ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446981

UNSUPERVISED DISPARITY ESTIMATION FOR LIGHT FIELD VIDEOS

Shansi Zhang and Edmund Y. Lam

Department of Electrical and Electronic Engineering,
The University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Light field (LF) videos contain not only the spatial-angular information but also the temporal information, which are useful for disparity estimation. The existing work on disparity estimation for LF videos relies on supervised training with disparity labels. To overcome this reliance, we develop an unsupervised disparity estimation framework for LF videos, which consists of a matching branch to perform feature matching and a refinement branch to refine the disparity maps. Our framework also includes a cross-feature fusion module with self-attention and cross-attention to fuse the multi-frame features, and a cost aggregation transformer with cross-depth self-attention blocks to explore the global depth dependencies. Moreover, we propose a left-right consistency strategy to estimate the occlusion regions for the input views and introduce a occlusion-aware photometric loss to solve the occlusion issue. Experimental results demonstrate that our method achieves superior performance compared to the existing supervised and unsupervised methods.

***Index Terms*—** LF videos, unsupervised disparity estimation, feature fusion, self-attention, occlusion

1. INTRODUCTION

A light field (LF) image contains multiple views to record the scene from different directions [1–3], providing rich geometric information that is useful for disparity (depth) estimation. This step is crucial for realistic sensing and 3D scene reconstruction [4, 5]. Although most existing LF cameras are limited to capturing static LF images, some advanced imaging devices have been developed to capture LF videos, which include a temporal dimension in addition to the spatial and angular dimensions. The temporal correlation among consecutive frames can also facilitate LF disparity estimation.

Most existing work for LF disparity estimation focuses on static LF images, and they can be divided into traditional methods, supervised learning-based methods and unsupervised learning-based methods. The traditional methods [6, 7] usually study the epipolar-plane images (EPIs) to infer the disparity values, or adopt the classical stereo matching approach to find correspondences in different views. However, they involve complex optimization problem with long execution time, and are usually sensitive to the occluded, texture-less and noisy regions. With the prevalence of deep learning, some supervised learning-based methods [8–10] have been proposed for LF disparity estimation, which significantly improve the accuracy and efficiency. However, these supervised methods depend on the disparity labels, which are costly to acquire, especially for the real-world LF images, and insufficiency of labeled data may lead to poor generalization capability. To eliminate the need for disparity labels, some unsupervised learning-based methods [11–14] have been developed, which are trained by the photometric loss that minimizes the warping errors. The above-mentioned methods can be applied to the individual frames of LF videos. However, the temporal information is not utilized to achieve better performance.

Disparity estimation for LF videos is a relatively new research area that has not been fully explored yet. Kinoshita et al. [15] developed a dataset of LF videos with disparity labels for every view, and proposed a framework based on Convolutional Long Short-Term Memory (CLSTM) to extract spatial-angular-temporal information from the LF videos. However, this architecture also requires disparity labels for supervised training.

To overcome the reliance on disparity labels, we develop an unsupervised disparity estimation framework for LF videos, called TStereoNet ('T' denotes temporal), which fully exploits the multi-view and temporal information. Our TStereoNet takes as input two views from the same row or column to estimate disparity maps for both of them. It consists of two branches, with a matching branch to estimate the initial disparity maps in a small scale by constructing matching costs, and a refinement branch to further refine the disparity maps in the full scale. Within the TStereoNet, we propose a cross-feature fusion (CFF) module with self-attention and cross-attention to fuse the features of multiple frames, and a cost aggregation transformer with cross-depth self-attention blocks to model the global depth dependencies, which is more lightweight and efficient than the stacked 3D convolution layers commonly used in the existing work [14, 16–18]. To address the occlusion issue, we propose a left-right consis-

This work was supported in part by the Research Grants Council of Hong Kong under Grant GRF 17201822, and in part by the ACCESS—AI Chip Center for Emerging Smart Systems, Hong Kong, SAR.

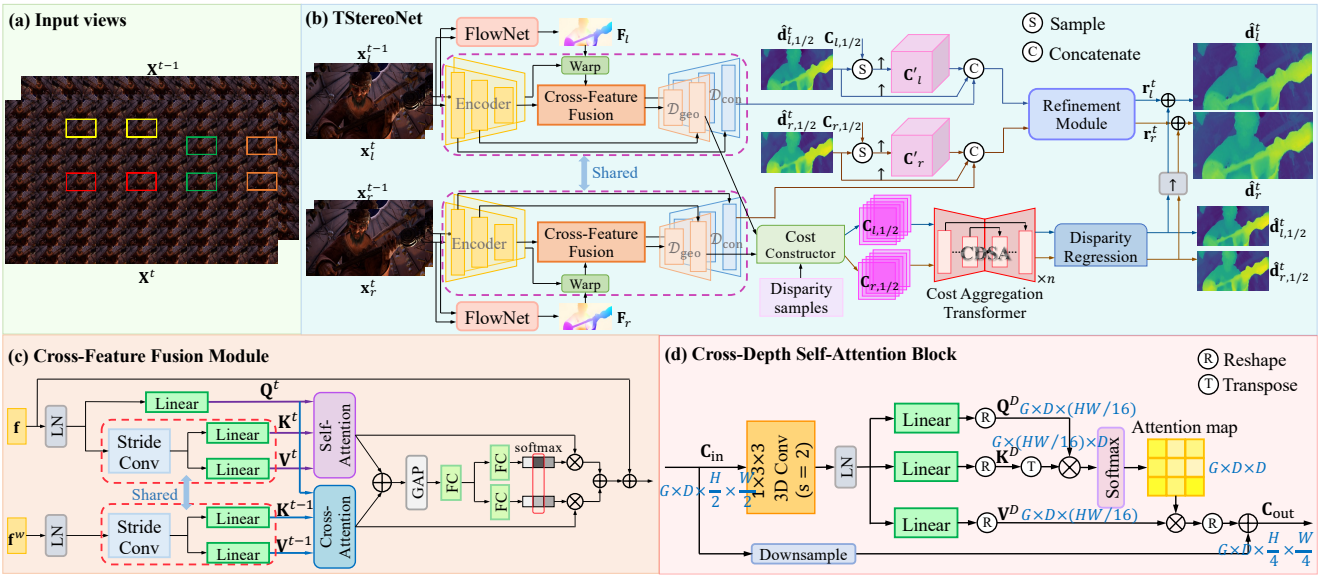

Fig. 1. (a) Input views. (b) Architecture of TStereoNet. (c) Cross-feature fusion module. (d) Cross-depth self-attention block.

tency strategy to obtain the occlusion regions for both input views and introduce a occlusion-aware photometric loss. Experimental results demonstrate that our framework outperforms the other unsupervised disparity estimation methods and achieves comparable performance with the supervised baseline model [15]. In addition, our framework can estimate disparity maps for every view, whereas most existing methods [13–15] are only capable of estimating disparity map for the central view of each LF frame.

2. PROPOSED METHOD

One frame of LF video at time t is represented as $\mathbf{X}^t \in \mathbb{R}^{U\times V\times 3\times H\times W}$ with angular resolution $U \times V$ and spatial resolution $H \times W$. The angular position and spatial position are indexed by (u, v) and (h, w), respectively. A single view at an arbitrary angular position, $\mathbf{X}^t(u, v) \in \mathbb{R}^{3\times H\times W}$, is denoted as $\mathbf{x}^t$.

2.1. Architecture of TStereoNet

Fig. 1 illustrates the architecture of our TStereoNet that consists of a matching branch and a refinement branch. It takes as input two views $\mathbf{x}_l^t$ and $\mathbf{x}_r^t$ from the same row or column in $\mathbf{X}^t$, and their corresponding views $\mathbf{x}_l^{t-1}$ and $\mathbf{x}_r^{t-1}$ in the previous frame $\mathbf{X}^{t-1}$. The views from the same column are rotated by 90° counterclockwise for horizontal disparity estimation. All the input views are first fed to a shared encoder to obtain the $\frac{1}{4}$-scale features. A pre-trained FlowNet is leveraged to estimate the optical flow $\mathbf{F}_l$ (from $\mathbf{x}_l^t$ to $\mathbf{x}_l^{t-1}$) and $\mathbf{F}_r$ (from $\mathbf{x}_r^t$ to $\mathbf{x}_r^{t-1}$), which are used to warp the features at $t-1$ to align with the features at t.

We propose a cross-feature fusion (CFF) module to fuse the feature $\mathbf{f}$ at t and the warped feature $\mathbf{f}^w$ at $t-1$, as shown in Fig. 1(c). The feature $\mathbf{f}$ after layer normalization (LN) passes through a liner layer to obtain $\mathbf{Q}^t$, and a stride convolution (for spatial reduction [19]) and linear layers to obtain $\mathbf{K}^t$ and $\mathbf{V}^t$. Then, self-attention within $\mathbf{f}$ is computed to explore the global spatial dependencies, with $\mathrm{softmax}(\frac{\mathbf{Q}^t(\mathbf{K}^t)^{\mathrm{T}}}{\tau})\mathbf{V}^t$, where τ is a learnable scale factor. The warped feature $\mathbf{f}^w$ is fed to the shared stride convolution and linear layers to obtain $\mathbf{K}^{t-1}$ and $\mathbf{V}^{t-1}$. Cross-attention between $\mathbf{f}$ and $\mathbf{f}^w$ is calculated by $\mathrm{softmax}(\frac{\mathbf{Q}^t(\mathbf{K}^{t-1})^{\mathrm{T}}}{\tau})\mathbf{V}^{t-1}$. The features after self-attention and cross-attention are added together, followed by the global average pooling (GAP), fully-connected (FC) layers and softmax activation to yield their channel-wise weights. Then, they are fused according to their corresponding weights, and the fused feature is added to the input feature $\mathbf{f}$ for a residual connection.

Although we take two-frame fusion as an example, our CFF module can realize fusion of multiple frames to fully utilize the temporal information. After that, a geometry decoder $\mathcal{D}_{\mathrm{geo}}$ and a context decoder $\mathcal{D}_{\mathrm{con}}$ are employed to upsample the feature output by the CFF module, with skip connections to merge the features from the encoder.

In the matching branch, the $\frac{1}{2}$-scale features from $\mathcal{D}_{\mathrm{geo}}$ are fed to a cost constructor, which performs warping operation according to the pre-defined disparity samples and calculates group-wise correlation [17] to measure the feature similarity, deriving the matching costs $\mathbf{C}_{l,1/2}$, $\mathbf{C}_{r,1/2} \in \mathbb{R}^{G\times D\times \frac{H}{2}\times \frac{W}{2}}$ (G and D are the number of groups and disparity samples, respectively) for $\mathbf{x}_l^t$ and $\mathbf{x}_r^t$, respectively. Then, $\mathbf{C}_{l,1/2}$ and $\mathbf{C}_{r,1/2}$ are input to a cost aggregation transformer, which consists of several cross-depth self-attention (CDSA) blocks with

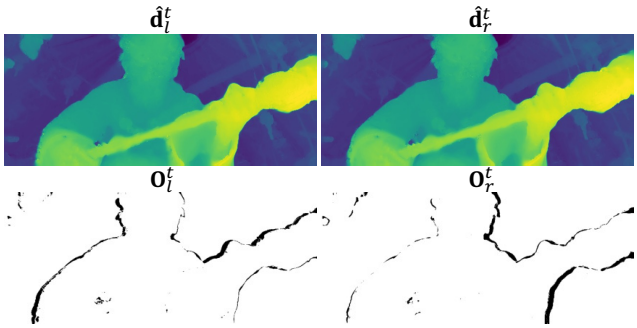

Fig. 2. The estimated disparity maps and their corresponding occlusion masks.

skip connections. We take a CDSA block with $\text{stride} = 2$ as an example to illustrate its structure, as shown Fig. 1(c). The input cost $\mathbf{C}_{\text{in}} \in \mathbb{R}^{G\times D\times \frac{H}{2}\times \frac{W}{2}}$ first passes through a $1 \times 3 \times 3$ 3D convolution layer with $\text{stride} = 2$. Then, it is processed by LN, linear layers and reshape operations to obtain $\mathbf{Q}^D$, $\mathbf{K}^D$ and $\mathbf{V}^D \in \mathbb{R}^{G\times D\times \frac{HW}{16}}$. In order to explore the global dependencies among the disparity costs for accurate probability estimation, self-attention is computed along D dimension to derive a $G \times D \times D$ attention map, with $\text{softmax}(\frac{\mathbf{Q}^D(\mathbf{K}^D)^{\text{T}}}{\tau})\mathbf{V}^D$, which is added to the downsampled input cost to obtain the output cost $\mathbf{C}_{\text{out}} \in \mathbb{R}^{G\times D\times \frac{H}{4}\times \frac{W}{4}}$. Multiple cost aggregation transformers can be cascaded, followed by disparity regression [16] to yield the $\frac{1}{2}$-scale disparity maps $\hat{\mathbf{d}}_{l,1/2}^t$ and $\hat{\mathbf{d}}_{r,1/2}^t$, which are upsampled and scaled to obtain the initial disparity maps.

In the refinement branch, the full-scale features from $\mathcal{D}_{\text{con}}$ providing contextual information are used to further refine the disparity maps. In addition, we use $\hat{\mathbf{d}}_{l,1/2}^t$ and $\hat{\mathbf{d}}_{r,1/2}^t$ to sample from $\mathbf{C}_{l,1/2}$ and $\mathbf{C}_{r,1/2}$ via linear interpolation as [18], followed by an upsampling operation to obtain the sampled cost $\mathbf{C}_l'$ and $\mathbf{C}_r'$. The initial disparity map, the contextual feature and the sampled cost for each view are concatenated and then fed to a refinement module, which comprises of two 3×3 convolution layers, to output the residual map $\mathbf{r}_l^t$ for $\mathbf{x}_l^t$, and $\mathbf{r}_r^t$ for $\mathbf{x}_r^t$. The final disparity maps $\hat{\mathbf{d}}_l^t$ and $\hat{\mathbf{d}}_r^t$ are obtained by adding the residual maps to the initial disparity maps. With our architecture, multiple input pairs are available within each LF to enhance data usage efficiency and the disparity map of any view can be obtained with proper input pair.

2.2. Occlusion Handling

The training of unsupervised disparity estimation is based on the photo-consistency assumption. However, it does not hold in the occlusion regions. To handle the occlusion issue, we propose a left-right consistency strategy to estimate the occlusion regions. For non-occluded pixels, the disparity value in $\hat{\mathbf{d}}_l^t$ should be identical to that in $\hat{\mathbf{d}}_r^t$ at the corresponding pixel. The pixels are marked as occluded ones if the mismatch exceeds a prescribe threshold. For the left view $\mathbf{x}_l^t$, its occlusion map $\tilde{\mathbf{O}}_l^t$ is derived with

$$\tilde{\mathbf{O}}_l^t = \exp(-|\mathcal{W}(\hat{\mathbf{d}}_r^t; \hat{\mathbf{d}}_l^t) - \hat{\mathbf{d}}_l^t|), \tag{1}$$

where $\mathcal{W}(\hat{\mathbf{d}}_r^t; \hat{\mathbf{d}}_l^t)$ denotes warping $\hat{\mathbf{d}}_r^t$ to align with $\hat{\mathbf{d}}_l^t$ using $\hat{\mathbf{d}}_l^t$. Then, the occlusion mask $\mathbf{O}_l^t$ is obtained by

$$\mathbf{O}_{l_{(h,w)}}^t = \begin{cases} 0, & \tilde{\mathbf{O}}_{l_{(h,w)}}^t < \gamma \\ 1, & \text{otherwise} \end{cases}, \tag{2}$$

where γ is a threshold to determine the occlusion pixels. The occlusion mask $\mathbf{O}_r^t$ for the right view $\mathbf{x}_r^t$ is obtained by the similar way but exchanging $\hat{\mathbf{d}}_l^t$ and $\hat{\mathbf{d}}_r^t$ in Eq. 1. Fig. 2 gives an example of the occlusion masks, where we can observe that the occlusion regions are usually near the object boundaries and lie in the opposite positions of the left and right views.

With the occlusion masks, we introduce an occlusion-aware photometric loss to exclude the occluded pixels, with

$$\begin{aligned}\mathcal{L}_{\text{opm}} = &\frac{1}{HW}\sum_{h,w} \mathbf{O}_{l_{(h,w)}}^t \odot |\mathcal{W}(\mathbf{x}_r^t; \hat{\mathbf{d}}_l^t) - \mathbf{x}_l^t| \\ &+ \frac{1}{HW}\sum_{h,w} \mathbf{O}_{r_{(h,w)}}^t \odot |\mathcal{W}(\mathbf{x}_l^t; \hat{\mathbf{d}}_r^t) - \mathbf{x}_r^t|,\end{aligned} \tag{3}$$

where $\odot$ denotes the element-wise multiplication. Note that the occlusion masks are not employed in the photometric loss during the initial training epochs in order to avoid inaccurate estimation of occlusion regions due to the inaccurate disparity maps. In addition, we apply the structure-aware smoothness loss [20] $\mathcal{L}_{\text{sm}}$ to $\hat{\mathbf{d}}_l^t$ and $\hat{\mathbf{d}}_r^t$. $\mathcal{L}_{\text{opm}}$ and $\mathcal{L}_{\text{sm}}$ are also applied to the initial disparity maps, which is omitted for simplicity. The full loss for training the TStereoNet is expressed as

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{opm}} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}}, \tag{4}$$

where λ_{sm} is the coefficient of $\mathcal{L}_{\text{sm}}$.

3. EXPERIMENTS

3.1. Dataset and Implementation Details

We used the Sintel 4D LFV dataset [15], which is developed from the open-source movie 'Sintel'. It contains 24 LF video sequences with $20 \sim 50$ frames. Each frame has 9×9 views with 436×1024 spatial resolution. The ground-truth disparity maps for every view are provided.

Regarding our TStereoNet, the initial $\frac{1}{2}$-scale disparity samples are equally spaced within $0 \sim 30$ pixels. The number of disparity samples D is set to 30, and the number of groups G is set to 16. During training, the input views were selected from an arbitrary row or column and randomly cropped to 256×256 patches. We set $\gamma = 0.3$ in Eq. 2 and $\lambda_{\text{sm}} = 0.05$ in Eq. 4. The initial learning rate was 5×10^{-4} and decayed by multiplying 0.9 every 50 epochs. Our TStereoNet was trained by the Adam optimizer [21] for about 300 epochs.

Table 1. Quantitative results on the scenes from the Sintel 4D LFV dataset. **Bold**: Best among the unsupervised methods.

Methods	shaman2				thebigfight3				bamboo2				ambushfight3			
	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)
CLSTM [15]	0.072	2.170	8.504	57.643	0.011	0.036	2.158	19.883	0.021	0.621	3.165	19.432	0.092	3.353	7.289	24.544
UnMono [12]	0.172	8.643	23.943	50.539	0.036	0.944	10.627	36.256	0.027	0.845	4.574	25.107	0.409	8.396	19.956	54.142
UnOcc [13]	0.082	2.271	14.192	49.504	0.025	1.154	5.002	16.438	0.029	0.982	5.459	24.221	0.210	3.330	18.697	53.500
Ours	**0.057**	**1.924**	**3.750**	**29.489**	**0.007**	**0.061**	**0.635**	**14.321**	**0.023**	**0.599**	**2.529**	**23.431**	**0.102**	**2.526**	**6.433**	**26.325**

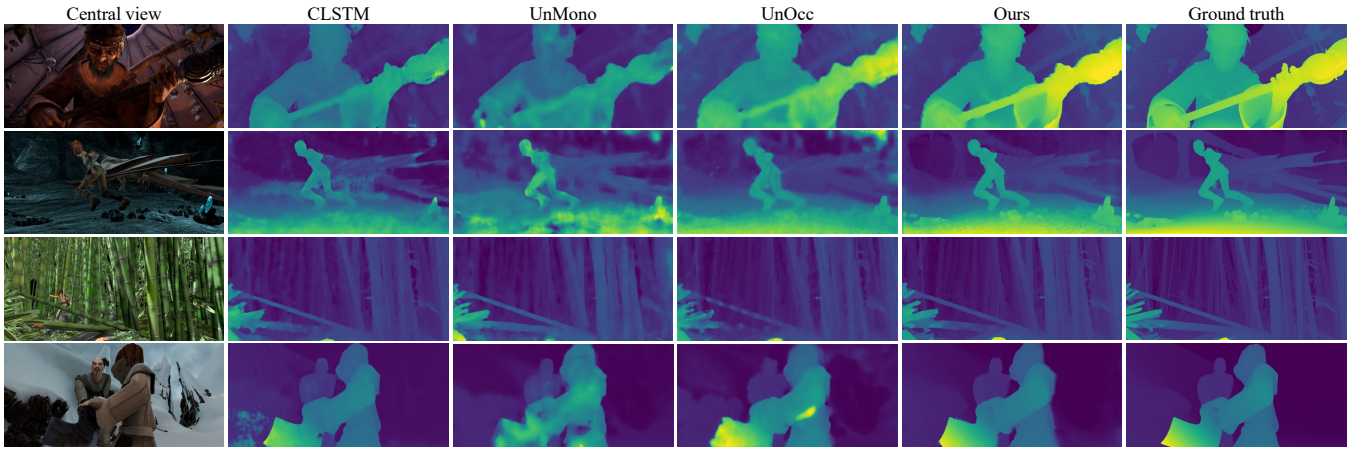

Fig. 3. Visual comparison of different methods on the four scenes from the Sintel 4D LFV dataset.

3.2. Comparison

We compared our method with the supervised baseline CLSTM [15], two unsupervised disparity estimation methods for static LF images, UnMono [12] and UnOcc [13], since we are the first to study unsupervised disparity estimation for LF videos. We evaluated these methods on four scenes from the Sintel 4D LFV dataset, and the quantitative results, including mean square error (MSE), bad pixel ratios (BPR) with threshold 0.07, 0.03 and 0.01, are listed in Table 1. Note that the unit of disparity for evaluation is millimeter as done in [15]. It can be seen that our method outperforms the unsupervised methods, UnMono and UnOcc, and achieves comparable or even better performance than the supervised method, CLSTM.

Fig. 3 shows the visual comparison on the four scenes, where we can see that our disparity maps are more visually compelling, with better smoothness and clearer object details. Note that we can further improve the performance by merging multiple disparity maps obtained from different input view pairs. However, we do not explore the disparity merging method in this study due to the page limit, and leave it for future research.

3.3. Ablation Studies

Several ablation studies were conducted to validate our CFF module, CDSA blocks and occlusion handling strategy. We trained additional models by removing the CFF module, replacing the CDSA blocks with 3D residual blocks, and removing the occlusion masks in Eq. 3, respectively. The average quantitative results with number of parameters (Param.) and FLOPs (for 256×256 inputs) are listed in Table 2. It can be seen that all these components help to improve the performance. In particular, our cost filter with CDSA blocks is more lightweight and involves much less computation compared to that with 3D convolution layers, which is commonly used in the existing stereo matching work.

Table 2. Ablation studies. **Bold**: Best.

Configurations	Param. (M)	FLOPs (G)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)
w/o CFF module	1.013	33.367	0.071	1.627	4.133	27.157
w/o CDSA blocks	1.101	57.187	0.068	1.409	3.728	26.021
w/o occlusion handling	1.020	37.953	0.063	1.322	3.479	25.487
Default	1.020	37.953	**0.047**	**1.278**	**3.337**	**23.392**

4. CONCLUSIONS

We develop an unsupervised disparity estimation framework for LF videos, called TStereoNet, with a matching branch and a refinement branch. It incorporates a cross-feature fusion module to fuse the multi-frame features and a cost aggregation transformer to model the global dependencies among the disparity costs. We also propose a left-right consistency strategy to address the occlusion issue during training. Our method achieves superior performance compared to the existing supervised and unsupervised methods.

5. REFERENCES

[1] Nan Meng, Hayden Kwok-Hay So, Xing Sun, and Edmund Y. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 873–886, 2021.

[2] Shansi Zhang and Edmund Y. Lam, "Learning to restore light fields under low-light imaging," *Neurocomputing*, vol. 456, pp. 76–87, 2021.

[3] Shansi Zhang and Edmund Y. Lam, "An effective decomposition-enhancement method to restore light field images captured in the dark," *Signal Processing*, vol. 189, pp. 108279, 2021.

[4] Xinjue Hu, Chenchen Wang, Yuxuan Pan, Yunming Liu, Yumei Wang, Yu Liu, and Lin Zhang, "4DLFVD: A 4d light field video dataset," in *Proceedings of the ACM Multimedia Systems Conference*, 2021, pp. 287–292.

[5] Shansi Zhang, Nan Meng, and Edmund Y. Lam, "LRT: An efficient low-light restoration transformer for dark light field images," *IEEE Transactions on Image Processing*, vol. 32, pp. 4314–4326, 2023.

[6] Hao Sheng, Pan Zhao, Shuo Zhang, Jun Zhang, and Da Yang, "Occlusion-aware depth estimation for light field using multi-orientation EPIs," *Pattern Recognition*, vol. 74, pp. 587–599, 2018.

[7] Chao-Tsung Huang, "Empirical bayesian light-field stereo matching by robust pseudo random field modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 552–565, 2019.

[8] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo 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)*, 2018, pp. 4748–4757.

[9] Jinglei Shi, Xiaoran Jiang, and Christine 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.

[10] Jiaxin Chen, Shuo Zhang, and Youfang Lin, "Attention-based multi-level fusion network for light field depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 1009–1017.

[11] Jiayong Peng, Zhiwei Xiong, Dong Liu, and Xuejin Chen, "Unsupervised depth estimation from light field using a convolutional neural network," in *Proceedings of the International Conference on 3D Vision*, 2018, pp. 295–303.

[12] Wenhui Zhou, Enci Zhou, Gaomin Liu, Lili Lin, and Andrew Lumsdaine, "Unsupervised monocular depth estimation from light field image," *IEEE Transactions on Image Processing*, vol. 29, pp. 1606–1617, 2020.

[13] Jing Jin and Junhui Hou, "Occlusion-aware unsupervised learning of depth from 4-d light fields," *IEEE Transactions on Image Processing*, vol. 31, pp. 2216–2228, 2022.

[14] Shansi Zhang, Nan Meng, and Edmund Y Lam, "Unsupervised light field depth estimation via multi-view feature matching with occlusion prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[15] Takahiro Kinoshita and Satoshi Ono, "Depth estimation from 4d light field videos," *arXiv preprint arXiv:2012.03021*, 2020.

[16] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 66–75.

[17] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3273–3282.

[18] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21919–21928.

[19] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 568–578.

[20] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 270–279.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.