*cm*	0.3129	12.9569	8.748	0.390	0.884	0.925	0.945
monoResMatch [ ]+pp	54*cm*	0.1004	3.3953	5.770	0.194	0.933	0.964	0.977
Monodepth2 [ ]	10*cm*	0.0843	1.7101	5.781	0.184	0.918	0.964	0.979
Monodepth2 [ ]	54*cm*	0.0864	2.3483	4.969	0.186	**0.939**	0.969	0.981
Ours	10*cm*,54 *cm*	**0.0643**	**0.9509**	**4.695**	**0.163**	0.936	**0.971**	**0.984**

Table 1: Evaluation on CARLA dataset. For comparison, we train previous methods separately with 10 *cm* and 54 *cm* baseline stereo images. Maximum predictions of all the networks are capped to 80 *m*. pp stands for post-processing.

4.1 Datasets

CARLA Dataset CARLA simulator [3] is used to acquire the multi-baseline dataset. We attach three cameras to the vehicle in a parallel arrangement to obtain horizontally aligned images. In addition, we also add a depth sensor to get ground truth depth maps for evaluation. We choose the large baseline to be 54 *cm*, which is equal to the baseline used in KITTI dataset [24]. The small baseline is chosen as 10 *cm*. The simulator is run in auto-pilot mode under clear weather conditions to gather the dataset. The dataset consists of approximately 14000 images, out of which 1300 images are used for evaluation while the remaining are used for training.

For evaluation, we use the metrics given in [5]: Abs Rel, Sq Rel, RMSE linear, RMSE log, and threshold-based metrics δ. The predicted disparity maps are converted to depth maps using baseline and focal length to compute these errors. For our approach, we use 54 *cm* baseline to get depth predictions. To solve the problem of unmatched regions on the left border of the disparity maps, we post-processed the results of Monodepth [10], monoResMatch [38], and 3Net [29] at test time using the method applied in [10]. In detail, we compute two disparity maps d and d‘ corresponding to the input image I and its flipped image I‘. The disparity d‘ is then flipped to get d“, which is aligned with the d. In d“, disparity ramps or unmatched regions will be located on the right border. To get the final prediction d^{pp}, we assign 10% left most pixels of d“ to left side of d^{pp}. Similarly, 10% right most pixels of d are assigned to right side of d^{pp}. The central pixels of d^{pp} are obtained by averaging central pixels of d“ and d. We report the results with and without post-processing.

Table 1 shows the detailed results on different depth ranges. From the results, it is seen

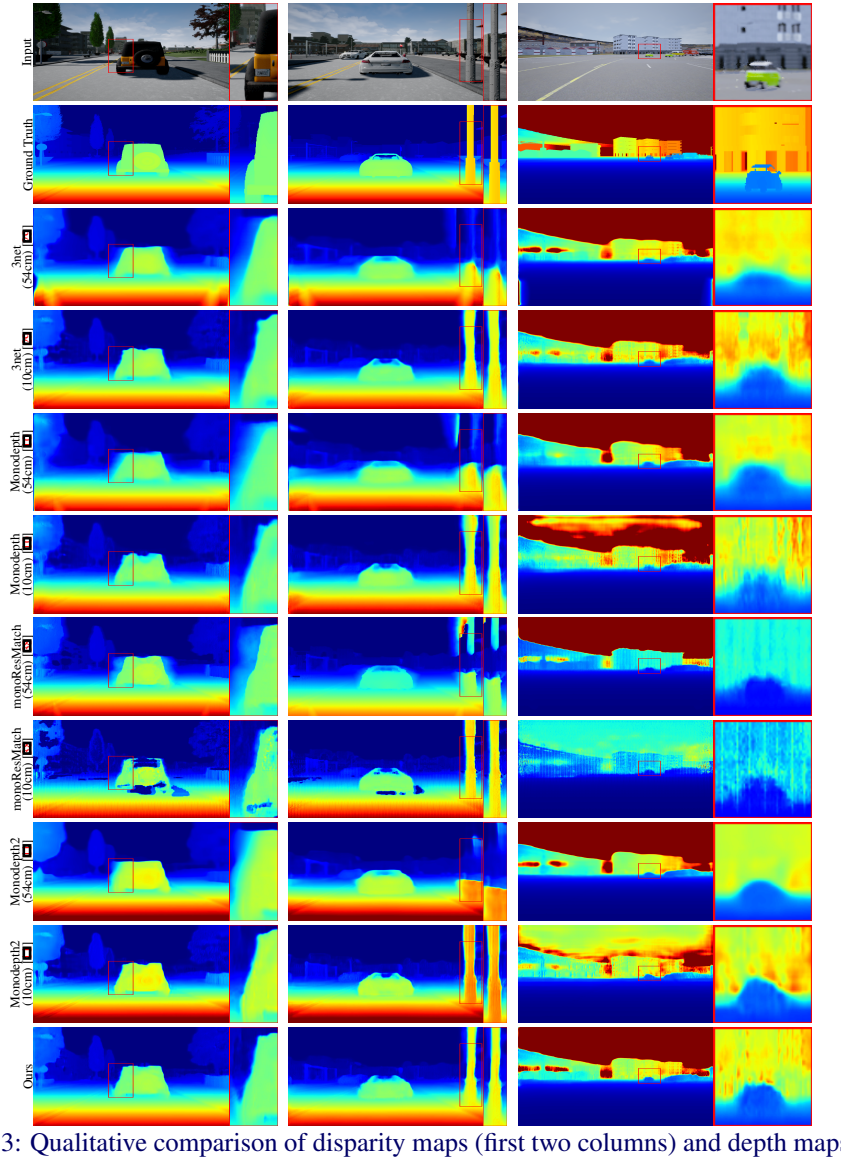

Figure 3: Qualitative comparison of disparity maps (first two columns) and depth maps (last column) on CARLA dataset. Zoomed-in views show that our method produces better results for both close and far objects.

that Monodepth [10], Monodepth2 [11] and 3Net [29] trained on wide baseline perform better at depth range greater than 10 meters. This is due to the fact that wide baseline can not deal with the occlusions caused by near objects. The monoResMatch [38] trained on 10 *cm* baseline performs worst among all the methods. For 0.1 to 80 meters depth range, our approach outperforms previous methods. This is expected as our method makes use of both narrow and wide baselines. Although post-processing considerably improves the results Monodepth, monoResMatch and 3Net, our approach does not require any post-processing.

For qualitative comparison, we provide disparity maps and zoomed-in views as shown

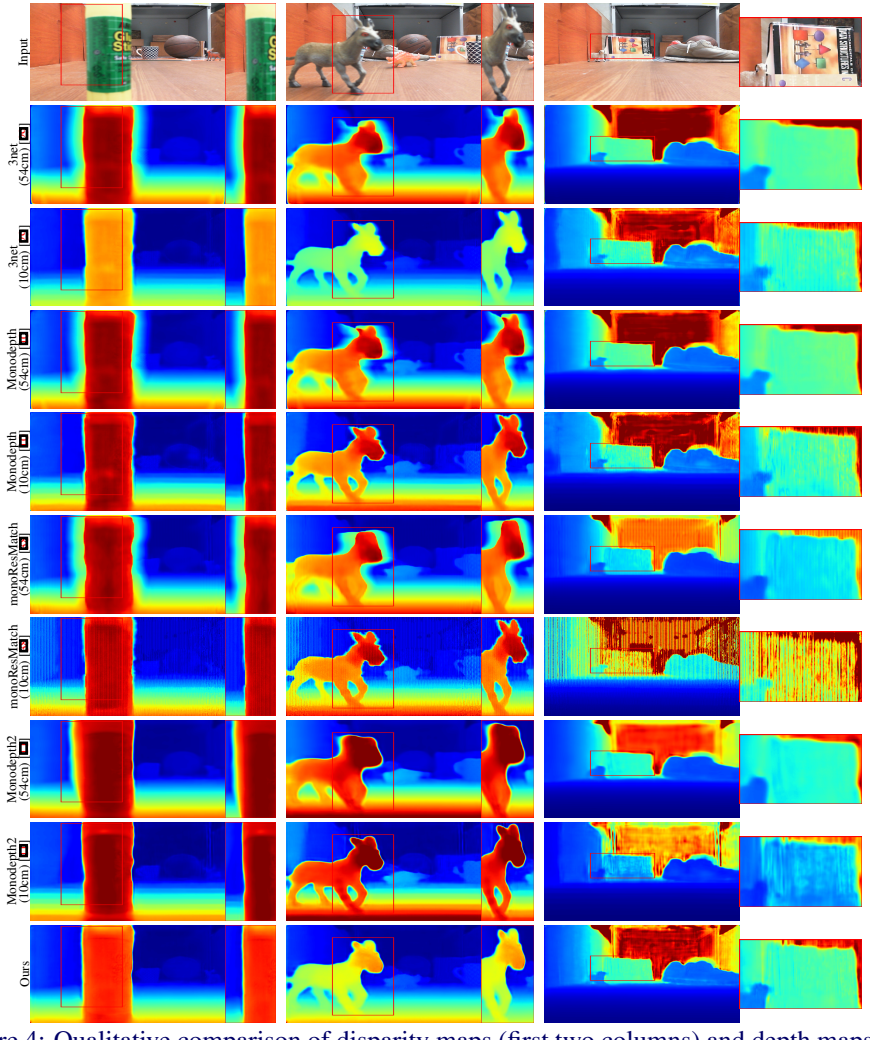

Figure 4: Qualitative comparison of disparity maps (first two columns) and depth maps (last column) on small objects dataset. Our method generates more accurate predictions with sharp boundaries.

in Figure 3. Results illustrate that small baseline tends to over-smooth the disparity predictions in far regions due to lower depth resolution. On the contrary, wide baseline provides more accurate predictions for far regions but generates severe occlusion artifacts for closer surfaces. Methods trained on wide baseline also perform worst in estimating the depth of thin surfaces such as poles. From the results, it is obvious that for close and far regions, our approach performs similar to small and large baselines, respectively. Hence, producing much improved depth estimates.

Small Objects Dataset We prepare another dataset to demonstrate the usefulness of our approach on real scenes. The dataset is captured using Microsoft LifeCam webcam. Instead of using three cameras, we employ single camera and displace it laterally to acquire three parallel images of each scene similar to [16]. Zaber's A-LSQ600D motorized linear

translation stage is used to control the position of camera. We capture the images of small objects to build the dataset of 5800 images. Training and test sets contain 5500 and 300 images respectively. All the objects are placed within two meters range. We set the small and wide baselines to 2 *mm* and 10 *mm*, respectively.

We show the results in Figure 4. Results clearly depict the superior performance of our approach over single baseline methods. The disparity maps of Figure 4 illustrate that for closer objects, our method produces crisp and accurate disparity maps similar to small baseline. In contrast, wide baseline generates serious artifacts near close object boundaries. The depth maps provide the clear picture of far objects. For far objects, small baseline fails to estimate accurate depth. On the other hand, multi-baseline method gives much accurate depth predictions similar to wide baseline. Again the results verify the efficacy of multi-baseline training over single baseline training.

4.2 Implementation Details

We implement our network in Tensorflow [1]. We train the model for 70 epochs with batch size of 8. All other hyperparameters are set as in [10]. We use Adam Optimizer with decay parameters β_1 and β_2 set to 0.9 and 0.999, respectively. Epsilon ε is set to 10^{-8}. We use a learning rate of 10^{-4} for first 35 epochs, 10^{-8} for 36 to 53 epochs, and 10^{-16} for last 17 epochs. The same hyperparameters were used for both datasets, showing generalization across datasets. All the images are sub-sampled to 256×512 before passing into the network. We also perform color augmentation as in [10]. Training on 13000 images for 70 epochs takes around 36 hours on Nvidia GTX 1080 GPU. The number of trainable parameters are approximately 66.2 million. It should be noted that at test time, we only use the output of decoder 3; therefore, depth prediction is fast and inference time of the network is 21 frames per second. A CPU implementation on an Intel Core i5 processor with 8GB RAM provides 2 frames per second.

5 Conclusion

In this work, we propose a novel multi-baseline technique for unsupervised monocular depth estimation. We overcome the shortcomings of single-baseline stereo supervision by training the model with two stereo baselines. Our model combines the advantages of small and wide baseline stereo systems. Unlike previous stereo approaches that work well in a certain range, our method generates accurate disparity maps both in near and far ranges. Furthermore, our method uses only a single camera at test time to predict multi-baseline depth. This is in contrast to traditional multi-baseline systems, which require more than two cameras to provide real time depth. Therefore, the proposed method is well-suited for practical applications.

Acknowledgement

This work was supported by Ministry of Science, ICT and Future Planning (CISS-2013073718).

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow,

Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

[2] Julie Delon and Bernard Rougé. Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 28(3):209–223, 2007.

[3] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[4] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *IEEE International Conference on Computer Vision*, pages 4384–4393, 2019.

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[7] David Gallup, Jan-Michael Frahm, Philippos Mordohai, and Marc Pollefeys. Variable baseline/resolution stereo. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[8] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.

[9] Christos Georgoulas and Ioannis Andreadis. A real-time occlusion aware hardware structure for disparity map computation. In *International Conference on Image Analysis and Processing*, pages 721–730. Springer, 2009.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *IEEE International Conference on Computer Vision*, pages 3828–3838, 2019.

[12] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *IEEE International Conference on Computer Vision*, pages 8977–8986, 2019.

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

[14] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005.

[15] Dominik Honegger, Torsten Sattler, and Marc Pollefeys. Embedded real-time multi-baseline stereo. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5245–5250. IEEE, 2017.

[16] Saad Imran, Sikander Bin Mukarram, Muhammad Umar Karim Khan, and Chong-Min Kyung. Unsupervised deep learning for depth estimation with offset pixels. *Optics Express*, 28(6):8619–8639, 2020.

[17] Minoru Ito and Akira Ishii. Three-view stereo analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4):524–532, 1986.

[18] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[19] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[20] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019.

[21] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015.

[22] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

[23] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

[24] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[25] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. *IEEE Transactions on pattern analysis and machine intelligence*, 15(4):353–363, 1993.

[26] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9250–9256. IEEE, 2019.

[27] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *2018 International Conference on 3D Vision (3DV)*, pages 587–595. IEEE, 2018.

[28] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. Towards real-time unsupervised monocular depth estimation on cpu. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5848–5854. IEEE, 2018.

[29] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International Conference on 3D Vision (3DV)*, pages 324–333. IEEE, 2018.

[30] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

[31] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[32] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, page 195–202. IEEE, 2003.

[33] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[34] Patrick Y Shinzato, Tiago C dos Santos, Luis Alberto Rosero, Daniela A Ridel, Carlos M Massera, Francisco Alencar, Marcos Paulo Batista, Alberto Y Hata, Fernando S Osório, and Denis F Wolf. Carina dataset: An emerging-country urban scenario benchmark for road detection systems. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 41–46. IEEE, 2016.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Peter Sturm and Bill Triggs. A factorization based algorithm for multi-image projective structure and motion. In *European conference on computer vision*, pages 709–720. Springer, 1996.

[37] Fabio Tosi, Matteo Poggi, Antonio Benincasa, and Stefano Mattoccia. Beyond local reasoning for stereo confidence estimation with deep learning. In *European Conference on Computer Vision (ECCV)*, pages 319–334, 2018.

[38] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.

[39] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. In *Advances in Neural Information Processing Systems*, pages 5871–5881, 2018.

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

[41] Pramod Yadati and Anoop M Namboodiri. Multiscale two-view stereo using convolutional neural networks for unrectified images. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 346–349. IEEE, 2017.

[42] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *The journal of machine learning research*, 17(1):2287–2318, 2016.

[43] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019.

[44] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*, 2015.

[45] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. In *IEEE International Conference on Computer Vision*, pages 6872–6881, 2019.

[46] Jiejie Zhu, Liang Wang, Ruigang Yang, and James Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.