Scene disparity estimation with convolutional neural networks

Anas, Essa, Guo, Li, Onsy, Ahmed and Matuszewski, Bogdan

Anas, Essa, Guo, Li ORCID: 0000-0003-1272-8480, Onsy, Ahmed ORCID: 0000-0003-0803-5374 and Matuszewski, Bogdan ORCID: 0000-0001-7195-2509 (2019) Scene disparity estimation with convolutional neural networks. SPIE Proceedings Multimodal Sensing: Technologies and Applications, 11059 . 110590T1-110590T9. ISSN 0277-786X

It is advisable to refer to the publisher's version if you intend to cite from the work.
http://dx.doi.org/10.1117/12.2527628

Scene disparity estimation with convolutional neural networks

Essa R. Anas, Li Guo, Ahmed Onsy, Bogdan J. Matuszewski

Essa R. Anas, Li Guo, Ahmed Onsy, Bogdan J. Matuszewski, "Scene disparity estimation with convolutional neural networks," Proc. SPIE 11059, Multimodal Sensing: Technologies and Applications, 110590T (21 June 2019); doi: 10.1117/12.2527628

SPIE.

Event: SPIE Optical Metrology, 2019, Munich, Germany

Scene Disparity Estimation with Convolutional Neural Networks

Essa R. Anas, Li Guo, Ahmed Onsy, Bogdan J. Matuszewski
Computer Vision and Machine Learning (CVML) Research Group, School of Engineering,
University of Central Lancashire, Preston, UK

ABSTRACT

Estimation of stereovision disparity maps is important for many applications that require information about objects' position and geometry. For example, as depth surrogate, disparity maps are essential for objects' 3D shape reconstruction and indeed other applications that do require three-dimensional representation of a scene. Recently, deep learning (DL) methodology has enabled novel approaches for the disparity estimation with some focus on the real-time processing requirement that is critical for applications in robotics and autonomous navigation. Previously, that constraint was not always addressed. Furthermore, for robust disparity estimation the occlusion effects should be explicitly modelled. In the described method, the effective detection of occlusion regions is achieved through disparity estimation in both, forward and backward correspondence model with two matching deep subnetworks. These two subnetworks are trained jointly in a single training process. Initially the subnetworks are trained using simulated data with the know ground truth, then to improve generalisation properties the whole model is fine-tuned in an unsupervised fashion on real data. During the unsupervised training, the model is equipped with bilinear interpolation warping function to directly measure quality of the correspondence with the disparity maps estimated for both the left and right image. During this phase forward-backward consistency constraint loss function is also applied to regularise the disparity estimators for non-occluding pixels. The described network model computes, at the same time, the forward and backward disparity maps as well as corresponding occlusion masks. It showed improved results on simulated and real images with occluded objects, when compared with the results obtained without using the forward-backward consistency constraint loss function.

Keywords: Deep Learning, Disparity Estimation, Occlusion Detection, Structural Similarity Index, Unsupervised Learning.

1. INTRODUCTION

With calibrated cameras, disparity can be used as a surrogate for scene depth, and has a number of possible applications, including: 3D scene reconstruction, robot navigation, dimensional inspection, or more generally can facilitate verification of manufacturing processes.

Numerous approaches have been proposed in literature to estimate disparity and various datasets are available for training and evaluations of these methods. While the disparity estimation for indoor scenes is an important problem (with relevant datasets being available, e.g. Middlebury [1]), the disparity estimation for outdoor scenes is a more challenging problem due to a large variability of scenes, objects, illumination conditions, as well as large range of estimated disparity values. Furthermore, popular inexpensive depth sensors based on an active infrared illumination (e.g. Kinect) cannot be normally used in that context and the dense depth maps need to be estimated using RGB cameras. Outdoor datasets as KITTI [2,3] and CityScapes [4] represent these challenges addressing critical applications like autonomous driving. Whereas traditional, variational based approaches [5,6], for disparity estimation still have an upper hand in terms of accuracy and regularity of the estimates. The recent applications of the deep learning models have showed promising results in terms of the accuracy, but more importantly, they significantly outperform the traditional methods in terms of computation time, which in that case it is also more predictable. This makes the DL approaches more suitable for real-time implementations. One of the remaining limitations preventing wider applicability of these deep learning methods is the requirement for very large training datasets that include ground truth. That said, the recent advances in DL unsupervised training and a recent proliferation of various representative synthetic datasets makes the used of deep learning methodology a viable alternative for the disparity estimation.

Synthetic datasets provide a convenient starting point to train deep models, e.g. FlyingThings3D [7] dataset. However, models trained with this or similar dataset may not fully represent real scenes, therefore leading to larger errors in practical applications. Training a convolutional neural network (CNN) on synthetic data may not provide the necessary variability and the generalization required to address the problem of the real scenes disparity estimation. Fine-tuning on

Multimodal Sensing: Technologies and Applications, edited by Ettore Stella,
Shahriar Negahdaripour, Dariusz Ceglarek, Christian Möller, Proc. of SPIE Vol. 11059
110590T · © 2019 SPIE · CCC code: 0277-786X/19/$21 · doi: 10.1117/12.2527628

a real dataset may represents an option to improve the model generalization properties. However, the issue with these realistic datasets is that, at best, they provide very limited sparse ground truth, as it is very difficult to obtain dense depth information for real scenes in large quantities. This limitation of the ground truth availability drives the problem to be solved in an unsupervised manner.

Another important problem which often appears when estimating disparity is the occlusion problem. Here, the occluded area between image pairs provide no information about correspondence which leads to ambiguity in the estimation of the disparity within occluded areas effecting the training process and eventually accuracy of the disparity estimation. As the disparity increases in the stereo image, the error introduced due to occlusion increases as well. To address this issue, an occlusion aware loss function is implemented. The occlusion areas are automatically identified for each image pair and the occluded areas do not contribute to the data (fidelity) term of the loss function. The occlusion is estimated for both forward and the backward disparity estimation.

In this work, the focus is to develop an occlusion aware DL model and therefore reduce effect the occlusion has on the accuracy of the loss function calculations during the training. The investigated network implementation is based on the DispNetCorr [7] network. The proposed model is equipped with disparity transformation and bilinear interpolation functions to allow photometric measurement and new Structural Similarity loss function component. It can be trained in both supervised and unsupervised setting by choosing the related loss functions.

Related Work

Convolutional neural networks have proven to be very successful in diverse application areas. For many image computing tasks, such as image classification or segmentation, the CNNs have become a preferred methodology. For other applications, including: direct depth [8], disparity [9], or optical flow [10] estimation, the CNN methods although gaining ground are not yet this dominant. This is despite already outperforming, in some aspects, the more traditional methods.

In [7], authors implemented a disparity estimation using convolutional neural network with supervised learning. They built synthetic dataset, called FlyingThings3D, to facilitate the network training process. Two FlowNet network structures, originally proposed in [10], were adopted. The first one is called DispNetS, it accepts stacked stereo image pair, while the second, called DispNetC (or DispNetCorr), initially uses two branches that meet at a correlation layer forming a single fused encoding branch. The correlation layer is used to aid matching features from two input images. The networks implement multiscale reconstruction (decoding) with explicit supervised loss at reconstruction levels contributing to the overall loss function being minimized during the training process. This arrangement allows the training to be performed at each stage improving network performance.

Other authors proposed methods for depth estimation utilizing single image within a supervised learning framework. In [11] authors suggested a patch-based model where the image is first divided into patches then the image planes orientations and the 3D features are extracted per each patch. The planes parameter representations are calculated using linear predictors then bundled using Markov Random Field MRF to incorporate local cues. The main problem, the authors reported, is the lack of the model to represent thin structures.

Another local approach, described in [12], utilises semantic information to provide better clues for the depth estimation. In [13] authors propose to use nearest neighbours query on the ground truth patches to warp the estimated depth map. However, the disadvantage in this method is that it needs entire training set to be available for the unknown depth estimation, which is not very practical.

Unlike previous studies, the work described in [8] produces dense pixel depth by utilising two scales deep network. The training performed on images and their corresponding depth values use raw pixels values for feature learning. Other studies built on this work by using conditional random field (CRF) to increase accuracy [14] and modify the loss function [15].

In [9], authors employed Siamese network for estimating matching distances between patches obtained from images. In [16] cross-based cost aggregation is employed to estimate disparity using the same concept. The authors in [17] utilized semi-global matching (SGM), however, this approaches does not train the network end-to-end.

A framework of a network consists of extended network that consist of three parts: of multiscale shared features (encoder), initial disparity estimation (decoder), disparity refinement is suggested in [18] that is trained end-to-end. The refinement part of the network is proposed to improve the outcome by smoothing the discontinuities and reject outliers

during the estimation. Disparity refinement performed using feature constancy, by calculating feature correlation, reconstruction error, and the initial disparity estimation.

This research is motivated by the already mentioned work reported in [7] and [10], as well as work reported in [18] where the concepts originally proposed in [7, 10] using supervised learning are extend to the unsupervised approach. In that work, the FlowNetS / FlowNetC architectures (originally proposed in [7] for optical flow estimation) are adopted with modification to compute two disparity maps one for each input image. The corresponding two FlowNets share parameters. This allows simultaneous and synchronized training of the two networks producing compatible features that can represent the forward and backward disparity maps. Subsequently, this allows an estimation for the corresponding forward and backward disparity occlusion masks. An access to the occlusion information, allows for more reliable data fidelity loss computation during training, improving the overall performance of the network.

The details of the proposed in this research network architecture are described in the next section.

2. NETWORK IMPLEMENTATION AND TRAINING

Network Implementation

The work described in this paper has been inspired by the results reported in [19]. The authors proposed there a novel network architecture, called UnFlow, for unsupervised estimation of optical flow from pair of images. Here, a very similar overall architecture have been implemented with a somewhat different encoder-decoder network which was originally proposed for estimation of disparity in the stereo images [7]. Figure 1 shows the graphical representation of the adopted architecture. Overall, the network consists of two channels. The first channel is fed with the two images in the forward order (say left-right), whereas the second channel takes input images in the reversed order (i.e. right-left). However, the model weights are shared across the two channels with leaky activation functions. This arrangement allows for a better features learning because each task influences the other [20], and encourages parameters regularization, while decreased number of the model parameters reduces overfitting [21]. Furthermore, weight regularisation has been also applied to further reduce overfitting.

With two disparity branches, the network can be trained for forward and backward disparity estimation in supervised fashion if a relevant ground truth is available. However, the model can be also trained in an unsupervised fashion by equipping the model with suitable loss function which does not require the ground truth data. For unsupervised training the loss function uses estimated displacements to maximise the photometric consistency (minimise the data fidelity term) between corresponding warped images as well as disparity smoothing to regularise the displacement field. This arrangement also allows estimating the occlusion masks for both the forward and backward disparities. The estimated occlusion masks allow for more accurate data fidelity loss estimation by excluding occluded pixels from loss calculations. To reduce overestimation of the occlusion, the number of estimated occluded pixels is also included in the loss function.

When compared with the original architecture proposed in [19], the network proposed in this paper has been adopted for disparity estimation on rectified images. Furthermore, the Structure Similarity Index Measurement loss function is employed to enhance loss estimation at the image boundaries [22] which is particularly useful especially in the warped images case.

Datasets

In this reported work, three datasets are utilized for training and evaluation, these are: FlyingThings3D [7], KITTI [2,3], and CityScapes [4]. The FlyingThings3D is a large detailed synthetic dataset specifically built to support work on disparity, optical flow and scene flow estimation. The FlyingThings3D stereo RGB images are renderings of scenes consisting of different randomly distributed 3D objects. Generally, the scenes background consists of cylinders and cuboids that varies in scale, texture and orientation. The foreground objects consist of 37927 detailed 3D models from Stanford's ShapeNet dataset [23]. Between five to twenty objects are randomly sampled from these foreground object, textured, resized and rotated along a smooth 3D trajectories with random displacement. The resulting rendered images are available as clean (cleanpass) or as more realistic (finalpass) versions. The latter include motion and depth of field blur effects. In this reported work 22000 stereo image for training and around 4000 for both validation and testing were selected from that dataset. The corresponding forward and backward ground truth disparity data are also available.

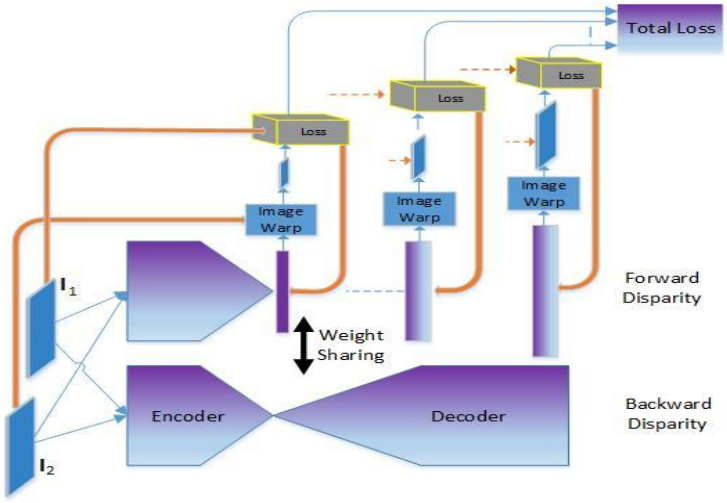

Figure 1. Diagram of the network architecture adopted in the reported work.

The KITTI disparity dataset consists of two 2012 and 2015 subsets. The scenes within these subsets are captured for real-world traffic situations, showing motorways, rural and urban areas. The "KITTI Stereo 2012" consists of about 50K undistorted and rectified image pairs that were collected for the purpose of developing mobile robotics and autonomous driving applications. The images are collections of about 6 hours of real traffic recordings. Moreover, the "KITTI Stereo 2015" set incorporates 200 image pairs with corresponding disparity, semantic segmentation and other gourd truth. In particular the disparity/depth ground truth is collected using Velodyne 3D laser scanner.

The CityScapes is another dataset that is designed to capture street scenes with high variability [4]. The dataset was collected for several months during different seasons and weather in 50 cities. The used RGB camera pair had 22 cm baseline with 2MP sensors and operated at 17 frames per second. The cameras were calibrated and the image pairs rectified. The dataset consists of around 20000 images pairs with forward disparity. In addition to the rectified image pairs and their corresponding data, the dataset includes vehicle odometry data obtained from vehicle sensors and GPS tracks.

The data used for training are further augmented following the same augmentation techniques as proposed in [7], which include random spatial translation, cropping, scaling, as well as colour, brightness and contrast transformations.

Loss Function

For the two stereo rectified RGB images, $I_1, I_2 : P \rightarrow R^3$, the task is to estimate the forward disparity $d^f(x,y)$ between pixel in I_1 and its corresponding pixel in I_2. Whereas the backward disparity $d^b(x,y)$ estimates respectively correspondence between pixels in I_2 and I_1. However, the pixel correspondence is violated in occluded areas since the corresponding part of the scene is not visible in one of the images and therefore correspondence cannot be obtained. Following the methodology proposed in [24], to avoid unreliable error estimations within the occluded areas, occlusion masks are estimated for the forward and backwards disparities. For the forward disparity the forward occlusion mask, $m^f(x,y)$, is estimated using the following inequality:

$$|d^f(x,y)+d^b(x+d^f(x,y),y)|^2 < \alpha_1(|d^f(x,y)|^2+|d^b(x+d^f(x,y),y)|^2)+\alpha_2 \quad (1)$$

Where, at location (x,y) the forward occlusion mask $m^f(x,y)$ has value 1 if the above inequality does not hold and zero otherwise. The backward occlusion mask $m^b(x,y)$ is estimated in a similar manner with forward and backward disparities swapping places in Equation 1. For the reported results, the values of α_1 and α_2 are set to 0.01 and 0.5, respectively.

The overall loss function consists of multiple components, including occlusion-aware data fidelity loss E_D, displacement smoothness constraint E_S, and the Structural Similarity loss E_{SSIM}.

The occlusion-aware data fidelity loss E_D function, similarly to [19], is estimated using photometric consistency assumption in the non-occluded areas, whereas the occluded pixels are not included in the calculation of that loss:

$$E_D = \sum_{(x,y)\in P} (1-m^f(x,y))\cdot g(I_1(x,y)-I_2(x+d^f(x,y),y)+ \\ (1-m^b(x,y))\cdot g(I_1(x+d^b(x,y),y)-I_2(x,y))+\beta_1 m^f(x,y)+\beta_2 m^b(x,y) \tag{2}$$

where: $\beta_1=\beta_2=0.01$, $g(x)=(x^2+\varepsilon^2)^\gamma$ is robust generalized Charbonnier penalty function with $\gamma=0.45$, and $\varepsilon=10^{-7}$; $m^f(x,y)$ and $m^b(x,y)$ are included to control the size of the estimated occlusion areas.

The Structural Similarity loss is defined as:

$$E_{SSIM}=\frac{1}{2}\left(\sum_{(x,y)\in P}(1-SSIM(I_1(x,y),I_2(x+d^f(x,y),y)))+(1-SSIM(I_1(x+d^b(x,y),y),I_2(x,y)))\right) \tag{3}$$

Where SSIM is defined in [22] and is often used to measures perceptual similarity between images. It operates on pixel neighborhood and therefore supplements operation of the E_D loss function.

The smoothness term loss is defined as:

$$E_s=\sum_{(x,y)\in P}|d_x{}^f(x,y)|\cdot\exp(-|I_{2,x}(x,y)|)+|d_y{}^f(x,y)|\cdot\exp(-|I_{2,y}(x,y)|)+ \\ |d_x{}^b(x,y)|\cdot\exp(-|I_{1,x}(x,y)|)+|d_y{}^b(x,y)|\cdot\exp(-|I_{1,y}(x,y)|) \tag{4}$$

where: $A_x=dA/dx$ and $A_y=dA/dy$ are the image derivatives, used as weights encouraging smoothing the disparity fields in areas with uniform image intensities. In the above equations, transformation and interpolation is performed as reported in [25]

The overall loss function E weights all the individual loss function components and is given as:

$$E=\lambda_D E_D+\lambda_{SSIM}E_{SSIM}+\lambda_S E_S \tag{5}$$

where: $\lambda_D=0.9$, $\lambda_{SSIM}=0.1$, $\lambda_S=0.1$

The model shown in Figure 1, trained in supervised scenario using the FlyingThings3D dataset with ground truth using Mean Absolute Error (MAE) loss function between the prediction and the ground truth for both the forward and the backward disparity. Adam optimizer with a starting learning rate of 10^{-4}. Since the model contains multi-scales at the decoding side, at each 5000 iterations one scale is training and the learning rate is reduced by a factor of 1.4. The training is performed for a maximum of 50K iterations. To fine-tune the model for other datasets the loss function of Equation 5 has been utilized for unsupervised training. During this phase, the model is trained using Equation 5 for total of 20K iterations and starting with training rate of 10^{-6}, and 4 images batch size. The dataset during the unsupervised phase included KITTI and CityScape datasets.

In the second scenario the training procedure describe above is employed but without ground truth. The experiment is performed with unsupervised setup (i.e. without a supervised pre-training) and the model trained with Equation 5 only as a loss function for 50K iteration on FlyingThings3D and KITTI separately.

3. RESULTS

Table 1 shows the results obtained using the network architecture and the loss function described in the previous sections. Three metrics have been used for the proposed method evaluation. The End Point Error (EPE) metric is an average absolute error measured between the estimated and the ground truth disparity maps and the two other metrics measure percentage of pixels with the disparity error bigger than 3 pixels (>3P) and 5 pixels (>5P). The proposed

architecture was tested using different combinations of the training/testing data and learning regimes, including supervised, unsupervised and fine-tuning learning scenarios.

The results for five different experiments are reported in Table 1 with the results from each experiment provided in the successive rows of the table. In the first experiment, the proposed network is trained in a supervised manner on the 22,000 synthetic image pairs selected from the FlyingThings3D dataset. The performance of the network is evaluated on different 4000 FlyingThings3D image pairs. The next two experiments use the pertained network from the first experiment and fine-tune the network using unsupervised training on the KITTI and CityScapes datasets. More specifically, the second experiment uses 40,000 image pairs from the 2012 KITTI subset for training, and 200 image pairs with known sparse ground truth from the 2015 KITTI subset for testing. The third test uses 20,000 CityScapes image pairs for training and different 1525 image pairs for testing. The fourth experiment uses the same data for training and testing as the first experiment, but this time the network is entirely trained in an unsupervised way, i.e. without any supervised pre-training. The last experiment uses the same training and test KITTI images as used in the second experiment, but again no supervised pre-training was used.

Compering results from the first and fourth experiments, it can be concluded that supervised learning provides better results than the unsupervised learning. This should be expected as the supervised learning uses much richer information. What is though interesting, when comparing results from the second and fifth experiments, is that a supervised pre-training using a synthetic dataset could somewhat improve results on the real data, even though the nature (properties) of the used synthetic data is very different from the real data. This is an important result as it is relatively easy to generate large synthetics dataset with the ground truth, i.e. suitable for supervised training, whereas it is very difficult or even impossible to obtain ground truth disparity maps for some scenes in real data acquisitions scenarios.

The results reported in the first line of Table 1 are better than corresponding results reported in [7]. The improvement in the performance could be explained due to the occlusion estimation and the shared parameters between the forward and the backward disparity networks. The loss function that the model trained with is the Mean Absolute Error (MAE) or L1 loss function that compares the predicted disparity with the ground truth. In addition, this network trained maximum displacement for the correlation layer of 40 pixels as suggested by [7] since the required displacement in FlyingThings3D dataset is high compare to other datasets. While in the case of KITTI model the maximum displacement parameter was set to 20 pixels.

Some additional work is planned to investigate network generalisation properties with supervised, unsupervised and fine-tuned scenarios. In that case the network is to be trained (fine-tuned) and tested on data from different datasets, including indoor and outdoor cases. Note that the fine-tuning in this context is to train the network without ground truth.

Table 1. Values of different metrics obtained for different training/testing scenarios.

	Dataset	**Training**	**EPE** [pixels]	**>3P** [%]	**>5P** [%]
1	FlyingThings3D	Supervised	1.4	13.43	3.99
2	FlyingThings3D fine-tuned on KITTI	Unsupervised	2.05	9.70	2.14
3	FlyingThings3D fine-tuned on CityScapes	Unsupervised	1.96	23.84	6.61
4	FlyingThings3D	Unsupervised	2.50	25.2	12.24
5	KITTI	Unsupervised	2.02	9.81	2.42

Qualitative results are shown in Figure 2-3. Figure 2, shows results obtained for KITTI dataset. In that case, only limited disparity ground truth is available which additionally is not dense. The pixels of the predicted disparity map are nullified where the corresponding pixels in the ground truth were null. The proposed model can inference the disparity maps with 24 ms (40 fps) on Nvidia Titan Xp and 60ms (16fps) on Nvidia GTX960 which showing real-time performance capability.

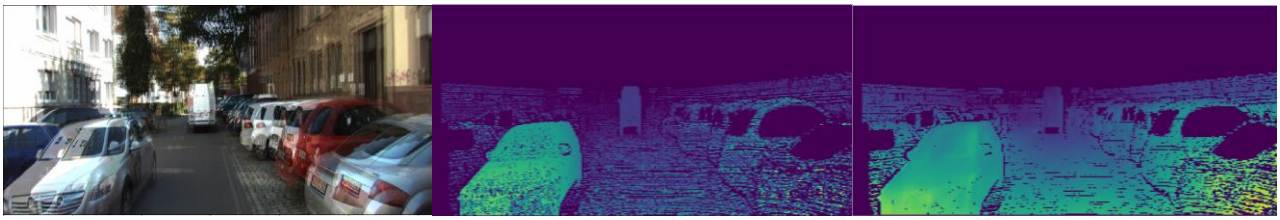

Figure 2. A sample result from the fifth experiment - the KITTI dataset. (Left): Input left-right images, (Middle): Kitti disparity ground truth, (Right): predicted disparity using unsupervised trained.

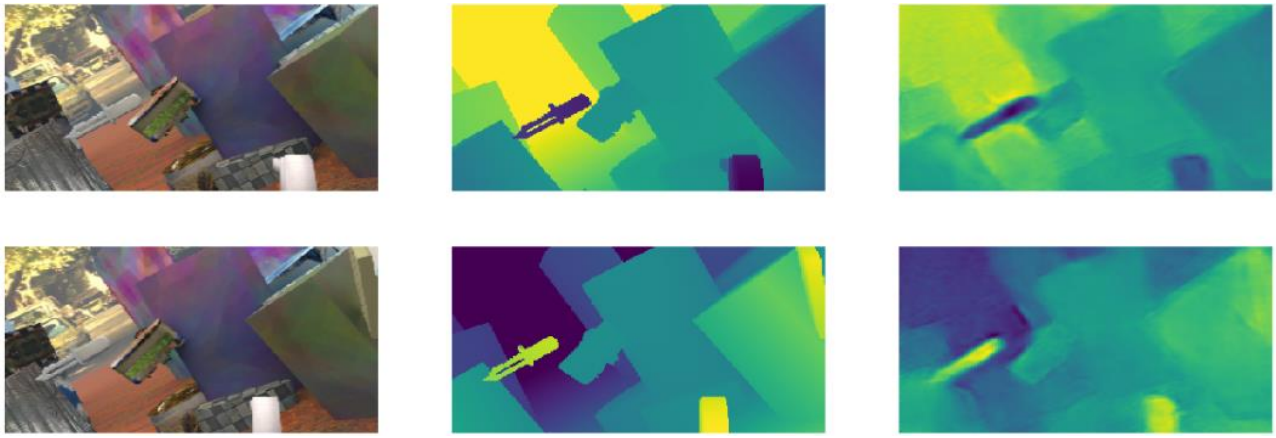

Figure 3. A sample result from the fourth experiment - the FlyingThings3D dataset. (Left): left and right images are shown in the top and bottom rows respectively, (Middle): corresponding forward and backward disparity maps ground truths, (Left): corresponding estimated forward and backward disparity maps using unsupervised training.

4. CONCLUSION

This paper describes a novel deep learning method for disparity estimation from a pair of rectified stereo images. Two previously proposed network architectures have been combined to address the problem of unsupervised network training for prediction of dense disparity maps. The method uses two prediction channels, with corresponding networks sharing weights, for estimation of the forward/backward (left/right) disparities. Using the disparities consistency assumption, forward/backward occlusion masks are also calculated. They are subsequently used to compute more accurately the data fidelity loss function. Two other components of the loss function have been also proposed. One is based on a structural similarity index measure, introduced to encompass higher order photometric similarities between input images. The final component of the loss function is introduced to regularize the disparity maps in the image uniform areas, where there is little information to guide the estimation of the disparity from the image contents alone. The proposed model is trained end-to-end and the results show improved prediction accuracy when tested on three popular dataset, FlyingThings3D, KITTI and CityScapes. Although the network can be trained in a supervised manner, its main advantage is ability to learn the regression model for disparity from data without grand truth being available. It has been shown that the model can achieve EPE of 2.5 pixels when trained without ground truth on the FlyingThings3D dataset, which is close to value of 1.6 pixels, achieved in [7] with the supervised learning. The model also shows improved result for KITTI with 2.05 after fine-tuning.

ACKNOWLEDGMENTS

Contributions to this paper by B.J. Matuszewski were in part supported by the UK EPSRC project EP/K019368/1: “Self-Resilient Reconfigurable Assembly Systems with In-Process Quality Improvement”

REFERENCES

[1] Scharstein D, Hirschmüller H, Kitajima Y, Krathwohl G, Nešić N, Wang X, Westling P. High-resolution stereo datasets with subpixel-accurate ground truth. In German conference on pattern recognition 2014 Sep 2 (pp. 31-42). Springer, Cham.

[2] Menze M, Geiger A. Object scene flow for autonomous vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015 (pp. 3061-3070).

[3] Geiger A , et. al. "KITTI Dataset," Slate, 18 April 2019, http://www.cvlibs.net/datasets/kitti/ (18 April 2019).

[4] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 3213-3223).

[5] Wanner S, Goldluecke B. Variational light field analysis for disparity estimation and super-resolution. IEEE transactions on pattern analysis and machine intelligence. 2014 Mar;36(3):606-19.

[6] Tran TH, Wang Z, Simon S. Variational disparity estimation framework for plenoptic images. In2017 IEEE International Conference on Multimedia and Expo (ICME) 2017 Jul 10 (pp. 1189-1194). IEEE.

[7] Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T. 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 2016 (pp. 4040-4048).

[8] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In Advances in neural information processing systems 2014 (pp. 2366-2374).

[9] Zbontar J, LeCun Y. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. Journal of Machine Learning Research. 2016 Jan 1;17(1-32):2.

[10] Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, Van Der Smagt P, Cremers D, Brox T. Flownet: Learning optical flow with convolutional networks. InProceedings of the IEEE international conference on computer vision 2015 (pp. 2758-2766).

[11] Saxena A, Chung SH, Ng AY. Learning depth from single monocular images. In Advances in neural information processing systems 2006 (pp. 1161-1168).

[12] Ladicky L, Shi J, Pollefeys M. Pulling things out of perspective. In Proceedings of the IEEE conference on computer vision and pattern recognition 2014 (pp. 89-96).

[13] Karsch K, Liu C, Kang SB. Depth transfer: Depth extraction from video using non-parametric sampling. IEEE transactions on pattern analysis and machine intelligence. 2014 Nov 1;36(11):2144-58.

[14] Li B, Shen C, Dai Y, Van Den Hengel A, He M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 1119-1127).

[15] Cao Y, Wu Z, Shen C. Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology. 2018 Nov;28(11):3174-82.

[16] Zhang K, Lu J, Lafruit G. Cross-based local stereo matching using orthogonal integral images. IEEE transactions on circuits and systems for video technology. 2009 Jul;19(7):1073-9.

[17] Hirschmüller H. Accurate and efficient stereo processing by semi-global matching and mutual information. Innull 2005 Jun 20 (pp. 807-814). IEEE.

[18] Liang Z, Feng Y, Guo Y, Liu H, Chen W, Qiao L, Zhou L, Zhang J. Learning for disparity estimation through feature constancy. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 2811-2820).

[19] Meiser S, Hur J, Roth S. UnFlow: Unsupervised Learning of optical flow with a bidirectional census loss. In Thirty Second AAAI Conference on Artificial Intelligence 2018 Apr 27.

[20] Girshick R. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 1440-1448).

[21] Ullrich K, Meeds E, Welling M. Soft weight-sharing for neural network compression. arXiv preprint arXiv:1702.04008. 2017 Feb 13.

[22] Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging. 2017 Mar;3(1):47-57.

[23] Savva M, Chang AX, Hanrahan P. Semantically-enriched 3d models for common-sense knowledge. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2015 (pp. 24-31).

[24] Fleet D, Weiss Y. Optical flow estimation. In Handbook of mathematical models in computer vision 2006 (pp. 237-257). Springer, Boston, MA.
[25] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. In Advances in neural information processing systems 2015 (pp. 2017-2025).