Light-Weight EPINET Architecture for Fast Light Field Disparity Estimation

Ali Hassan
Department of Information Systems and Technology
Mid Sweden University
Sundsvall, Sweden
ali.hassan@miun.se

Mårten Sjöström
Department of Information Systems and Technology
Mid Sweden University
Sundsvall, Sweden
marten.sjostrom@miun.se

Tingting Zhang
Department of Information Systems and Technology
Mid Sweden University
Sundsvall, Sweden
tingting.zhang@miun.se

Karen Egiazarian
Faculty of Information Technology and Communication Sciences
Tampere University
Tampere, Finland
karen.eguiazarian@tuni.fi

***Abstract*—Recent deep learning-based light field disparity estimation algorithms require millions of parameters, which demand high computational cost and limit the model deployment. In this paper, an investigation is carried out to analyze the effect of depthwise separable convolution and ghost modules on state-of-the-art EPINET architecture for disparity estimation. Based on this investigation, four convolutional blocks are proposed to make the EPINET architecture a fast and light-weight network for disparity estimation. The experimental results exhibit that the proposed convolutional blocks have significantly reduced the computational cost of EPINET architecture by up to a factor of 3.89, while achieving comparable disparity maps on HCI Benchmark dataset.**

***Index Terms*—Light Field, Deep Learning, Disparity Estimation, Compression, Depthwise Separable Convolution**

I. Introduction

Recent advancement in image acquisition devices has enabled users to capture the spatial and angular information of the scene, known as Light Field (LF) [1]–[3]. The LF data enables the freedom of viewpoint selection, focal plane changing, and object refocusing in the entire captured scene. Hence, the light field can provide an immersive experience in multimedia applications [4] such as 3D modeling [5], medical imaging [6], telemedicine [7] and movies [8].

With the rise of consumer-level LF cameras [2], [3], disparity estimation from LF has become a promising way to find the pixels in the multiple views that correspond to the same 3D point in the scene. Numerous post-processing applications [4], [8] utilizing LF data rely on disparity information [10] of the scene, e.g. view synthesis [11] and super-resolution [12]. In recent studies, many deep learning (DL) based algorithms are proposed and have achieved significant improvement in the estimation of disparity information [13]–[15]. Convolutional layers in DL architectures plays a vital role in feature extractions from input data, using 2D and 3D convolution operations.

The work was supported by the European Joint Doctoral Programme on Plenoptic Imaging (PLENOPTIMA) through the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 956770.

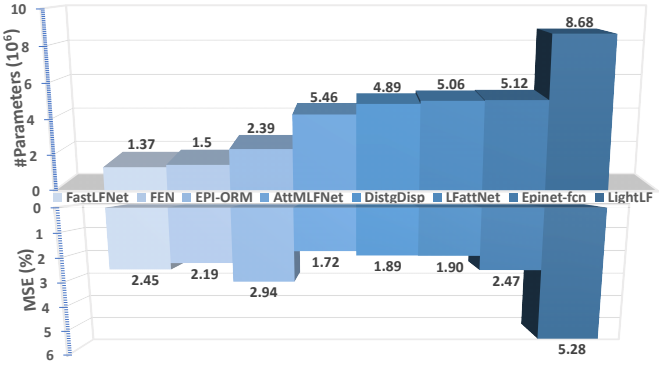
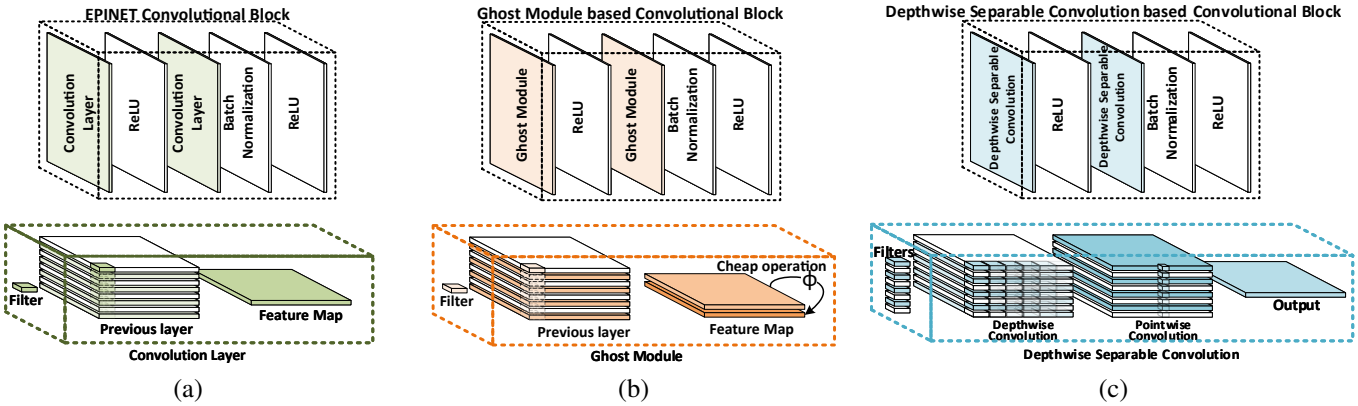


Fig. 1. Comparison of number of parameters in recent deep learning architectures for light field disparity estimation along with their average mean square error (MSE) [9]

3D convolutions have shown promising performance in the disparity estimation task. Alperovich *et al.* [13] propose a fully convolutional autoencoder using 3D convolutions for disparity estimation of light fields. Similarly, Tsai *et al.* [14] and Chen *et al.* [15] utilize a mixture of 3D and 2D convolutional layers in learning-based disparity estimation networks. Although the 3D convolution extracts spatio-temporal information and is beneficial for volumetric data, it is computationally expensive and leads to significant memory consumption [16]. Therefore, to avoid memory burden, recent DL architectures instead employed 2D convolutions.

2D convolutional layers are computationally less costly than 3D convolutional layers. Shi *et al.* [19] proposed a disparity estimation algorithm that uses 2D convolution operations to extract features information. Li *et al.* [20] proposed an end-to-end fully convolutional network to estimate the depth value of the intersection point on the horizontal and vertical Epipolar



(a) (b) (c)

Fig. 2. Illustration of convolutional blocks for EPINET architecture. (a) The state-of-the-art convolutional block of EPINET architecture, which extract the features information using ordinary convolution operation [10] (b) The proposed Ghost Module based Convolutional Block, in which the half feature maps are extracted using ordinary convolution operation and remaining feature maps are generated using linear operation [17] (c) The proposed Depthwise Separable Convolution based Convolutional Block, which is composed of depthwise convolution operation, followed by pointwise convolution operation [18].

Plane Images (EPIs). Leistner *et al.* [21] proposed encoder-decoder based U-shaped network architectures to extract geometric and disparity information from LF images. Although these architectures use the less costly 2D convolutional layers for feature extraction, standard convolutional operations are still computationally expensive [17], [18].

The majority of recent successful DL-based disparity estimation architectures require millions of parameters, as illustrated in Figure 1. This makes them energy, computation, and memory intensive. As a result, such architectures require long inference time and power-consuming computational resources (e.g., GPU). Typically, the file size of DNN models are enormous (e.g., more than 200 Megabytes [22]), which limit their deployment. For these reasons, compression of DL architecture is essential for training and deployment of such models in practice.

Several strategies have been proposed to reduce the complexity of spatial convolutions [23]. Howard *et al.* [18] adopted group convolutions and depthwise separable convolutions as alternatives to standard spatial convolutions, which require less computational cost for image classification problems. Similarly, Han *et al.* [17] proposed the Ghost module to generate more features map by applying linear operations on previously extracted feature information. As a result, the required number of parameters and the computational complexity are decreased without changing the size of the output feature map. Although these compression techniques can significantly reduce the complexity of DL networks designed for tasks such as classification, detection, and segmentation [17], [18], [24], their impact on LF rendering and disparity estimation algorithms has not been evaluated.

The state-of-the-art disparity estimation architecture (EPINET), proposed by Shin *et al.* [10] is the focus of our study. It is composed of four multi-stream networks, where each stream is used to extract feature maps in a specific angular direction of LF images using ordinary convolutional layers. The approach achieved state-of-the-art results on the HCI 4D Light Field Benchmark [9]. The EPINET architecture consists of 20 convolutional blocks, where each block consists of 2 ordinary convolutional layers as shown in Figure 2a. As a result, 40 ordinary convolution layers yield massive multiplication operation and requires 5.13 million parameters [10] to predict a disparity map. Since one convolutional block is repeated across the entire EPINET architecture, compressing it will result in a significant reduction in computational cost.

In this paper, an investigation is carried out to analyze the effect of ghost modules and depthwise separable convolution on EPINET architecture. Based on this investigation, following main contributions are made:

1) Four different convolutional blocks are proposed to make EPINET a fast and lightweight architecture for LF disparity estimation.
2) The proposed convolutional blocks have significantly reduced the model parameters, computational cost, and inference time.
3) The compressed architecture estimate the disparity map with a slight improvement compared to the original architecture on the HCI Benchmark dataset [9].

II. Proposed Method

Inspired by the achievement of significant complexity reduction of GhostNet [17] and MobileNet [18], two convolutional blocks are proposed to minimize the computational cost of EPINET architecture for fast disparity estimation. The proposed ghost module (GM) based convolutional blocks and depthwise separable convolution (DWSC) based convolutional blocks are discussed in Sections II-A and II-B respectively.

A. GM-based Convolutional Block

In order to reduce the complexity of EPINET architecture, the ordinary convolution layers in the convolutional blocks are replaced with the ghost feature map extraction module [17]. The theoretical analysis of the ghost module is available in

[17]. Since our objective is to minimize the model parameters, half of the convolutional features are generated using ordinary convolution operations, and the remaining features are produced using simple learning operations. As a result, the incorporation of GM in EPINET convolutional block should reduce the parameters and computational cost by up to 2 times. In Figure 2b, the top image shows the proposed GM-based convolutional blocks, and the bottom image shows the feature extraction process of the GM.

B. DWSC-based Convolutional Block

DWSC is a mixture of Depthwise Convolutions and Pointwise Convolutions [18]. In Depthwise Convolution, a single filter is applied to each input channel to extract features, followed by a point convolution operation, which computes the linear combination of these extracted features using 1×1 convolutions across the channel. In the proposed DWSC-based Convolutional Block, the ordinary convolutional layers are replaced with DWSC to compress the EPINET architecture. The theoretical analysis of the DWSC is available in [18]. In Figure 2c, the block on the top of the figure shows the proposed DWSC-based Convolutional Block, whereas the bottom side of the figure shows the process of the Depthwise and Pointwise convolution operation.

III. Experimental Setup

Dataset: In order to train and evaluate the performance of the proposed convolutional blocks-based EPINET architecture, HCI Benchmark dataset [9] is used with the same configuration as in [10]. This dataset has a spatial resolution of 512×512 and angular resolution of 9×9. The grayscale LF images are fed into the proposed compressed variants of EPINET architecture, to predict a disparity map as an output [10].

Evaluation Metrics: For the quantitative evaluation, the predicted disparity maps are compared with the ground truth of the HCI test data set. The bad pixel ratio (BPR) [9] with three thresholds (0.01, 0.03 and 0.07 pixels) and mean square errors (MSE) [9] are used to evaluate the performance of the proposed EPINET variant. For computational complexity measurement, the Tensorflow [25] profiler library is used to measure Floating Point Operations per second (FLOPS). The average prediction time of test sets is also reported as the actual inference time.

Implementation Details: In [10], the EPINET architecture with 7×7 input views and data augmentation technique (Epinet-fcn) achieved the least MSE among the other variants of EPINET having 3×3, 5×5 and 9×9 input. Therefore, the proposed convolutional blocks are incorporated in Epinet-fcn architecture and named Ghost Module based EPINET (EPINET-G) and DWSC based EPINET (EPINET-D). The number of iterations (10 Million), batch size (16), loss function (Mean Average Error [26]), optimizer (Root Mean Squared Propagation [27]), and learning rate (LR) are the same as stated in [10]. It is not clear how and when LR is decreased in [10], so we reduced the LR from 10^{-5} to 10^{-6} at 5 million iterations.

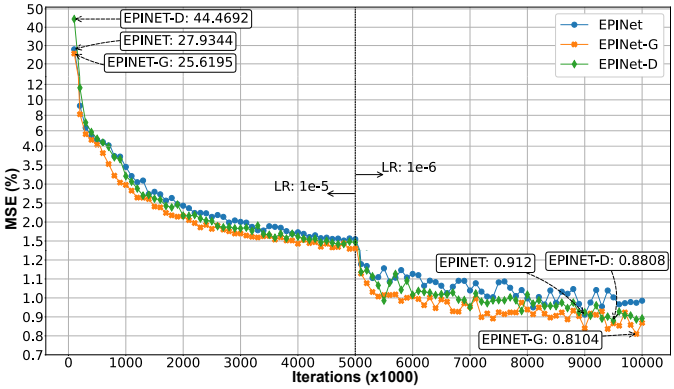


Fig. 3. Model training performance on 10M Iterations. The y-axis uses customized scaling for better visibility of MSE, and the x-axis reports the number of iterations.

Network Training: The network training and evaluation are performed on hardware equipped with two NVIDIA GeForce RTX 2070 graphics cards and 64GB of Random Access Memory (RAM) with Ubuntu 18.04 operating system. Since EPINET source code is publicly available (https://github.com/rgmueller/EPINET-tf2) and written using Tensorflow library [25], the proposed convolutional blocks are also implemented under the same library. The training results are presented in Figure 3. It can be seen that the EPINET-D starts with higher MSE but converges and aligns with the other variant of EPINET after 8 Million iterations, and becomes stable at the end of training.

IV. Results

Since the scope of this work is to investigate to what extent the computational complexity of EPINET architecture can be reduced while preserving its performance, the comparison of the proposed architectures is made with the uncompressed original architecture only. The qualitative comparisons are presented in Figure 4. It can be seen that the EPINET-D and EPINET predict disparity map close to ground truth with comparable MSE and BPR, whereas EPINET-G has high MSE and BPR in the output disparity map. It is reported that the ghost operation can reduce the computational cost with performance degradation and affect the ranking of their architecture [28]. The redundancy in the feature maps generated using ghost operation is limiting the representational ability of convolutional layers [29], hence GM is not very suitable for such application. It is clear from the analysis that there is a trade-off between computational cost and MSE using EPINET-G architecture. Although the ghost operation does not perform well, the research findings are presented for readers to understand the efficacy of these approaches for light field disparity estimation.

Based on the comparable performance of MSE using DWSC, two additional variants of the EPINET-D are also considered in the analysis. The first convolution layer in the

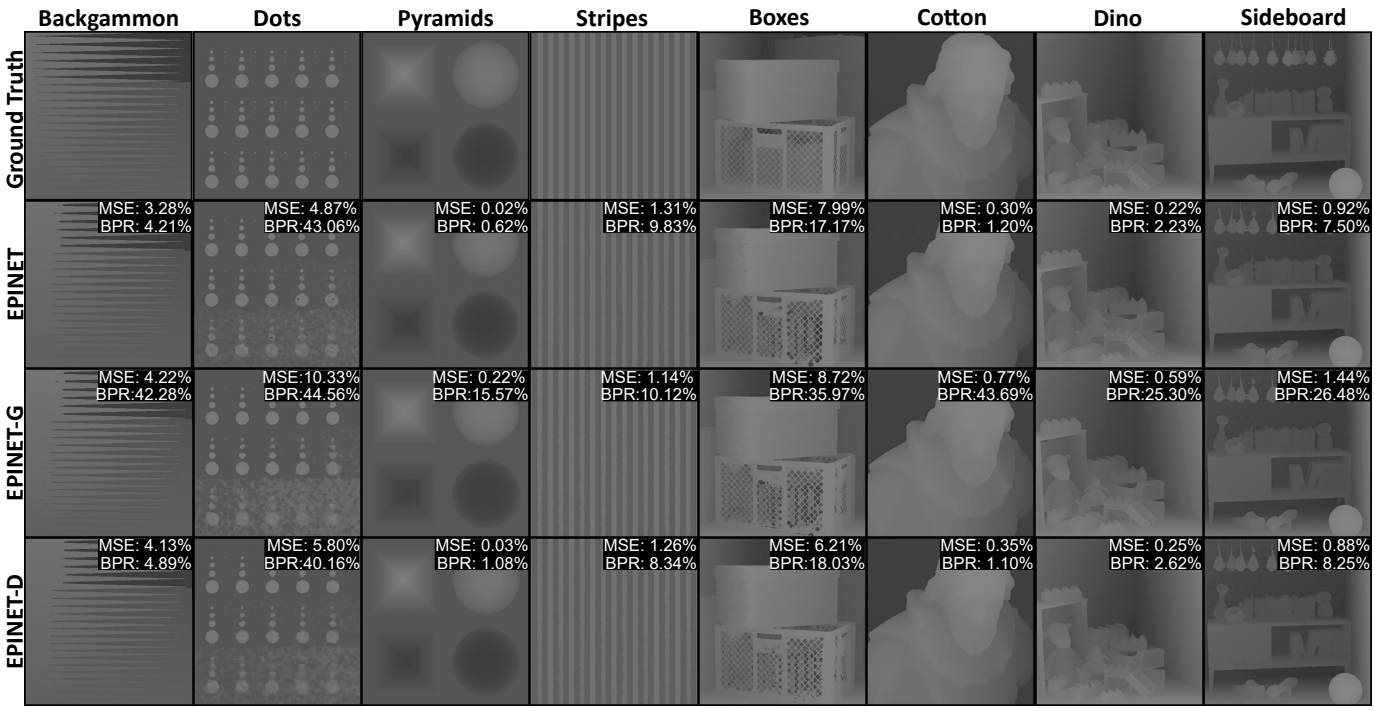


Fig. 4. Qualitative comparison of proposed variants of EPINET architecture. The value on each disparity map report MSE and BPR (Error<0.7) with respect to ground truth disparity map.

TABLE I
QUANTITATIVE EVALUATION OF THE PROPOSED VARIANT OF EPINET ARCHITECTURE ON HCI TEST DATASET*.

Model	Parameters	FLOPs	BPR (%)			MSE	Inference Time	Compression
	Millions (M)	Giga (G)	<0.01	<0.03	<0.07	(%)	Second (s)	Times (×)
EPINET [10]	*5.1243*	*2544.4052*	62.0530	23.5477	7.5029	0.9249	*3.2135*	-
EPINET-G	2.5771	1278.3409	*88.1097*	*59.5703*	*26.4755*	*1.4444*	1.5999	1.9884
EPINET-DC	3.1289	1552.0681	**57.0940**	22.8431	8.6336	0.8467	1.5941	*1.6377*
EPINET-CD	3.0786	1529.4374	57.7908	**21.7012**	**7.1723**	**0.8286**	2.5471	1.6645
EPINET-D	**1.3173**	**649.9651**	61.8916	25.4563	8.2497	0.8777	**0.8947**	**3.8901**

*In Table I, the **bold** numbers shows the best value and *italic* number shows the least value among the column.

reference EPINET convolutional block is replaced with a DWSC to propose the EPINET-DC variant. Similarly, another combination is created by replacing the second convolutional layer of reference EPINET with DWSC, and it is referred to as EPINET-CD architecture. Based on these two combinations, the resulting EPINET variants is trained using same training configurations [10] and their evaluation results are presented in Table I. It can be seen that the EPINET-CD and EPINET-DC have similar number of parameters and FLOPs, but they have small MSE and BPR compared to EPINET, EPINET-D and EPINET-G variants.

The DWSC based EPINET-D architecture outperforms the other variants in terms of parameters, FLOPS, and inference time, and has an MSE in par with EPINET, EPINET-CD and EPINET-DC. Since each channel of the input stack represents one grayscale LF image, depthwise convolution can extract useful feature maps on each channel before applying point-wise convolution to mix information across the feature maps. Therefore, DWSC can play a vital role in feature extraction modules for the LF disparity estimation.

V. CONCLUSION

In this paper, four convolutional blocks were proposed to make EPINET a fast and lightweight architecture for disparity estimations. The evaluation results state that the depthwise separable convolution plays a vital role in the features extraction module, which leads to significant computational cost reduction up to 3.89 times, even with slight quality improvement of the disparity map. In contrast, the ghost module based EPINET-G architecture was unable to estimate competitive disparity map with respect to EPINET. In future work, further compression methods are lined up for evaluation for disparity estimation. It is also planned to further test the effectiveness of depthwise separable convolution in other LF applications.

REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 31–42.

[2] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Ph.D. dissertation, Stanford University, 2005.

[3] R. GmbH, "Raytrix: 3d light field camera technology," https://raytrix.de/, 2016, [Online] (Accessed: 2022-01-30).

[4] M. Martínez-Corral and B. Javidi, "Fundamentals of 3d imaging and displays: a tutorial on integral imaging, light-field, and plenoptic systems," *Adv. Opt. Photon.*, vol. 10, no. 3, pp. 512–566, Sep 2018. [Online]. Available: http://opg.optica.org/aop/abstract.cfm?URI=aop-10-3-512

[5] C. Perra, F. Murgia, and D. Giusto, "An analysis of 3d point cloud reconstruction from light field images," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2016, pp. 1–6.

[6] Z. Wang, L. Zhu, H. Zhang, G. Li, C. Yi, Y. Li, Y. Yang, Y. Ding, M. Zhen, S. Gao *et al.*, "Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning," *Nature methods*, vol. 18, no. 5, pp. 551–556, 2021.

[7] G. Wang, W. Xiang, and M. Pickering, "A cross-platform solution for light field based 3d telemedicine," *Computer Methods and Programs in Biomedicine*, vol. 125, pp. 103–116, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260715002953

[8] S. Zhou, T. Zhu, K. Shi, Y. Li, W. Zheng, and J. Yong, "Review of light field technologies," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, 2021. [Online]. Available: https://doi.org/10.1186/s42492-021-00096-8

[9] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Computer Vision – ACCV 2016*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham: Springer International Publishing, 2017, pp. 19–34.

[10] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] W. Zhou, G. Liu, J. Shi, H. Zhang, and G. Dai, "Depth-guided view synthesis for light field reconstruction from a single image," *Image and Vision Computing*, vol. 95, p. 103874, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885620300068

[12] X. Liu, M. Wang, A. Wang, X. Hua, and S. Liu, "Depth-guided learning light field angular super-resolution with edge-aware inpainting," *The Visual Computer*, pp. 1–13, 2021.

[13] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Light field intrinsics with a deep encoder-decoder network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, "Attention-based view selection networks for light-field disparity estimation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 095–12 103, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6888

[15] J. Chen, S. Zhang, and Y. Lin, "Attention-based multi-level fusion network for light field depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 1009–1017.

[16] S. Mittal and Vibhu, "A survey of accelerator architectures for 3d convolution neural networks," *Journal of Systems Architecture*, vol. 115, p. 102041, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1383762121000400

[17] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1580–1589.

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[19] J. Shi, X. Jiang, and C. Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5867–5880, 2019.

[20] K. Li, J. Zhang, R. Sun, X. Zhang, and J. Gao, "Epi-based oriented relation networks for light field depth estimation," 2020.

[21] T. Leistner, H. Schilling, R. Mackowiak, S. Gumhold, and C. Rother, "Learning to think outside the box: Wide-baseline light field depth estimation with epi-shift," in *International Conference on 3D Vision (3DV)*, sep 2019.

[22] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," *arXiv preprint arXiv:1412.6115*, 2014.

[23] L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485–532, 2020.

[24] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," 2018.

[25] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous systems," 2015.

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[27] T. Tieleman, G. Hinton *et al.*, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[28] W. Jia, J. Gao, W. Xia, Y. Zhao, H. Min, and J.-T. Lu, "A performance evaluation of classic convolutional neural networks for 2d and 3d palmprint and palm vein recognition," *International Journal of Automation and Computing*, vol. 18, no. 1, pp. 18–44, 2021.

[29] A. He, T. Li, N. Li, K. Wang, and H. Fu, "Cabnet: Category attention block for imbalanced diabetic retinopathy grading," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 143–153, 2021.