Revisiting Dead Leaves Model: Training With Synthetic Data

Article *in* Signal Processing Letters, IEEE · December 2021

DOI: 10.1109/LSP.2021.3132289

3 authors, including:

Pavan Madhusudana
University of Texas at Austin
20 PUBLICATIONS **297** CITATIONS

SEE PROFILE

Revisiting Dead Leaves Model: Training with Synthetic Data

Pavan C. Madhusudana, Seok-Jun Lee and Hamid R. Sheikh

***Abstract*—Deep neural networks targeting stereo disparity estimation have recently surpassed the performance of hand-crafted traditional models. However, training these networks require large labeled databases for obtaining accurate disparity estimates. In this paper, we address the large data requirement by generating synthetic data using natural image statistics. Images generated using dead leaves model have been shown to share many statistical characteristics commonly seen in natural images. In this work, we created a synthetic dataset using the 3D dead leaves model consisting of occluding spheres, and projected them onto parallel camera planes to obtain stereo image pairs along with ground-truth disparity map. This generated data was subsequently used to train a deep neural network in a supervised manner to estimate disparity. Through experiments we show that this trained model achieves competitive performance across real-world and synthetic stereo datasets, even without any additional fine-tuning. The proposed method for dataset generation is simplistic in nature, computationally inexpensive and can be easily scaled for large scale data generation.**

***Index Terms*—dead leaves model, natural scene statistics, disparity estimation, stereo matching, disparity map**

I. Introduction

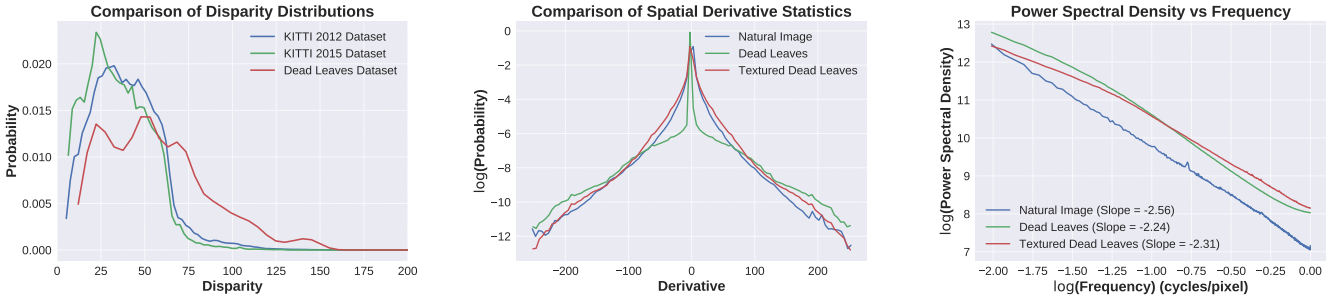
OVER the last decade supervised deep neural networks (DNN) have made significant strides towards achieving human-level performance over a variety of tasks such as image classification [1], [2], object detection, semantic segmentation [3], [4], optical flow estimation [5], [6], network representation learning [7], [8] etc. These deep models typically contain millions of trainable parameters, and require large datasets in order to achieve superior performance. Availability of large labeled datasets such as Imagenet [9], COCO [10], SUN [11] etc., in combination with better computational capabilities have contributed to the successes of these models. However, designing high quality labeled datasets is an expensive and challenging task, and does not scale well due to large amount of time associated with labeling. For certain problems such as disparity/depth estimation, optical flow etc., acquiring training data is a demanding procedure, requiring synchronised capture of images and 3D scene model (e.g.using a laser scanner or structured light sensor), followed by a careful registration of all the acquired data. Additionally, manual post-processing might be required to mask occluded regions, invalid depth estimates or sensor inaccuracies [12].

Here, we consider the task of creating a new synthetic dataset for disparity estimation problem, which has inspirations from natural image statistics. Existing real world stereo datasets [12]–[14] are small in size, lack dense ground-truth disparity, and are limited in their content diversity. One alternative approach that has been explored in the past is to employ synthetic datasets [15]–[17], obtained using computer animation, and rendered using 3D graphics software such as Blender[1]. The contents are designed with the goal of imparting sufficient realism, diversity as well as to be semantically meaningful. Using synthetic data has many obvious advantages over real data: (i) It is inexpensive to generate; ideally infinite amount of data can be generated. (ii) It is easy to obtain ground-truth for synthetic data. DNNs trained using synthetic data [16], [18], [19] has been shown to generalize quite well on realistic datasets, even without fine-tuning. However, computer animation based datasets often require a careful design of scenes in terms of background, objects, shapes, color, amount of texture etc. Here, we investigate whether images obtained from computer animation can be effectively replaced with images generated purely based on natural image statistics. We also explore whether such synthetic images can be used as a reasonable proxy for real-data to train DNNs.

Dead leaves (DL) model, initially proposed by Matheron [20] was motivated by the effect of occlusion of physical objects. In [21] images generated using DL model was shown to exhibit similar statistical regularities as that of natural images. Under the DL model, synthetic images are generated by adding independent shapes such as disks in a layered manner. Recently, in [22], a DNN trained using DL images was observed to achieve impressive performance on image restoration tasks. In this contribution, we extend the original 2D DL model to 3D by replacing disks with occluding spheres. These spheres are then projected onto the camera plane to obtain DL images. Since the location of the spheres is known apriori, depth map for the projected image can be constructed, and used as the ground-truth for disparity estimation. We show that a DNN trained using this generated DL data generalizes surprisingly well on real-world data, even without fine-tuning. To the best of our knowledge, this is first such work employing the 3D DL model for generating synthetic data aimed at disparity estimation. Our proposed method has several advantages: (i) It is simplistic in nature, and does not require any computationally intensive process for generation. (ii) The model is based on statistical properties of natural images and is independent of any computer graphics/animation models. (iii) The proposed model is generic and flexible, can be easily extended to other applications such as optical flow, multiview stereo, depth estimation etc.

P. C. Madhusudana is with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, USA (e-mail: pavancm@utexas.edu). Seok-Jun Lee and Hamid R. Sheikh are with Samsung Research America (SRA). (e-mail: seokjun1.lee@samsung.com; hr.sheikh@samsung.com).This work was done during an internship at SRA by the first author.

[1]https://www.blender.org/



(a) Disparity Distribution (b) Marginal Distribution (c) Power Spectrum

Fig. 1. Statistical comparison of dead leaves dataset. (a) Comparison of disparity distributions across real and dead leaves datasets. (b) Comparison of distributions of spatial derivatives. (c) Plot illustrating the variation of power spectrum with frequency.

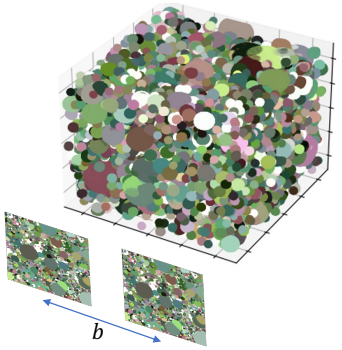
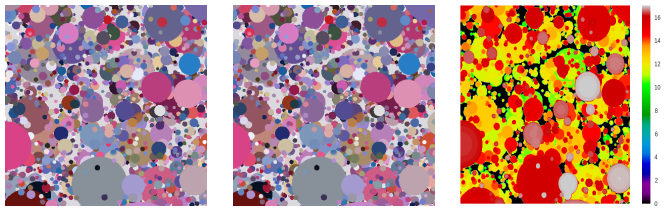
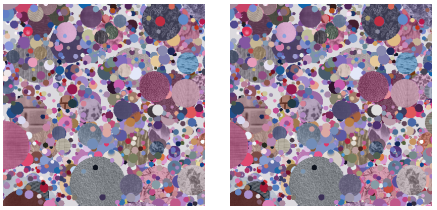


Fig. 2. Illustration of the setup employed for obtaining stereo images from 3D Dead Leaves space.



(a) left image (b) right image (c) disparity map

Fig. 3. Sample dead leaves stereo data



(a) left image (b) right image

Fig. 4. Sample textured dead leaves data. This image pair is obtained by blending textures to stereo data shown in Fig. 3. Viewed better when zoomed.

II. Method

A. 3D Dead Leaves Model

DL model was initially proposed in 2D by adding circular disks in a layered manner. Extending DL model to 3D has been previously investigated in [23], [24] to study the statistics of range and disparity of natural scenes. For constructing a 3D DL space, we follow similar procedure as described for 2D case in [21], [22], whereby circular disks are replaced with spheres. The space was populated with N opaque spheres of random radius and uniformly distributed in space. Intersection between the spheres were allowed. The radii r of the spheres were randomly sampled from distribution $f(r) = Kr^{-3}$, where K is a normalizing constant. This constraint results in images which share many statistical properties of natural images [21]. Additionally, the radii values r were restricted to lie in range $[r_{min}, r_{max}]$ in order to have well defined models [21]. The spheres were assigned colors by randomly sampling from histograms of natural images [22] (for different scenes different natural images were employed). This was done to ensure that the resulting color distribution was similar to that of natural images.

The colored dead leaves space is then projected onto the two parallel camera planes as illustrated in Fig. 2 to obtain stereo image pairs. For simplicity we assume a pin-hole camera model, and the stereo camera planes to be perfectly parallel. Parallel camera planes eliminate the need for image rectification. Thus, disparity d at pixel (x, y) in the left image can be calculated as $d(x, y) = \frac{fb}{D(x,y)}$ where f is camera's focal length, b is the baseline width between camera centres and $D(x, y)$ is the depth value corresponding to pixel (x, y) in the left image. A sample stereo data along with ground-truth disparity map is shown in Fig. 3. By varying focal length f and baseline width b, stereo image pairs spanning wide disparity levels can be generated. Our motive was to replicate disparity levels that are generally observed in real world scenes. The resulting distribution of disparity values is compared against that of real world datasets KITTI 2012 and KITTI 2015 in Fig. 1a. From the figure it may be observed that the DL dataset has approximately similar disparity distribution as the KITTI datasets, implying that the chosen range of f and b leads to realistic disparity levels.

B. Textured Dead Leaves

We enhance the original DL model by introducing texture to the shapes of DL model. We refer this modified version as the Textured DL model. Adding texture has several

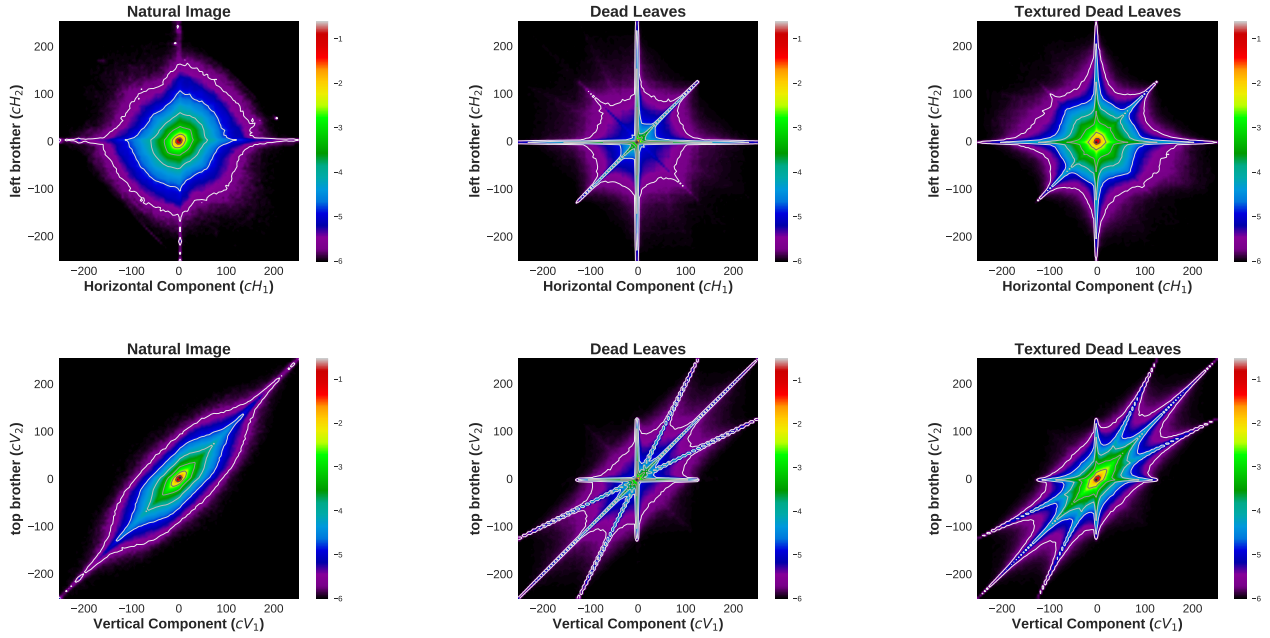


Fig. 5. Comparison of joint statistics (derivative) of natural and dead leaves images. Plots show the log(probability) distributions for different neighboring coefficients. Top row contains horizontal neighbors and bottom row vertical neighbors. Texture addition makes dead leaves distributions to be closer to that of natural images.

advantages: (i) It improves image gradients, particularly in smooth regions. (ii) It reduces ambiguity when learning the mapping between stereo images and corresponding disparity map. (iii) The statistics of the resulting synthetic images have more closer resemblance to natural images than that of the original DL model. Additionally, we also show in Sec. III-D that employing textured DL data leads to better disparity estimates. The textures were chosen randomly from a texture database and blended separately for each DL disk present in the generated image. Brodatz texture database [25] consisting of 112 different patterns was employed as the texture database. For blending, we used alpha blending with equal weightage ($\alpha = 0.5$) to texture and background color. Note that, addition of textures does not change the disparity map as depth values remain unchanged. We also ensured that same texture patterns were used for corresponding disks present in the left and right images. A sample textured DL data is shown in Fig. 4.

C. Significance of Dead Leaves Model

Our primary goal was to generate synthetic disparity data with similar statistical characteristics as that of natural images. The DL model has been extensively studied in [21], and was observed to have considerable overlap in terms of statistical properties with natural images. In particular, the marginal and bivariate distributions of linear filtered natural and DL images were observed to be very similar. Additionally, the power spectrum of DL images exhibited inverse square variation with frequency, commonly seen in natural scenes [26], [27]. These favorable characteristics coupled with computationally inexpensive generation process make the DL model a good candidate for obtaining synthetic data. Although the properties studied in [21] were for the 2D DL model, similar behavior was observed for projected images obtained from the 3D DL model. This is illustrated in Figs. 1b and 5, where we compare marginal and joint statistics of spatial derivatives respectively. From the plots it may be observed that the addition of textures make distributions more closer to that of natural images. The variation of power spectrum with frequency is shown in Fig. 1c, and all compared datasets approximately have a slope close to -2, which is analogous to a $1/f^2$ falloff.

TABLE I
PERFORMANCE COMPARISON OF THE DIFFERENT TRAININGS OF PSMNET. EVALUATION IS PERFORMED ON KITTI [12], [28] AND SCENE FLOW [17] DATASETS. VALUES REPRESENT AVERAGE END POINT ERRORS (EPE). FT* DENOTES FINETUNING ON KITTI 2015 DATASET

Training Dataset	KITTI 2012 [12]	KITTI 2015 [14]	Scene Flow [17]	
			Train	Test
Scene Flow	1.35	1.83	-	1.09
Scene Flow + ft*	0.96	-	7.69	6.76
Dead Leaves	3.01	3.14	13.26	11.52
Textured Dead Leaves	3.38	2.29	9.97	8.3

III. EXPERIMENTS

A. Dataset Generation

Using $N = 20,000$ spheres for each scene, we generated a total of 480 scenes of 1024×1024 resolution using the 3D DL model. The colors for the spheres were sampled from natural images present in the Waterloo database [29]. The images were rendered using PyTorch3D [30] framework employing Pulsar renderer [31]. For each scene, we used 3 focal length values and 9 baseline widths to obtain a total of 27 stereo image pairs, and corresponding disparity maps. Given the 480 scene contents, we arrived at $480 \times 27 = 12,960$ stereo image pairs, which constituted the entire generated DL dataset.

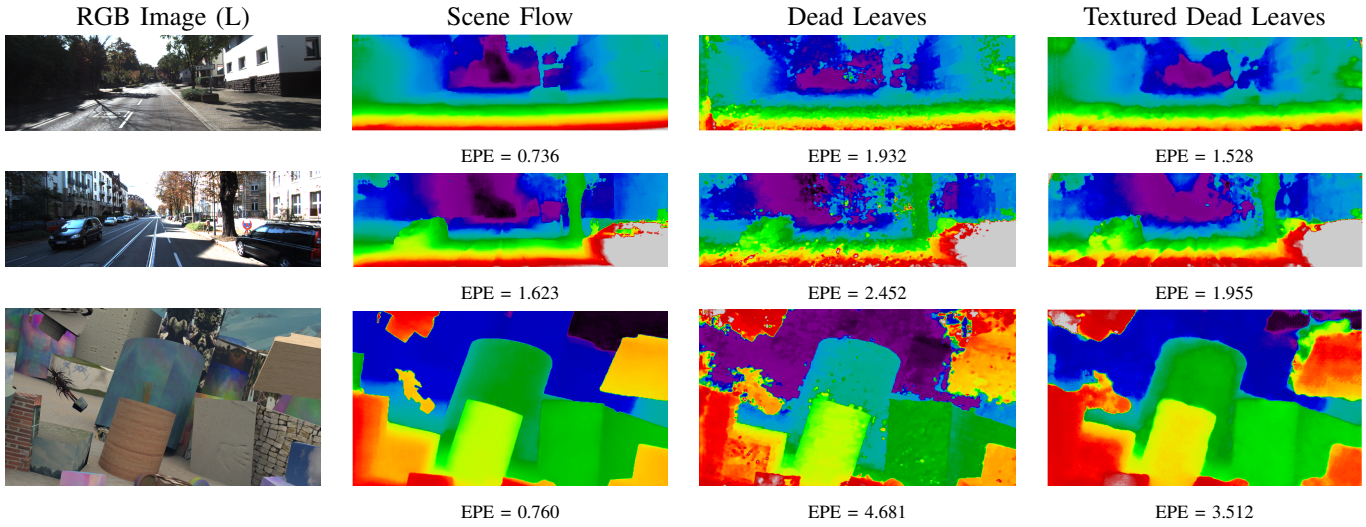


Fig. 6. Comparison of disparity predictions across different datasets. Rows from top to bottom: KITTI 2012, KITTI 2015 and Scene Flow. For each predicted disparity map, the corresponding EPE value is reported at the bottom. It can be observed that the addition of texture to DL images makes the disparity estimates more smoother and noise free.

B. Training Details

In order to evaluate the proposed synthetic dataset, we chose Pyramid Stereo Matching Network (PSMNet) [18] architecture for disparity estimation. PSMNet is a state-of-the-art network for disparity estimation which uses spatial pyramid pooling to incorporate feature maps at multiple scales followed by a 3D Convolutional Neural Network (CNN) for regularization. PSMNet was trained from scratch using the generated DL dataset for 3250 iterations with a batch size of 12. During training, the images were randomly cropped to 256×512 size. Adam [32] optimizer with a learning rate of 0.0001 (β_1 = 0.9, β_2 = 0.999) was used to train the network.

C. Datasets

We evaluated the performance of PSMNet trained using DL dataset on three datasets : KITTI 2012 [12], KITTI 2015 [14] and Scene Flow [17]. KITTI datasets contain real-world data captured from a driving car, while Scene Flow consists of synthetic images obtained from computer animation. All the images were evaluated at their original resolution in RGB color space. Although our primary goal was to obtain better performance on real-world data, we evaluated on Scene Flow data to analyze the generalizability of DL dataset on synthetic animated data. Predicted disparity maps were evaluated using average End Point Error (EPE) given by $EPE = \frac{1}{M}\sum_{i=1}^{M} |disp^i_{pred} - disp^i_{GT}|$, where M is the total number of pixels present in the image, $disp^i_{pred}$ and $disp^i_{GT}$ are predicted and ground truth disparity at pixel i respectively.

D. Performance Comparison

For comparison purposes we included two additional models : (a) PSMNet trained on Scene Flow, (b) PSMNet trained on Scene Flow and fine-tuned on KITTI 2015. Both these models were trained as described in [18]. The performance of PSMNet under different trainings is compared in Table I where average EPE values are reported, and lower EPE values denote superior disparity estimates. The values in Table I for KITTI datasets correspond to the training set of the publicly available data, while for Scene Flow we report performance on both training and testing sets. From the Table, the effect of adding textures to DL data is clearly visible as it significantly boosts performance, especially on KITTI 2015 and Scene Flow datasets. Note that, the values reported in Table I were obtained by training only on the DL dataset with no additional fine-tuning. Thus, it will be hard to match the performance of those obtained by training/fine-tuning using well-supplied real/animation data.

The predicted disparity maps are visually compared in Fig. 6. The impact of using texture on DL images is evident from the Fig. 6, where training with textured DL data leads to smoother and noise free estimates when compared to pure DL data. It may also be observed from Fig. 6 that the estimates for KITTI 2015 data obtained from Scene Flow and textured DL trained models are visually closer when compared to pure DL trained model, reinforcing the observations made in Table I. Note that the disparity prediction for Scene Flow trained model in third row of Fig. 6 (containing Scene flow test image) has very low EPE since the model was trained on Scene Flow training set.

IV. Conclusion and Future Work

We presented an image statistics based 3D dead leaves model for generating synthetic data for disparity estimation problem. We evaluated the generated dataset by using it as a proxy for real data during training. The trained model showed good generalizability, and achieved competitive performance on real-world as well as synthetic stereo datasets.

For 2D projection from 3D DL model, a simple pin-hole camera was used, which can be restrictive as real-world cameras differ significantly from a pin-hole camera. As part of future work we plan to incorporate effects observed in real camera like lens blur, shot noise etc. which can result in statistics more closer to natural images.

REFERENCES

[1] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[5] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[6] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[7] G. Xue, M. Zhong, J. Li, J. Chen, C. Zhai, and R. Kong, "Dynamic network embedding survey," *arXiv preprint arXiv:2103.15447*, 2021.

[8] J. Chen, M. Zhong, J. Li, D. Wang, T. Qian, and H. Tu, "Effective deep attributed network representation learning with topology adapted smoothing," *IEEE Transactions on Cybernetics*, 2021.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[11] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.

[12] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[13] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*. Springer, 2014, pp. 31–42.

[14] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3061–3070.

[15] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European conference on computer vision*. Springer, 2012, pp. 611–625.

[16] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[17] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[18] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[19] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 185–194.

[20] G. Matheron, "Mod'ele s´equentiel de partition al´eatoire," *Centre de Morphologie Mathématique*, 1968.

[21] A. B. Lee, D. Mumford, and J. Huang, "Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model," *International Journal of Computer Vision*, vol. 41, no. 1, pp. 35–59, 2001.

[22] R. Achddou, Y. Gousseau, and S. Ladjal, "Synthetic images as a regularity prior for image restoration neural networks," in *Eighth International Conference on Scale Space and Variational Methods in Computer Vision (SSVM).*, 2021.

[23] Z. Yang and D. Purves, "A statistical explanation of visual space," *Nature neuroscience*, vol. 6, no. 6, pp. 632–640, 2003.

[24] P. B. Hibbard, "A statistical model of binocular disparity," *Visual Cognition*, vol. 15, no. 2, pp. 149–165, 2007.

[25] P. Brodatz, *Textures: a photographic album for artists and designers, by Phil Brodatz*. Dover publications, 1966.

[26] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer. A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[27] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[28] M. Menze, C. Heipke, and A. Geiger, "Joint 3D estimation of vehicles and scene flow," *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, vol. 2, p. 427, 2015.

[29] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2016.

[30] J. Johnson, N. Ravi, J. Reizenstein, D. Novotny, S. Tulsiani, C. Lassner, and S. Branson, "Accelerating 3d deep learning with pytorch3d," in *SIGGRAPH Asia 2020 Courses*. Association for Computing Machinery, 2020.

[31] C. Lassner and M. Zollhofer, "Pulsar: Efficient sphere-based neural rendering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 1440–1449.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int'l Conf. Learning Representations*, pp. 1–15, 2015.