# ContextNet: Learning Context Information for Texture-less Light Field Depth Estimation

Wentao Chao, Xuechun Wang, Yiming Kan, Fuqing Duan ✉

School of Artificial Intelligence, Beijing Normal University, Beijing, China
{chaowentao, wangxuechun, 202221081044}@mail.bnu.edu.cn
fqduan@bnu.edu.cn

**Abstract.** Depth estimation in texture-less regions of the light field is an important research direction. However, there are few existing methods dedicated to this issue. We find that context information is significantly crucial for depth estimation in texture-less regions. In this paper, we propose a simple yet effective method called ContextNet for texture-less light field depth estimation by learning context information. Specifically, we aim to enlarge the receptive field of feature extraction by using dilated convolutions and increasing the training patch size. Moreover, we design the Augment SPP (AugSPP) module to aggregate features of multiple-scale and multiple-level. Extensive experiments demonstrate the effectiveness of our method, significantly improving depth estimation results in texture-less regions. The performance of our method outperforms the current state-of-the-art methods (e.g., LFattNet, DistgDisp, OACC-Net, and SubFocal) on the UrbanLF-Syn dataset in terms of MSE $\times 100$, Bad-Pix 0.07, BadPix 0.03, and BadPix 0.01. Our method also ranks third place of comprehensive results in the competition about LFNAT Light Field Depth Estimation Challenge at CVPR 2023 Workshop without any post-processing steps.

**Keywords:** Light field · depth estimation · texture-less regions.

## 1 Introduction

Light field (LF) can simultaneously record the spatial and angular information of the scene with a single shot, which has many practical applications, for example refocusing [20, 30], super-resolution [36, 14, 7, 6, 5, 26, 28], view synthesis [33, 19, 15], semantic segmentation [22], 3D reconstruction [17], virtual reality [35], especially depth (disparity) estimation [38, 4, 39, 23, 25, 21, 3, 27, 29, 2, 1].

By utilizing the additional angle information of LF images, researchers on LF depth estimation have made great progress. Texture-less regions are common and intractable in LF images, restricting the performance of depth estimation, especially in synthetic datasets such as UrbanLF-Syn[22]. However, there are few methods for texture-less LF depth estimation, specifically for large texture-less regions. We analyze the characteristics of existing methods, which can be divided into traditional and deep learning-based methods. Traditional methods[24, 13,

32, 38, 4, 41, 39] need to rely on various prior assumptions, are very vulnerable to texture-less regions, and the algorithms are time-consuming.

In recent years, with the development of deep learning, deep learning-based methods[11, 10, 23, 25, 3, 27, 29, 2, 1] increasingly are used for LF depth estimation tasks and have advantages over traditional methods in terms of efficiency and accuracy. Nowadays, the mainstream deep learning-based methods[25, 3, 29, 27, 2, 1] are mainly based on multi-view stereo matching, including four steps: feature extraction, cost volume construction, cost aggregation, and disparity regression, such as SubFocal[2], which has achieved the best overall performance on the HCI 4D benchmark[12]. Currently, existing methods[25, 3, 29, 27, 2, 1] often adopt the LF image patch (e.g., 32×32) for efficient training, which is effective when there are few or no texture-less regions in the image. However, its side effect causes the receptive field of the model to be too limited and unable to learn abundant context information. When the texture-less region is too large, it may exceed the size of the model's receptive field, making it impossible for the model to infer reasonable results.

We find that the depth estimation results of textured regions around texture-less regions are reliable, so it is important to learn large and abundant context information effectively to alleviate the difficulty of depth estimation in large-scale texture-less regions. As for context information learning, there are two aspects to achieving it. On the one hand, we aim to enlarge the receptive field of feature extraction, including using dilated convolutions and increasing the patchsize of training images. On the other hand, we design an Augment Spatial Pyramid Pooling (AugSPP) module to aggregate features from multiple scales and levels. Combining the expanded receptive field with the AugSPP module alleviates the difficulty of depth estimation in large-scale texture-less regions. Our contributions are as follows:

- We analyze the importance of context information in texture-less regions and present a simple yet effective method called ContextNet to learn context information for texture-less LF depth estimation.
- We utilize dilated convolutions and increase the patchsize of training images to enlarge the receptive field of feature extraction. Moreover, an augment SPP (AugSPP) module is designed to effectively aggregate features from multiple scales and levels.
- Extensive experiments validate the effectiveness of our method. In comparison with state-of-the-art methods, our method achieves superior performance in terms of MSE ×100, BadPix 0.07, BadPix 0.03 and BadPix 0.01 metrics on the UrbanLF-Syn dataset. Furthermore, our method ranks third place of overall scores in the LFNAT LF Depth Estimation Challenge at CVPR 2023 Workshop without any post-processing.

## 2   Related Work

The related work on LF depth estimation can generally be divided into two categories: traditional methods and deep learning-based methods. Below we review each category in detail.

## 2.1  Traditional Methods

Traditional methods can generally be subdivided into three categories based on the representations of LF images: multi-view stereo (MVS), epipolar plane image (EPI), and defocus-based methods. MVS-based methods [13] use multi-view information from SAIs for stereo matching to obtain depth. Jeon *et al.*[13] employed phase translation theory to describe the sub-pixel shift between SAIs and utilized matching processes for stereo matching. EPI-based methods [37] implicitly estimate the depth of the scene by computing the slope of the EPI. Wanner *et al.*[31] proposed a structure tensor that can estimate the slope of lines in horizontal and vertical EPIs, and refined these results through global optimization. Additionally, Zhang *et al.*[37] introduced the Spinning Parallelogram Operator (SPO), which can compute the slope of straight lines in EPI with minimal sensitivity to occlusion, noise, and spatial blending. Defocus-based methods [24] obtain depth by measuring how blurred pixels are on different focus stacks. Tao *et al.*[24] combined scattering and matching cues to generate a local depth map through Markov random field for global optimization. Williem *et al.*[32] improved the robustness of occlusion and noise for depth estimation by using information entropy between different angles and adaptive scattering. Zhang *et al.*[38] used the special linear structure of an EPI and locally linear embedding (LLE) for LF depth estimation. Zhang *et al.*[39] proposed a two-stage method for LF depth estimation that utilized graph-based structure-aware analysis. Zhang *et al.*[40] combined an undirected graph with occluded and un-occluded SAIs in corner blocks to exploit the structural information of the LF. Han *et al.*[9] introduced an occlusion-aware vote cost (OAVC) to enhance the accuracy of edge preservation in the depth map. However, these methods rely on hand-designed features and subsequent optimization, which are time-consuming and have limited accuracy for text-less regions.

**Deep Learning-based Methods** Deep learning has experienced rapid development and has been widely applied in various LF processing tasks, particularly depth estimation. Heber *et al.*[11] were the first to use a CNN to extract features from an EPI and calculate the scene's depth. Shin *et al.*[23] proposed EPINet, which used four directional (0°, 90°, 45°, and 135°) EPIs as input and a center sub-aperture image (SAI) disparity map as output. Tsai *et al.*[25] introduced the LFAttNet network, which employed a view selection module based on an attention mechanism to calculate the importance of each view and served as the weight for cost aggregation. Guo *et al.*[8] designed an occlusion region detection network (ORDNet) for explicit estimation of occlusion maps and subsequent networks focus on non-occluded and occluded regions, respectively. Chen *et al.*[3] designed the AttMLFNet, an attention-based multilevel fusion network that combines features from different perspectives hierarchically through intra-branch and inter-branch fusion strategies. Wang *et al.*[29] extended the spatial-angular interaction mechanism to the disentangling mechanism and proposed DistgDisp for LF depth estimation. They also developed the OACC-Net [27],

which uses dilated convolution instead of shift-and-concat operation and iterative processing with occlusion masks to build an occlusion-aware cost volume. Chao *et al.*[2] proposed the SubFocal method for sub-pixel disparity distribution learning by constructing a sub-pixel cost volume and leveraging disparity distribution constraints to obtain a high-precision disparity map. Chao *et al.*[1] presented a method called OccCasNet by constructing the occlusion-aware cascade cost volume for depth estimation and achieved a better trade-off between accuracy and efficiency. At present, existing methods often utilize LF image patches (e.g., 32×32) for training and have achieved high accuracy on textured regions of LF. However, this setting can result in a model with a too-limited receptive field, as the texture-less region may be too large and exceed the size of the model's receptive field, leading to unreasonable results and low accuracy.

We observe that the depth estimation of textured regions around texture-less regions can be accurate. Therefore, we propose a method called ContextNet to learn large and abundant context information in an efficient and effective manner, which will be beneficial for addressing the challenge of depth estimation in large-scale texture-less regions. To achieve this, we utilize dilated convolution and increase the patch size of training images to efficiently enlarge the receptive field of feature extraction. Additionally, we design an AugSPP module to effectively aggregate features at multiple-scale and multiple-level.

## 3   Method

### 3.1   Overview

An overview of our ContextNet is shown in Fig. 1 (a). First, the features of each SAI are extracted and aggregated using a shared feature extraction based on dilated convolution [34] and AugSPP module [16]. Second, the sub-pixel view shift is performed to construct the sub-pixel cost volume [2]. Third, the cost aggregation module is used to aggregate the cost volume information. The predicted disparity map is produced by attaching a disparity regression module. We will describe each module in detail below.

### 3.2   Feature Extraction and AugSPP module

**Enlarging Receptive Field** Dilated convolution expands the receptive field by adding holes in the convolution filter, and has a larger receptive field under the same parameter amount and calculation amount without using downsampling. Therefore, we utilize dilated convolution and increase training patchsize correspondingly to enlarge the receptive field. Figure 1 (b) shows the structure of feature extraction. First, two $3 \times 3$ convolutions are employed to extract the initial feature with a channel of 4. Then, we use a feature extraction based on dilated convolution[34] module to extract multi-level features of SAI. The feature extraction module contains four stages, and the dilated ratios are set to 1, 2, 4, and 8, respectively.
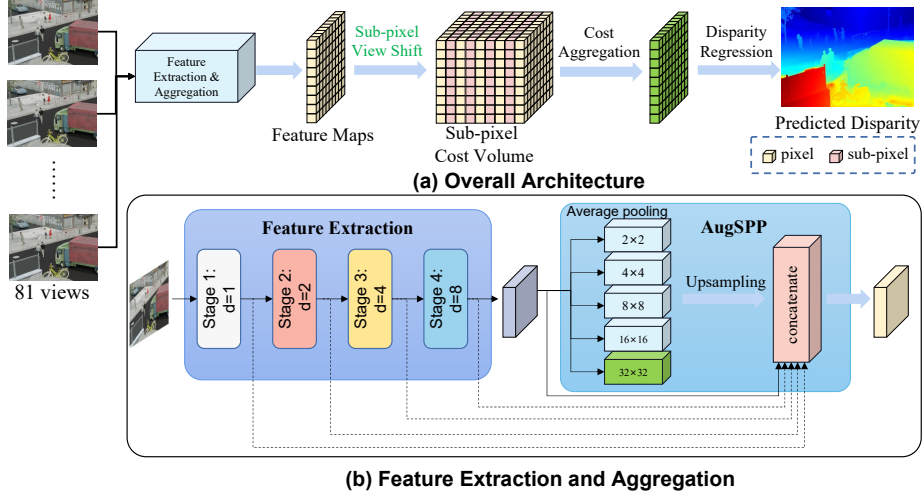
**(a) Overall Architecture**

**(b) Feature Extraction and Aggregation**

**Fig. 1.** The specific network design of ContextNet. (a) Overall Architecture. (b) Feature Extraction and AugSPP module.

**Multi-scale and Multi-level Features Aggregation** Based on the original SPP module[16], we propose the AugSPP module for aggregating multi-scale and multi-level features. Specifically, we add an extra pooling size $32 \times 32$ in the AugSPP module different from previous methods[25, 2]. So five average pooling operations at multi-scale are used to compress the features. The sizes of the average pooling blocks are $2 \times 2$, $4 \times 4$, $8 \times 8$, $16 \times 16$ and $32 \times 32$, respectively. Bilinear interpolation is adopted to upsample these low-dimensional feature maps to the same size. We also aggregate multi-level features of feature extraction using skip connections to further improve the discrimination of features. A $1 \times 1$ convolution layer is used for reducing the feature dimension. The features output by the AugSPP module contain multi-scale and multi-level discriminative context information and are concatenated to form a feature map $F$. The feature extraction and aggregation module incorporates additional textured features from large neighboring regions for challenging regions, such as texture-less and reflection areas.

### 3.3   Sub-pixel Cost Volume

Cost volume is constructed by *shift-and-concat*[25, 3, 2, 1] operation within a predefined disparity range (such as from -0.5 to 1.6 in the UrbanLF-Syn dataset). In order to alleviate the narrow baseline of LF images, Different from the previous method [13] using phase shift theorem to construct sub-pixel cost volume, we follow [2, 1] to construct a sub-pixel feature level cost volume based on bilinear interpolation, which can save memory-consuming. After shifting the feature maps, we concatenate these feature maps into a 4D cost volume $D \times H \times W \times C$.

It is worth noting that a smaller sampling interval can generate a finer sub-pixel cost volume but will increase computation time and slows down inference. Therefore, in order to the trade-off between accuracy and speed, we adopt 22 disparity levels ranging from -0.5 to 1.6, where the sub-pixel interval is 0.1.

### 3.4   Cost Aggregation and Disparity Regression

The shape of the sub-pixel cost volume is $D \times H \times W \times C$, where $H \times W$ denotes the spatial resolution, $D$ is the disparity number, and $C$ is the channel number of feature maps, and we employ 3D CNN to aggregate the sub-pixel cost volume. Following [25, 2], our cost aggregation consists of eight $3 \times 3 \times 3$ convolutional layers and two residual blocks. After passing through these 3D convolutional layers, we obtain the final cost volume $C_f \in D \times H \times W$. We normalize $C_f$ by using the softmax operation along dimension $D$. Finally, the output disparity $\hat{d}$ can be calculated as follows:

$$\hat{d} = \sum_{d_k=D_{min}}^{D_{max}} d_k \times \text{softmax}(-C_{d_k}), \qquad (1)$$

where $\hat{d}$ denotes the estimated center view disparity, $D_{min}$ and $D_{max}$ stand for the minimum and maximum disparity values, respectively, and $d_k$ is the sampling value between $D_{min}$ and $D_{max}$ according to the predefined sampling interval.

## 4   Experiments

In this section, we first introduce the UrbanLF-Syn datasets and implementation details. Then, we compare the performance of our method with the state-of-the-art methods. Finally, we conduct an extensive ablation study to analyze the proposed ContextNet.

### 4.1   Datasets and Implementation Details

UrbanLF-Syn dataset[22] contains 230 synthetic LF samples, with 170 training, 30 validation, and 30 test samples for LFNAT LF Depth Estimation Challenge at the CVPR 2023 Workshop. Each sample consists of 81 SAIs with a spatial resolution of 480×640 and an angular resolution of 9×9.

We employ the $L1$ loss as the loss function, as it is robust to outliers. We use the SubFocal as the baseline model. We follow the same data augmentation strategy as in [25, 2, 1] to improve the model performance, which includes random horizontal and vertical flipping, 90-degree rotation, and adding random noise. It is important to note that the spatial and angular dimensions need to be flipped or rotated jointly to maintain the LF structures. We randomly crop LF images into 64×64 grayscale patches to provide more context information for our model. We remove texture-less regions where the mean absolute difference between the center pixel and other pixels is less than 0.02. The batchsize is set to 16, and we

**Table 1.** Quantitative comparison results with state-of-the-art methods on the validation of UrbanLF-Syn dataset[22] in terms of BadPix 0.07, BadPix 0.03, BadPix 0.01, and MSE×100. The best results are shown in boldface.

| Method | Img11 | Img27 | Img34 | Img50 | Img54 | Img68 | Img69 | Img70 | Avg. BP 0.07 |
|---|---|---|---|---|---|---|---|---|---|
| LFattNet[25] | 10.179 | 23.648 | 22.637 | 9.668 | 13.533 | 9.819 | 12.075 | 7.467 | 13.629 |
| DistgDisp[29] | **6.825** | 20.240 | 24.522 | 8.304 | 13.830 | 9.367 | 10.118 | 7.691 | 12.612 |
| OACC-Net[27] | 6.845 | 24.058 | 27.792 | 10.390 | 13.922 | 11.799 | 13.234 | 8.750 | 14.599 |
| SubFocal[2] | 9.194 | 22.223 | 22.619 | 6.008 | 11.476 | 8.385 | 9.357 | 6.542 | 11.976 |
| Ours | 8.456 | **12.149** | **17.104** | **5.396** | **5.771** | **6.954** | **5.790** | **2.875** | **8.062** |

| Method | Img11 | Img27 | Img34 | Img50 | Img54 | Img68 | Img69 | Img70 | Avg. BP 0.03 |
|---|---|---|---|---|---|---|---|---|---|
| LFattNet[25] | 14.902 | 34.327 | 30.583 | 15.486 | 19.495 | 14.758 | 18.967 | 15.744 | 20.533 |
| DistgDisp[29] | **12.240** | 33.657 | 32.753 | 15.282 | 21.297 | 15.938 | 18.350 | 15.770 | 20.661 |
| OACC-Net[27] | 14.095 | 38.117 | 34.853 | 18.251 | 21.962 | 23.187 | 21.917 | 18.052 | 23.804 |
| SubFocal[2] | 13.342 | 30.742 | 28.696 | 10.661 | 15.413 | 12.176 | 14.103 | 13.739 | 17.359 |
| Ours | 15.032 | **23.512** | **24.472** | **10.643** | **11.308** | **10.758** | **11.187** | **11.723** | **14.830** |

| Method | Img11 | Img27 | Img34 | Img50 | Img54 | Img68 | Img69 | Img70 | Avg. BP 0.01 |
|---|---|---|---|---|---|---|---|---|---|
| LFattNet[25] | 24.113 | 49.134 | 43.789 | 30.819 | 33.616 | 28.407 | 34.609 | 31.762 | 34.531 |
| DistgDisp[29] | 30.141 | 51.105 | 50.838 | 33.585 | 38.073 | 46.023 | 36.557 | 36.934 | 40.407 |
| OACC-Net[27] | 33.191 | 57.538 | 51.364 | 38.486 | 39.505 | 56.173 | 42.212 | 41.426 | 44.987 |
| SubFocal[2] | **20.913** | 41.808 | 37.782 | 19.375 | 22.847 | 19.485 | 23.257 | 25.075 | 26.318 |
| Ours | 23.614 | **39.654** | **35.729** | **19.788** | **20.960** | **18.555** | **21.352** | **22.688** | **25.293** |

| Method | Img11 | Img27 | Img34 | Img50 | Img54 | Img68 | Img69 | Img70 | Avg. MSE |
|---|---|---|---|---|---|---|---|---|---|
| LFattNet[25] | 0.662 | 1.977 | 2.514 | 0.283 | 1.142 | 1.083 | 0.495 | 0.210 | 1.046 |
| DistgDisp[29] | **0.304** | 1.558 | 2.509 | 0.192 | 0.763 | 0.743 | 0.313 | 0.185 | 0.820 |
| OACC-Net[27] | 0.342 | 3.025 | 11.739 | 1.799 | 3.527 | 2.790 | 2.095 | 0.273 | 3.199 |
| SubFocal[2] | 0.644 | 1.735 | 2.975 | 0.200 | 1.376 | 1.066 | 0.515 | 0.165 | 1.085 |
| Ours | 0.324 | **0.795** | **1.038** | **0.105** | **0.252** | **0.418** | **0.172** | **0.080** | **0.398** |

use the Adam optimizer [18]. The disparity range is set from -0.5 to 1.6, and the disparity interval is set to 0.1. Our ContextNet is implemented in the framework of TensorFlow on an NVIDIA A100 GPU. The learning rate is initially set to $1 \times 10^{-3}$ and decreased by a factor of 0.5 after every 30 epochs. The training is stopped after 120 epochs, and we select the best model based on its performance on the validation set.

We evaluate our method using two metrics: Mean Squared Error (MSE × 100), and BadPix($\epsilon$). The MSE ×100 measures the mean square errors of all pixels, multiplied by 100. BadPix ($\epsilon$) represents the percentage of pixels whose absolute disparity error exceeds a predefined threshold, commonly set to 0.01, 0.03, or 0.07.

### 4.2 Comparison of State-of-the-art Methods

**Qualitative Comparison** We compare our method with four state-of-the-art methods, including LFattNet [25], DistgDisp [29], OACC-Net [27], and SubFocal [2]. Figure 2 shows qualitative comparison results on *Img27*, *Img54*, *Img69*, and
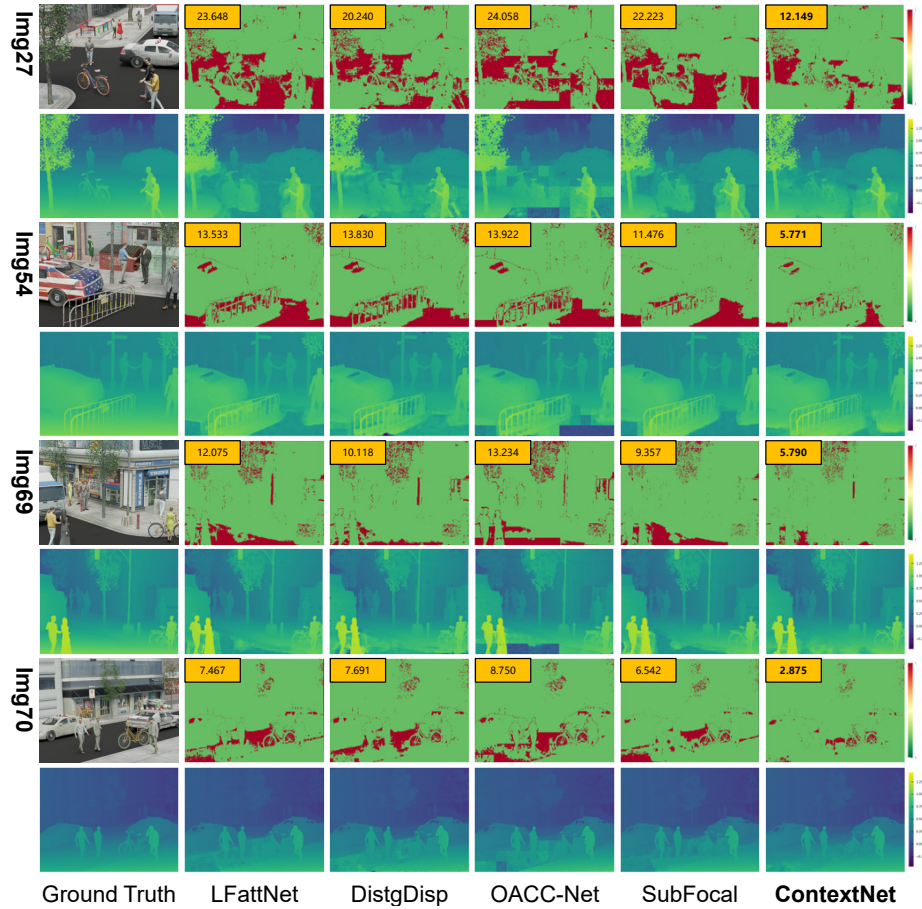
**Fig. 2.** Visual comparisons between our method and state-of-the-art methods on the UrbanLF-Syn dataset[22] scenes, i.e, *Img27*, *Img54*, *Img69* and *Img70*, including LFattNet [25], DistgDisp [29], OACC-Net [27] and SubFocal[2], with the corresponding BadPix 0.07 error maps. Lower is better. The best results are shown in boldface. Please zoom in for a better comparison.

*Img70* scenes of UrbanLF-Syn dataset validation set. Note that these scenes contain large texture-less areas, such as road surfaces. Compared with other methods, our method has less error overall, especially in texture-less regions, which verifies the effectiveness of our method, which can help texture-less regions for depth estimation by learning context information.

**Quantitative Comparison** We have also conducted quantitative comparison experiments with four state-of-the-art methods [25, 29, 27, 2]. Table 1 shows the comparison results on the UrbanLF-Syn dataset[22] for four metrics: BadPix

**Table 2.** The benchmark in the average comparison on the testing set of the UrbanLF-Syn [22] dataset in terms of BadPix 0.07, BadPix 0.03, BadPix 0.01, and MSE ×100. The best results are shown in boldface.

| Method | MSE ×100 | BadPix 0.07 | BadPix 0.03 | BadPix 0.01 |
|---|---|---|---|---|
| MultiBranch | 2.776 | 86.35 | 64.915 | 43.402 |
| MTLF | 1.373 | 41.156 | 21.452 | 13.034 |
| UOAC | 0.953 | 53.926 | 28.211 | 15.582 |
| EPI-Cost | 1.175 | 47.927 | 24.15 | 14.637 |
| MS3D | 0.559 | 31.066 | 14.664 | 7.917 |
| CBPP | 0.394 | 27.385 | **12.628** | **5.907** |
| HRDE | **0.368** | 27.802 | 12.825 | 6.205 |
| Ours | 0.416 | **24.681** | 12.649 | 6.75 |

**Table 3.** The average results of different disparity intervals and numbers on 8 scenes of the UrbanLF-Syn dataset[22] dataset validation set in terms of BadPix 0.07, and MSE×100. The best results are shown in boldface.

| Disparity Interval | Disparity Number | Disparity Range | MSE ×100 | BadPix 0.07 |
|---|---|---|---|---|
| 1 | 4 | [-1, 2] | 1.277 | 14.367 |
| 0.5 | 6 | [-0.5, 2] | 1.141 | 13.873 |
| 0.3 | 8 | [-0.5, 1.6] | 1.072 | 13.614 |
| 0.15 | 15 | [-0.5, 1.6] | **1.063** | 13.254 |
| 0.1 | 22 | [-0.5, 1.6] | 1.065 | **12.946** |
| 0.05 | 43 | [-0.5, 1.6] | 1.085 | 13.201 |

0.07, BadPix 0.03, BadPix 0.01, and MSE ×100. Our method ranks first in most scenes and achieves the top metrics in average BadPix 0.07, BadPix 0.03, BadPix 0.01, and MSE ×100, significantly outperforming current state-of-the-art methods by a large margin. We have submitted our results to the UrbanLF-Syn dataset website. Table 2 shows that our method is competitive and also ranks third place of comprehensive results in the competition for LFNAT LF Depth Estimation Challenge at CVPR 2023 Workshop without any post-processing steps[1].

### 4.3    Ablation Study

**Disparity Interval of Sub-pixel Cost Volume** We also conduct extensive ablation experiments to validate our method. First, we experiment with different disparity intervals and disparity numbers for the sub-pixel cost volume. The training patchsize of the default settings is set to 32, and the feature channel of cost aggregation is set to 170. It can be seen from Table 3 that as the disparity

---

[1] `http://www.lfchallenge.com/dp_lambertian_plane_result/`. On the benchmark, the name of our method is called SF-Net.

**Table 4.** The average results of different variants on 8 scenes of the UrbanLF-Syn dataset[22] dataset validation set in terms of BadPix 0.07, and MSE×100. The best results are shown in boldface.

| Variants | MSE ×100 | BadPix 0.07 |
|---|---|---|
| baseline | 1.157 | 13.230 |
| +add input training patchsize: 64×64 | 0.806 | 11.298 |
| +change dilated ratio: [1,2,4,8] | 0.416 | 8.758 |
| +AugSPP: 32×32 average pooling | 0.459 | 8.675 |
| +AugSPP: concat multi-level feature | 0.440 | 8.223 |
| +finer interval: 0.1 | 0.460 | 8.184 |
| +more feature channel: 170 | **0.398** | **8.062** |

interval decreases, the corresponding MSE and Badpix 0.07 are totally decreased. Considering efficiency and accuracy, we finally chose a disparity interval of 0.1.

**Context Information Learning** We validate different components for context information learning. The training patchsize of the baseline model is set to 32, the feature channel of cost aggregation is set to 96, the disparity interval is 0.15, the number of disparities is 15, and the disparity range is -0.5 to 1.6. As shown in Table 4, the different components we proposed can improve the metrics step by step. Compared with the baseline model, our method achieves improvements of 65.6% and 39% on the MSE ×100 and BadPix 0.07 metrics, respectively.

## 5    Conclusion and Limitations

In this paper, we propose a method, namely ContextNet, to learn context information for texture-less LF depth estimation. On the one hand, we use dilated convolution and increase the patchsize of training images to enlarge the receptive field of feature extraction. On the other hand, an AugSPP module is designed to improve the overall performance of our method by effectively aggregating features from multi-scale and multi-level. Extensive experiments validate the effectiveness of our method. Our method outperforms state-of-the-art methods on the UrbanLF-Syn dataset and also ranks third place of comprehensive results in the competition about LFNAT LF Depth Estimation Challenge at CVPR 2023 Workshop.

While our method achieves competitive results in texture-less regions, there is still room for improvement. Regarding the LF depth estimation of texture-less regions, we plan to start from the following aspects in the future. We can further expand the receptive field by fusing the results of monocular depth estimation. Additionally, we may design a post-processing step by diffusing the depth of textured regions to texture-less regions. Finally, we can utilize shape priors to texture-less depth estimation with the help of semantic segmentation maps.

# References

1. Chao, W., Duan, F., Wang, X., Wang, Y., Wang, G.: Occcasnet: Occlusion-aware cascade cost volume for light field depth estimation. arXiv preprint arXiv:2305.17710 (2023)
2. Chao, W., Wang, X., Wang, Y., Chang, L., Duan, F.: Learning sub-pixel disparity distribution for light field depth estimation. arXiv preprint arXiv:2208.09688 (2022)
3. Chen, J., Zhang, S., Lin, Y.: Attention-based multi-level fusion network for light field depth estimation. In: AAAI. pp. 1009–1017 (2021)
4. Chen, J., Chau, L.: Light field compressed sensing over a disparity-aware dictionary. TCSVT **27**(4), 855–865 (2017)
5. Chen, Y., Zhang, S., Chang, S., Lin, Y.: Light field reconstruction using efficient pseudo 4d epipolar-aware structure. TCI **8**, 397–410 (2022)
6. Cheng, Z., Liu, Y., Xiong, Z.: Spatial-angular versatile convolution for light field reconstruction. TCI **8**, 1131–1144 (2022)
7. Cheng, Z., Xiong, Z., Chen, C., Liu, D., Zha, Z.J.: Light field super-resolution with zero-shot learning. In: CVPR. pp. 10010–10019 (2021)
8. Guo, C., Jin, J., Hou, J., Chen, J.: Accurate light field depth estimation via an occlusion-aware network. In: ICME. pp. 1–6 (2020)
9. Han, K., Xiang, W., Wang, E., Huang, T.: A novel occlusion-aware vote cost for light field depth estimation. TPAMI **44**(11), 8022–8035 (2022)
10. He, L., Wang, G., Hu, Z.: Learning depth from single images with deep neural network embedding focal length. TIP **27**(9), 4676–4689 (2018)
11. Heber, S., Pock, T.: Convolutional networks for shape from light field. In: CVPR. pp. 3746–3754 (2016)
12. Honauer, K., Johannsen, O., Kondermann, D., Goldluecke, B.: A dataset and evaluation methodology for depth estimation on 4d light fields. In: ACCV. pp. 19–34. Springer (2016)
13. Jeon, H.G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.W., So Kweon, I.: Accurate depth map estimation from a lenslet light field camera. In: CVPR. pp. 1547–1555 (2015)
14. Jin, J., Hou, J., Chen, J., Kwong, S.: Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In: CVPR. pp. 2260–2269 (2020)
15. Jin, J., Hou, J., Chen, J., Zeng, H., Kwong, S., Yu, J.: Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion. TPAMI (2020)
16. K. He, X. Zhang, S. Ren, and J. Sun: Spatial pyramid pooling in deep convolutional networks for visual recognition. TPAMI **37**(9), 1904–1916 (2015)
17. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.H.: Scene reconstruction from high spatio-angular resolution light fields. TOG **32**(4), 73–1 (2013)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Meng, N., So, H.K.H., Sun, X., Lam, E.Y.: High-dimensional dense residual convolutional neural network for light field reconstruction. TPAMI **43**(3), 873–886 (2019)
20. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Ph.D. thesis, Stanford University (2005)

21. Peng, J., Xiong, Z., Wang, Y., Zhang, Y., Liu, D.: Zero-shot depth estimation from light field using a convolutional neural network. TCI **6**, 682–696 (2020)
22. Sheng, H., Cong, R., Yang, D., Chen, R., Wang, S., Cui, Z.: Urbanlf: A comprehensive light field dataset for semantic segmentation of urban scenes. TCSVT **32**(11), 7880–7893 (2022)
23. Shin, C., Jeon, H.G., Yoon, Y., Kweon, I.S., Kim, S.J.: Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In: CVPR. pp. 4748–4757 (2018)
24. Tao, M.W., Hadap, S., Malik, J., Ramamoorthi, R.: Depth from combining defocus and correspondence using light-field cameras. In: ICCV. pp. 673–680 (2013)
25. Tsai, Y.J., Liu, Y.L., Ouhyoung, M., Chuang, Y.Y.: Attention-based view selection networks for light-field disparity estimation. In: AAAI. pp. 12095–12103 (2020)
26. Van Duong, V., Huu, T.N., Yim, J., Jeon, B.: Light field image super-resolution network via joint spatial-angular and epipolar information. TCI (2023)
27. Wang, Y., Wang, L., Liang, Z., Yang, J., An, W., Guo, Y.: Occlusion-aware cost constructor for light field depth estimation. In: CVPR. pp. 19809–19818 (2022)
28. Wang, Y., Wang, L., Liang, Z., Yang, J., Timofte, R., Guo, Y.: Ntire 2023 challenge on light field image super-resolution: Dataset, methods and results. arXiv preprint arXiv:2304.10415 (2023)
29. Wang, Y., Wang, L., Wu, G., Yang, J., An, W., Yu, J., Guo, Y.: Disentangling light fields for super-resolution and disparity estimation. TPAMI (2022)
30. Wang, Y., Yang, J., Guo, Y., Xiao, C., An, W.: Selective light field refocusing for camera arrays using bokeh rendering and superresolution. SPL **26**(1), 204–208 (2018)
31. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. TPAMI **36**(3), 606–619 (2014)
32. Williem, W., Park, I.K.: Robust light field depth estimation for noisy scene with occlusion. In: CVPR. pp. 4396–4404 (2016)
33. Wu, G., Liu, Y., Fang, L., Dai, Q., Chai, T.: Light field reconstruction using convolutional network on epi and extended applications. TPAMI **41**(7), 1681–1694 (2018)
34. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
35. Yu, J.: A light-field journey to virtual reality. TMM **24**(2), 104–112 (2017)
36. Zhang, S., Lin, Y., Sheng, H.: Residual networks for light field image super-resolution. In: CVPR. pp. 11046–11055 (2019)
37. Zhang, S., Sheng, H., Li, C., Zhang, J., Xiong, Z.: Robust depth estimation for light field via spinning parallelogram operator. CVIU **145**, 148–159 (2016)
38. Zhang, Y., Lv, H., Liu, Y., Wang, H., Wang, X., Huang, Q., Xiang, X., Dai, Q.: Light-field depth estimation via epipolar plane image analysis and locally linear embedding. TCSVT **27**(4), 739–747 (2016)
39. Zhang, Y., Dai, W., Xu, M., Zou, J., Zhang, X., Xiong, H.: Depth estimation from light field using graph-based structure-aware analysis. TCSVT **30**(11), 4269–4283 (2019)
40. Zhang, Y., Dai, W., Xu, M., Zou, J., Zhang, X., Xiong, H.: Depth estimation from light field using graph-based structure-aware analysis. TCSVT **30**(11), 4269–4283 (2020)
41. Zhu, H., Wang, Q., Yu, J.: Occlusion-model guided antiocclusion depth estimation in light field. J-STSP **11**(7), 965–978 (2017)