Attention-based Multi-Level Fusion Network for Light Field Depth Estimation

Jiaxin Chen,[1,2] **Shuo Zhang,**[1,2,3*] **Youfang Lin**[1,2,3,4]
[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing, China
[3]CAAC Key Laboratory of Intelligent Passenger Service of Civil Aviation, Beijing, China
[4]Key Laboratory of Transport Industry of Big Data Appalication Technologies for Comprehensive Transport, Beijing, China
{chenjiaxin, zhangshuo,yflin}@bjtu.edu.cn

Abstract

Depth estimation from Light Field (LF) images is a crucial basis for LF related applications. Since multiple views with abundant information are available, how to effectively fuse features of these views is a key point for accurate LF depth estimation. In this paper, we propose a novel attention-based multi-level fusion network. Combining with the four-branch structure, we design intra-branch fusion strategy and inter-branch fusion strategy to hierarchically fuse effective features from different views. By introducing the attention mechanism, features of views with less occlusions and richer textures are selected inside and between these branches to provide more effective information for depth estimation. The depth maps are finally estimated after further aggregation. Experimental results show the proposed method achieves state-of-the-art performance in both quantitative and qualitative evaluation, which also ranks first in the commonly used HCI 4D Light Field Benchmark.

Introduction

Light Fields (LFs) (Adelson and Bergen 1991) record light in different directions and describe scenes with more rich information than traditional images. Lytro (Ng 2018) and Raytrix (Perwaß and Wietzke 2018) are successful instances of commercial LF cameras, which capture scenes from different directions in one shot by placing the micro-lens array (Ng et al. 2005a) in front of the imaging sensor. As one crucial step, depth estimation provides the structure information and is used for various researches, such as digital refocusing (Ng et al. 2005b), image segmentation (Wanner, Straehle, and Goldluecke 2013), view synthesis (Jin et al. 2020), saliency detection (Li et al. 2014) and super resolution (Zhang, Lin, and Sheng 2019).

In order to estimate accurate depth (or equal to disparity) information, lots of traditional approaches have been proposed. Based on the photo-consistency of views in LFs, some methods (Jeon et al. 2015; Williem, Park, and Lee 2017; Sheng et al. 2017) are designed by constructing cost volumes based on traditional stereo matching. Due to the narrow baseline of LFs, other approaches analyze specific linear structures in Epipolar Plane Images (EPIs) (Wanner, Straehle, and Goldluecke 2014; Zhang et al. 2016) or the focusness in focal stack (Lin et al. 2015; Tao et al. 2013) for depth estimation. However, since the local depth estimation is sensitive to occlusion, noise and texture-less regions, further complex optimization methods (Wanner, Straehle, and Goldluecke 2014; Wang, Efros, and Ramamoorthi 2015) are necessary in order to obtain smooth depth maps. Moreover, these methods also suffer from high computational cost due to the discretization of depth space.

Recently, some learning-based methods (Shin et al. 2018; Luo et al. 2017) have been introduced for depth estimation in LFs. Similar to traditional approaches, some (Luo et al. 2017; Feng et al. 2018) are designed by learning the slope of lines in EPIs, while others (Shi, Jiang, and Guillemot 2019; Guo et al. 2020) directly explore the correspondences among all views in LFs. Since LFs provide a wealth of viewing angle information, how to select suitable views for matching calculation in different regions becomes an important issue. However, most of these methods choose to directly fuse all features together (Shin et al. 2018) or according to one simple attention for the whole image (Yu-Ju et al. 2020). Therefore, some matching errors in complex areas with occlusions and less textures are introduced when performing depth estimation.

In this work, we introduce a multi-level fusion network based on attentions for LF depth estimation. We consider four directions (0°, 90°, 45° and 135°) of LFs and group them into four branches. Combining with the four-branch structure, two different feature fusion methods are proposed:

- The intra-branch feature fusion based on channel attention, in which features of views that contain less occlusions are selected within one branch.
- The inter-branch feature fusion based on branch attention, in which features between branches are further fused by choosing branches that have less occlusions and richer textures.

With further feature extraction and cost aggregation, the final depth maps are generated. Experimental results on both synthetic and real-world datasets show the proposed method recovers more accurate estimated depth maps than other state-of-the-art methods. The proposed feature fusion strategies are also proved that are effective to improve the depth estimation results, especially in occlusion boundaries. The

[*] Corresponding author: zhangshuo@bjtu.edu.cn

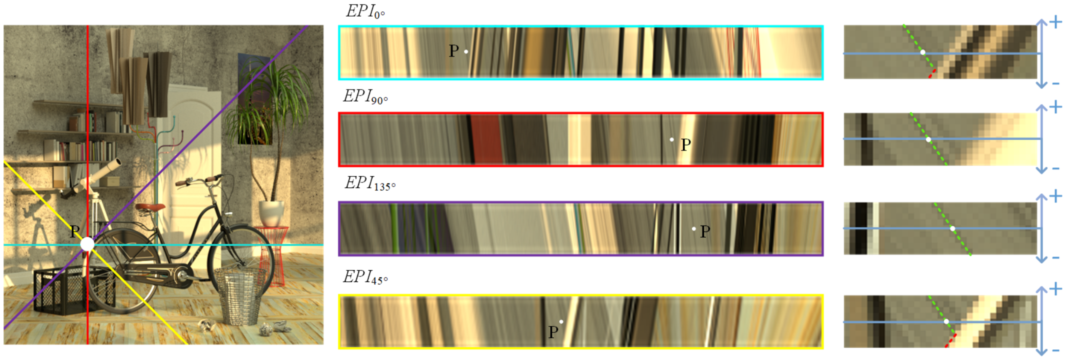


Figure 1: One typical example with the related EPIs, which is occluded when views change along the lower-right direction. Compared with $90°$ and $135°$ EPIs, the line of point P is broken in the $0°$ and $45°$ EPIs. For $0°$ and $45°$ EPIs, if views are separated into two groups with '+' and '-' two sides, the line in the '+' side is more complete than the '-' side.

evaluation metrics on the commonly used HCI 4D Light Field Benchmark also show that our method achieves the highest accuracy on average and gets the rank one position to date.

Related Work

In this section, previous depth estimation methods are reviewed, which include conventional methods and deep learning based methods.

Conventional Methods

By slicing views of LFs in one direction, the depth information are intuitively displayed in the EPIs (Levoy and Hanrahan 1996; Gortler 1996). Wanner et al. (2014) first proposed to estimate direction of lines on EPIs based on structure tensors and the local estimation is then integrated using fast denoising and global optimization. Zhang et al. (2016) proposed a spinning parallelogram operator to estimate the slope of lines on EPIs by assuming that the difference between the two sides of the line is the largest. Zhang et al. (2017) proposed the locally linear embedding for depth estimation, which improves the accuracy of the estimation results and reduces the calculation time without global optimization.

Different from EPI-based methods, some methods are designed based on view images and estimate the disparity by matching pixels in different views. Jeon et al. (2015) estimated disparity by computing matching cost volumes between the central view image and view images displaced using the phase shift theorem. Wang et al. (2015) introduced a depth estimation method which treats the non-occluded and occluded region differently to handle occlusions. Williem and Park (2016) used angle entropy measurement and adaptive defocus response to construct data costs, which is robust to occlusion and less sensitive to noise.

However, most of these conventional methods suffer from long calculation time with complex optimization and are sensitive to occlusion, texture-less and noisy regions.

Deep Learning Based Methods

Recently, deep learning methods are used in LF depth estimation. Based on EPIs, several methods have been proposed to estimate orientations of lines in EPIs. Luo et al. (2017) used EPI (horizontal and vertical) patch pairs to train a CNN. Then post-processing is used to obtain better results. Feng et al. (2018) leveraged on synthetic LFs and proposed a two-stream CNN network that learns to estimate the disparities of multiple neighborhood pixels from EPIs. Leistner et al. (2019) introduced the idea of EPI-Shift that virtually shifts the LF stack. The proposed network predicts the integer and offsets of the disparity separately and then combines them. Li et al. (2020) designed a novel module to construct the relationship between oriented lines in horizontal and vertical EPIs. However, these EPI-based methods only consider the EPI characteristics of horizontal and vertical directions, so that the information is not sufficient and reduces the reliability of the results. Moreover, due to the lack of global information constraints, subsequent optimization processing is required.

The other methods are designed by direct exploring the correspondences between all views in LFs. Shi (2019) estimated disparity for all LF views based on the finetuned FlowNet 2.0 (Ilg et al. 2017), which is suitable for both densely and sparsely sampled LF data. Shin et al. (2018) introduced a multi-stream input structure that concatenates views from four directions in different branches to explore EPI information for depth estimation. Guo et al. (2020) proposed an occlusion-aware network by leveraging the explicitly learned occlusion maps, which is capable of estimating accurate depth maps with sharp edges. The recently proposed attention-based view selection network (Yu-Ju et al. 2020) used all LF images to construct cost volumes and then generated an attention map indicating the importance of each view. However, since all pixels in one view are assigned the same weight, it is difficult to extract special features for different occluded regions. Most of these methods do not make full use of the large amount of rich viewing angle information provided by LFs, which makes it difficult to extract effective information especially in the case of occlu-

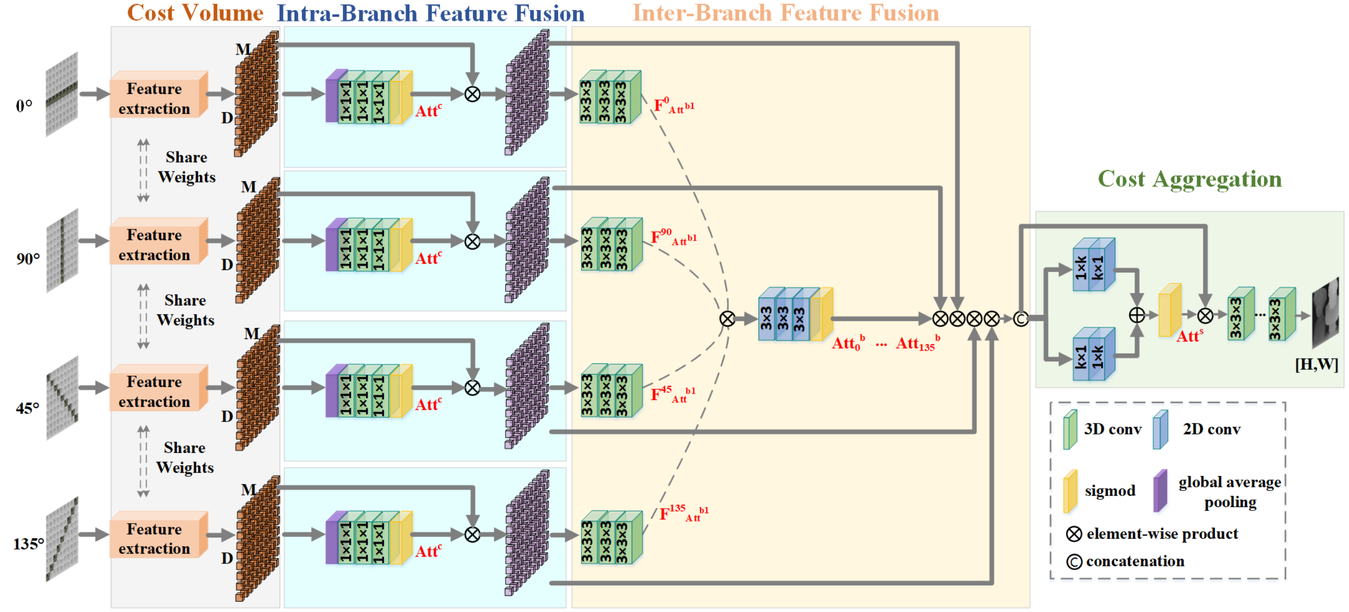


Figure 2: The architecture of our four-branch LF depth estimation network. The cost volume is first constructed after feature extraction in the four branches. Then features of views are fused in each branch based on the intra-branch feature fusion. After that, features of the four branches are further merged through the inter-branch fusion. Finally, the depth map of center view is generated through the cost aggregation module.

sion and weak texture areas.

Motivation

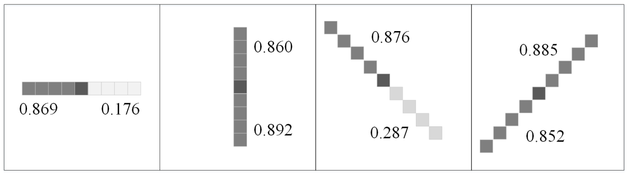
Since the LF contains many views from different perspectives, a lot of redundant information can be used to find the correct color consistency. Therefore, how to choose the effective angular information from LF images for depth estimation is a problem worthy of further investigation. We choose one typical example with occlusions in Figure 1 for further analyses. As shown, P is occluded when view changes along the lower-right direction (↘) with one specific degree in $(0^\circ, 45^\circ)$. The related EPIs in four directions are also shown, which are able to directly reflect the correspondence through the slope of lines. If views in one direction are grouped, the EPIs can be regarded as the slices along the specific direction in the group of views. We then explore the special properties within and between these groups.

- In this example, the lines of P in the 0° and 45° EPIs are broken by the foreground objects. By contrast, in other 90° and 135° EPIs, no occlusions exist and the diagonal lines in the related EPIs are complete. Therefore, the group with more complete and clear correspondence should be given more attention.
- Since occlusions exist in the lower-right direction (↘), for the upper-left (↖) views in 0° and 45° EPIs, P is not occluded and the related line in the EPI is well kept. Therefore, when we slice the EPIs into '$+$' and '$-$' two sides, the correspondence in one side '$+$' can still be used to provide correct information without occlusion errors.

For different points in one image, occlusions may come from different directions. Therefore, we design our network with two attentions at different levels within and between these branches for each point. Based on the above observations, we first separate the views in one branch into two sides and introduce the attention mechanism to calculate the importance of each side for fusion, i.e. the intra-branch fusion strategy. Then the four branches are fused to generate attention maps through the inter-branch fusion module, in which the branches with less occlusions and clear correspondences are assigned with large weights.

The Proposed Method

Figure 2 shows the proposed network structure in detail, which consists of four parts. Suppose that $L(x, y, u, v, c) \in \mathbb{R}^{H\times W\times M\times M\times C}$, where (x, y) and (u, v) represents the coordinates in spatial domain and angular domain, respectively (Levoy and Hanrahan 1996). The M views in four different directions $(0^\circ, 45^\circ, 90^\circ, 135^\circ)$ are grouped and fed into four branches respectively. The feature extraction is first performed on the input views to obtain feature maps and the initial cost volumes are constructed in the four branches. We then design an intra-branch feature fusion module based on channel attention to fuse features of each view together in one branch. The cost volumes in four branches are further fused in the inter-branch feature fusion module using a branch attention strategy, in which the cost volumes with less occlusions are preferred. Finally, the disparity map of the center view is calculated after the cost aggregation module.



(a) 0° (b) 90° (c) 45° (d) 135°

Figure 3: The calculated channel attentions in the four branches of point ”P” in the scene ”Bicycle” in Figure 1. $(a \sim d)$ shows the attention map inside the $0, 45, 90, 135$ direction branch, respectively.

Feature Extraction and Cost Volume Construction

In order to fully extract features from views in different scales, the commonly used Spatial Pyramid Pooling module (SPP) (Yu-Ju et al. 2020; Chang and Chen 2018) is introduced. Same with (Yu-Ju et al. 2020), the feature extraction module is designed with several 2D convolutions, residual blocks and SPP module with 4 levels. The output that have F features of the SPP module with hierarchical context information are then concatenated, which can be denoted as $\mathcal{F}_{spp} \in \mathbb{R}^{M \times H \times W \times F}$. Through the hierarchical feature extraction strategy, more useful features from nearby regions is supplemented for challenging areas, e.g. texture-less and reflection areas.

After extracting the features of each view, we further implement the view shifting strategy as (Yu-Ju et al. 2020) to enlarge the receptive field. Specifically, the features $\mathcal{F}_{spp}$ of each view are manually shifted with $D = 9$ disparity levels from -4 to 4 according to their relative position with the central view. The cost volumes $\mathcal{F}_{spp}$ are then expanded as $\mathcal{F}^{s}_{spp} \in \mathbb{R}^{M \times D \times H \times W \times F}$. When the views are shifted, the large disparities are reduced in one specific cost volume and the relationships between adjacent views are easier to extract using a relatively small receptive field.

Intra-Branch Fusion based on Channel Attention

When the features of each view are extracted, previously methods (Shin et al. 2018) directly concatenate all features together. However, as analyzed in Figure 1, some regions are visible in some views but occluded in other views. If all the features are directly fused, the cost volumes of these occluded regions become ambiguous and the correct disparities are difficult to find. As analyzed before, in simple occlusion scenes with one occlusion, when the views change in one direction, the points can only be occluded on the one side of central view. Therefore, we propose the Intra-Branch Fusion module, in which features are fused inside the branch to choose views on one side which are less likely to have occluded regions.

Specifically, we design the Intra-Branch Fusion module H_{Att^c} with one 3D global average pooling layer, three 1×1 convolutional layers and a sigmoid layer. The channel attention is calculated as:

$$Att^c = H_{Att^c}(\mathcal{F}^{s}_{spp}), \tag{1}$$

where $Att^c \in \mathbb{R}^{3 \times H \times W}$ represents the importance of the feature maps on the central view and two sides of the central view.

Instead of estimating M weight for each view in one branch, we divide the M views into two groups and estimate two weights for the two groups. Using this constraint, the learnable parameters are reduced and the network is easier to train with reasonable results. We show one example in Figure 3, where the attentions of the two sides of point P in Figure 1 in four directions are shown. Since occlusions exist in $0°$ and $45°$ EPIs, the attentions of the two sides in these two branches have a large difference. For views on the non-occluded side (↖), the related weights (0.869 and 0.876) are higher compared with weights (0.176 and 0.287) of the occluded side (↘). By contrast, since in $90°$ and $135°$ EPIs, views on both sides have no occlusions and the related weights are assigned with similar values.

The attention Att^c is then multiplied with the related cost volume using the element-wise product:

$$\mathcal{F}_{Att^c} = \mathcal{F}^{s}_{spp} \otimes Att^c, \tag{2}$$

where $\mathcal{F}_{Att^c}$ has the same size with $\mathcal{F}^{s}_{spp}$ and $\otimes$ is element-wise product.

Inter-Branch Fusion based on Branch Attention

After the intra-branch fusion, $\mathcal{F}^{i}_{Att^c}$, $i =$ 0°,90°,45°,135°, from four different branches are obtained. In this section, we further fuse features from four branches to integrate the information effectively. As previous analyses, the features of the same pixel in the four branches are different, in which some may have occlusions and some may have insufficient textures. Therefore, instead of using a simple concatenation operation, we design the inter-branch fusion module to fuse the features from different branches.

In order to calculate the attentions of four branches for each point, the cost volume $\mathcal{F}^{i}_{Att^c}$ first passes through three 3D convolutional layers, labelled as $H_{Att^{b1}}$:

$$\mathcal{F}^{i}_{Att^{b1}} = H_{Att^{b1}}(\mathcal{F}^{i}_{Att^c}), \tag{3}$$

where $\mathcal{F}^{i}_{Att^{b1}} \in \mathbb{R}^{D \times H \times W}$. The four features $\mathcal{F}^{i}_{Att^{b1}}$ are then fused through corresponding point multiplication, so that the information of the same pixel in the four branches can better interact. Then the attention maps are generated after several 2D convolutional layers and one sigmoid layer, labelled as $H_{Att^{b2}}$:

$$[Att^{b}_{0}, ..., Att^{b}_{135}] = H_{Att^{b2}}(\mathcal{F}^{0}_{Att^{b1}} \otimes ... \otimes \mathcal{F}^{135}_{Att^{b1}}), \tag{4}$$

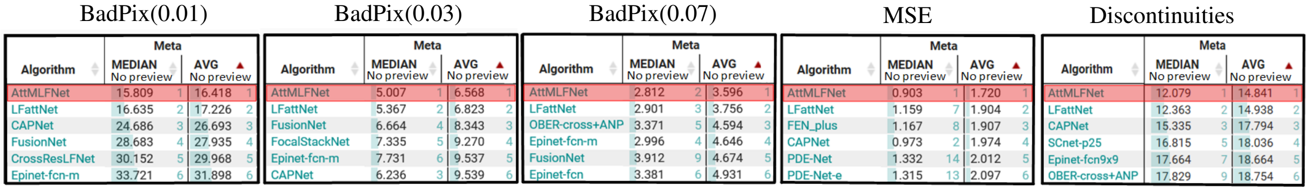
where the $Att^{b}_{i} \in \mathbb{R}^{H \times W}$ and $i = 0°, 90°, 45°, 135°$ represents branch attention of four different angular branch respectively. Finally, the four cost volumes are multiplied by the four branch attentions Att^{b}_{i}:

$$\mathcal{F}_{Att^b} = [\mathcal{F}^{0}_{Att^c} \otimes Att^{b}_{0}, ..., \mathcal{F}^{135}_{Att^c} \otimes Att^{b}_{135}], \tag{5}$$

where $\mathcal{F}_{Att^b} \in \mathbb{R}^{4M \times D \times H \times W \times F}$ is the fused cost volume for further aggregation. In this way, the information provided by the four branches is selectively merged and branches with more clear matching information make more contribution to the cost volume. More experiments for the proposed inter-branch fusion are conducted in the following Ablation Study.

	Backgammon				Dots				Pyramids				Stripes				Time
	0.07	0.03	0.01	mse	0.07	0.03	0.01	mse	0.07	0.03	0.01	mse	0.07	0.03	0.01	mse	
Epinet-fcn-m	3.501	5.563	19.43	3.705	2.490	9.117	35.61	1.475	0.159	0.874	11.42	0.007	**2.457**	**2.711**	**11.77**	0.932	10.66
EPI-Shift	22.89	40.53	70.58	12.79	43.92	53.18	74.55	13.15	1.242	7.315	40.48	0.037	22.72	47.70	78.95	1.686	29.43
EPN+OS+GC	3.328	10.56	55.98	3.699	39.25	82.74	84.91	22.37	0.242	3.169	28.56	0.018	18.54	19.596	28.17	8.731	297.3
PS_RF	7.142	13.93	74.65	6.892	7.975	17.54	78.80	8.338	**0.107**	6.235	83.23	0.043	2.964	5.790	41.64	1.382	993.4
SPO	3.781	8.639	49.94	4.587	16.27	35.06	58.07	5.238	0.861	6.263	79.20	0.043	14.97	15.46	21.87	6.955	2134
LFattNet	**3.126**	**3.985**	**11.58**	**3.648**	**1.432**	3.012	15.05	1.425	0.195	0.488	2.063	0.004	2.933	5.417	18.21	0.892	5.746
ours	3.228	4.625	13.73	3.863	1.606	**2.021**	**10.61**	**1.035**	0.174	**0.429**	**1.767**	**0.003**	2.932	4.743	15.44	**0.814**	**4.551**
	Boxes				Cotton				Dino				Sideboard				Time
	0.07	0.03	0.01	mse	0.07	0.03	0.01	mse	0.07	0.03	0.01	mse	0.07	0.03	0.01	mse	
Epinet-fcn-m	12.34	**18.11**	46.09	5.968	0.447	2.076	25.72	0.197	1.207	3.105	19.39	0.157	4.462	10.86	36.49	0.798	10.65
EPI-Shift	25.95	44.15	74.36	9.790	2.176	10.68	46.86	0.475	5.964	22.15	64.16	0.392	11.79	36.64	73.42	1.261	18.97
EPN+OS+GC	15.30	29.01	67.35	9.314	2.060	9.767	54.85	1.406	2.877	12.79	58.79	0.565	7.997	23.87	66.35	1.744	261.9
PS_RF	18.94	35.23	76.39	9.043	2.425	14.98	70.40	1.161	4.379	16.44	75.96	0.751	11.75	36.28	79.97	1.945	1187
SPO	15.89	29.52	73.23	9.107	2.594	13.71	69.05	1.313	2.184	16.36	69.87	0.310	9.297	28.81	73.36	1.024	2068
LFattNet	**11.04**	18.97	**37.04**	3.996	0.271	0.697	3.644	0.209	0.848	2.339	12.22	0.093	2.869	7.243	**20.73**	0.530	5.868
ours	11.14	18.65	37.66	**3.842**	**0.195**	**0.374**	**1.522**	**0.059**	**0.440**	**1.193**	**4.559**	**0.045**	**2.691**	**6.951**	21.56	**0.398**	**4.542**

Table 1: Numerical comparison of $BadPix(0.07, 0.03, 0.01)$, $MSE * 100$ and running time on different scenes.



BadPix(0.01)

Algorithm	Meta			
	MEDIAN No preview		AVG ▲ No preview	
AttMLFNet	15.809	1	16.418	1
LFattNet	16.635	2	17.226	2
CAPNet	24.686	3	26.693	3
FusionNet	28.683	4	27.935	4
CrossResLFNet	30.152	5	29.968	5
Epinet-fcn-m	33.721	6	31.898	6

BadPix(0.03)

Algorithm	Meta			
	MEDIAN No preview		AVG ▲ No preview	
AttMLFNet	5.007	1	6.568	1
LFattNet	5.367	2	6.823	2
FusionNet	6.664	4	8.343	3
FocalStackNet	7.335	5	9.270	4
Epinet-fcn-m	7.731	6	9.537	5
CAPNet	6.236	3	9.539	6

BadPix(0.07)

Algorithm	Meta			
	MEDIAN No preview		AVG ▲ No preview	
AttMLFNet	2.812	2	3.596	1
LFattNet	2.901	3	3.756	2
OBER-cross+ANP	3.371	5	4.594	3
Epinet-fcn-m	2.996	4	4.646	4
FusionNet	3.912	9	4.674	5
Epinet-fcn	3.381	6	4.931	6

MSE

Algorithm	Meta			
	MEDIAN No preview		AVG ▲ No preview	
AttMLFNet	0.903	1	1.720	1
LFattNet	1.159	7	1.904	2
FEN_plus	1.167	8	1.907	3
CAPNet	0.973	2	1.974	4
PDE-Net	1.332	14	2.012	5
PDE-Net-e	1.315	13	2.097	6

Discontinuities

Algorithm	Meta			
	MEDIAN No preview		AVG ▲ No preview	
AttMLFNet	12.079	1	14.841	1
LFattNet	12.363	2	14.938	2
CAPNet	15.335	3	17.794	3
SCnet-p25	16.815	5	18.036	4
Epinet-fcn9x9	17.664	7	18.664	5
OBER-cross+ANP	17.829	9	18.754	6

Figure 4: The screenshot of the benchmark (http://hci-lightfield.iwr.uni-heidelberg.de) ranking result in September 2020. Our method "AttMLFNet" is highlighted and ranks the average first on five different metrics.

Cost Volume Aggregation

In order to further aggregate the cost volume information, we also implement the widely used spatial attention strategy (Zhao and Wu 2019) in the proposed network structure. As shown in Figure 2, the spatial attention branch H_{Att^s} consists of two parallel branches, which includes 2D convolutional layers with $1 \times k$ and $k \times 1$ kernels ($k = 9$ in the experiment). The result attention is then multiplied with the cost volume using the element-wise product. The spatial attention further strengthens the connection between adjacent pixels and fully extracts context information. After that, the cost volume then passes through eight 3D CNN layers H_{3D}:

$$\mathcal{F}_{\mathit{final}} = H_{3D}(\mathcal{F}_{Att^b} \otimes H_{Att^s}(\mathcal{F}_{Att^b})), \tag{6}$$

where $\mathcal{F}_{\mathit{final}} \in \mathbb{R}^{D \times H \times W}$. Finally, the disparity regression is used to estimate continuous disparity maps:

$$\widehat{d} = \sum_{d \in D} d \times \sigma(-c_d), \tag{7}$$

where c_d is the slice of F_{final} along dimension D and represents the cost of disparity label d. The probability of each disparity label is calculated by softmax $\sigma(\cdot)$. Then the disparity $\widehat{d}$ is estimated according to the probabilities.

Experiments

In this section, we first introduce the detailed implementation of the experiments. Then the performance of our proposed method is compared with other state-of-the-art methods. Finally, we verify the effectiveness of the proposed attention module through ablation study.

Details of Implementation

We use the 4D synthetic LF Dataset (Honauer et al. 2016) in our experiment, in which images have 9×9 views and 512×512 spatial resolution. Same with other networks (Shin et al. 2018; Yu-Ju et al. 2020), 16 images in "Additional" are used for training, 8 images in "Stratified" and "Training" for validating and 4 images in "Test" for testing.

During training, patch-wise training is used by randomly cropping 32×32 gray-scale patches from the LF images. Same with the training strategy in (Shin et al. 2018; Yu-Ju et al. 2020), the reflection, refraction and texture-less areas are manually removed during training in order not to confuse the consistency of matching. The L1 loss that measures the difference of estimated disparity $\widehat{d}$ and ground truth disparity d_{gt} is used in our network. We use Adam optimizer (Kingma and Ba 2014) to optimize the network and set the batch size to 16. The learning rate is kept at $1e^{-3}$. The tensorflow is used to implement the proposed network. The model is trained on an NVIDIA GTX 1080Ti GPU and takes about one week for training.

Evaluation

In the experiments, the $BadPix(\varepsilon)$ defined in (Honauer et al. 2016) and Mean Square Errors (MSE) are used for quantitative evaluation. $BadPix(\varepsilon)$ measures the percentage of wrongly estimated pixels whose errors exceed ε, i.e.

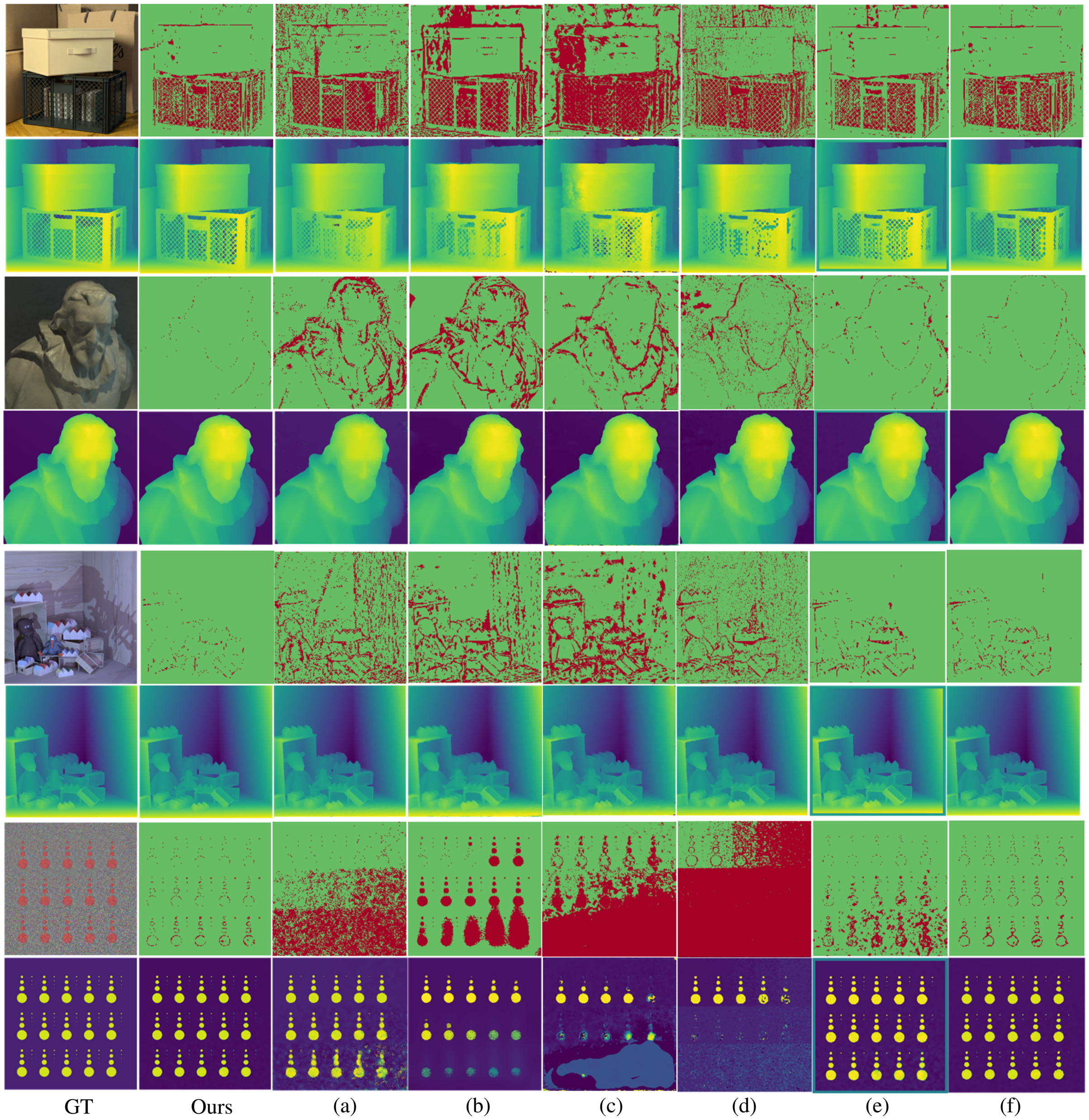


Figure 5: Visual comparison of scenes ”Boxes”, ”Cotton”, ”Dino”, ”Dots” with state-of-the-art methods, including (a) SPO (b) PS_RF (c) EPI-Shift (d) EPN+OS+GC (e) Epinet-fcn-m (f) LFattNet. The $BadPix(0.03)$ error maps and the disparity maps are shown. In the error map, green regions represent the correct estimation, while red parts represent error pixels. Our results have fewer error pixels especially on flat surfaces and along occlusion boundaries.

$|d_{gt}(i) - d_e(i)| > \varepsilon$. The ε is set as 0.07, 0.03 and 0.01. The ground truth depth in “Stratified” and “Training” is public and the ground truth depth in “Test” is only used to test results that are submitted to the benchmark website.

We compare the proposed method with several state-of-the-art methods which are top-ranked on the 4D Light Field Benchmark, including SPO (Zhang et al. 2016), Epinet-fcn-m (Shin et al. 2018), EPI-Shift (Leistner et al. 2019), EPN+OS+GC (Luo et al. 2017), PS_RF (Jeon et al. 2017), and LFattNet (Yu-Ju et al. 2020). Table 1 illustrates specific numerical comparison results. Our method achieves the lowest errors in most scenes, especially for images with lots of occlusions, such as “Dino” and “Sideboard”. For “Cotton” with unclear texture, “Dots” with lots of noise and “Pyramids” with slanted surface, our method obviously outperforms other approaches. The average comparison of the whole validation and testing images is shown in Figure 4, which is the screenshot of the top six on the Benchmark website in September 2020. Our method “AttMLFNet” ranks the first among all methods in four mainly com-

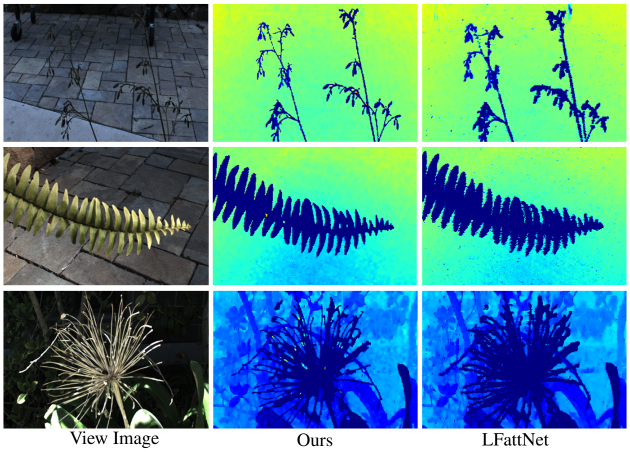


Figure 6: Depth estimation results of real world LF images captured by Lytro Illum. Our depth maps show better performance especially in edges of thin objects than LFattNet.

	BaseNet	BattNet	Proposed
Avg BadPix(0.07)	4.010	3.364	**2.801**
Avg MSE	1.926	1.561	**1.257**

Table 2: Results with the BaseNet, BattNet and the proposed network. The average evaluation metrics are calculated on scenes of Table 1.

pared metrics ($BadPix007$, $BadPix003$, $BadPix001$, $MSE^{*}100$) in average. We also show that the proposed outperforms other methods in discontinuity regions defined in (Honauer et al. 2016) .

Some example estimated depth maps and the related $BadPix(0.03)$ error maps are shown in Figure 5. The Epinet (Shin et al. 2018) does not consider the relationship between the views of different branches and directly concatenates all features together for further processing. In LFattNet (Yu-Ju et al. 2020), all points in one view are assigned with one weight, in which different features cannot be specifically chosen for different regions. Therefore, these methods have estimation errors in occluded areas, texture-less areas and along object edges. By contrast, our method effectively fuses features inside and between the four branches for different regions, which fully utilizes information in the angular domain for depth estimation. As shown, our results show sharper object boundaries and more smooth surfaces than Epinet and LFattNet.

The real-world LFs (Wang, Efros, and Ramamoorthi 2015) captured by Lytro Illum cameras (Ng 2018), are also used for evaluation. Compared with synthetic images, real-world images contain a lot of noise. Since the ground truth depth of real-world LF images is unavailable, we use the trained models same with the last experiment. The related depth maps are compared in Figure 6. As shown, our depth maps also show obviously more clear occlusion boundaries and more smooth planes than LFattNet (Yu-Ju et al. 2020) in the real-world dataset.

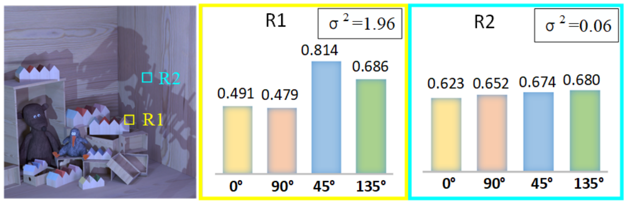


Figure 7: Comparison of two different types of areas ($R1$: occlusion; $R2$: no occlusion). We show the distribution of average weight values of four branches in the two regions.

Ablation Study

In order to validate the effectiveness of our proposed intra-branch and inter-branch fusion strategies, we design three experiments for ablation study as: (1) BaseNet is the base network without intra-branch and inter-branch fusion, in which we directly concatenate cost volumes from four branches after feature extraction. (2) BattNet is the network with only inter-branch fusion. (3) AttMLFNet is our final proposed network including both intra-branch and inter-branch fusion. Other experiment settings are kept the same. As in Table 2, Our proposed network BattNet with inter-branch fusion based on branch attention performs much better than the BaseNet, which integrates features from different directions more effectively. The intra-branch fusion based on channel attention further improves the estimation results of BattNet, which uses channel attention to choose more correct features within the branch.

We further conduct another experiment to show the effectiveness of the proposed inter-branch fusion. We first choose two specific areas with occlusions ($R1$) or without occlusions ($R2$), shown in Figure 7. The weight of each pixel in different branches is separately calculated. We then calculate the average weights of points in the specific areas in different branches as $w_{0,90,45,135}$. The variance σ^2 of $w_{0,90,45,135}$ is also calculated, which quantitatively reflects the attention difference of the four branches. Note that $R1$ contains occlusion structure, but there is no occlusion in the 45° direction. In the weight histogram of $R1$, w_{45} is much higher than the other weights. Conversely, in the non-occlusion region $R2$, the average weight values of all branches are comparable. At the same time, σ^2 of the $R1$ is much larger than $R2$, which implies that our network has indeed learned the different effective information among different branches.

Conclusion

This paper proposed a multi-level fusion network based on two types of attentions for light field depth estimation. Through the analysis of EPIs in different regions, the intra-branch and inter-branch feature fusion strategies are proposed to select features that contain less occlusions and clear correspondence cues for cost volume construction. Experimental results demonstrated the effectiveness of the proposed attention mechanism for accurate depth estimation. The quantitative and qualitative comparison also showed that our method achieves state-of-the-art performance in the HCI 4D Light Field Benchmark and real-world images.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No.61902018), and in part by China Postdoctoral Science Foundation (Grant No. 2019M650470).

References

Adelson, E. H.; and Bergen, J. R. 1991. The plenoptic function and the elements of early vision. Computational Models of Visual Processing. *Int. J. Comput. Vis* 20.

Chang, J.-R.; and Chen, Y.-S. 2018. Pyramid Stereo Matching Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5410–5418.

Feng, M.; Wang, Y.; Liu, J.; Zhang, L.; Zaki, H. F. M.; and Mian, A. 2018. Benchmark Data Set and Method for Depth Estimation From Light Field Images. *IEEE Transactions on Image Processing* 27(7): 3586–3598.

Gortler, S. J. 1996. The lumigraph. In *Conference on Computer Graphics and Interactive Techniques*.

Guo, C.; Jin, J.; Hou, J.; and Chen, J. 2020. Accurate Light Field Depth Estimation via an Occlusion-Aware Network. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*.

Honauer, K.; Johannsen, O.; Kondermann, D.; and Goldluecke, B. 2016. A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 19–34.

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1647–1655.

Jeon, H.-G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.-W.; and Kweon, I. S. 2017. Depth from a Light Field Image with Learning-based Matching Costs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99): 1–14.

Jeon, H.-G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.-W.; and So Kweon, I. 2015. Accurate Depth Map Estimation From a Lenslet Light Field Camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1547–1555.

Jin, J.; Hou, J.; Yuan, H.; and Kwong, S. 2020. Learning Light Field Angular Super-Resolution via a Geometry-Aware Network. In *In Proceedings of the Association for the Advance of Artificial Intelligence (AAAI)*.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer ence* .

Leistner, T.; Schilling, H.; Mackowiak, R.; Gumhold, S.; and Rother, C. 2019. Learning to Think Outside the Box: Wide-Baseline Light Field Depth Estimation with EPI-Shift. *2019 International Conference on 3D Vision (3DV)* 249–257.

Levoy, M.; and Hanrahan, P. 1996. Light Field Rendering. In *ACM Proceedings of Annual Conference on Computer Graphics and Interactive Techniques*, 31–42.

Li, K.; Zhang, J.; Sun, R.; Zhang, X.; and Gao, J. 2020. EPI-based Oriented Relation Networks for Light Field Depth Estimation. In *British Machine Vision Conference (BMVC)*.

Li, N.; Ye, J.; Ji, Y.; Ling, H.; and Yu, J. 2014. Saliency Detection on Light Field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2806–2813.

Lin, H.; Chen, C.; Bing Kang, S.; and Yu, J. 2015. Depth Recovery from Light Field Using Focal Stack Symmetry. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3451–3459.

Luo, Y.; Zhou, W.; Fang, J.; Liang, L.; and Dai, G. 2017. EPI-Patch Based Convolutional Neural Network for Depth Estimation on 4D Light Field. In *International Conference on Neural Information Processing*.

Ng, R. 2018. Lytro Redefines Photography with Light Field Cameras. http://www.lytro.com. Accessed: Oct. 22, 2018.

Ng, R.; Levoy, M.; Brédif, M.; Duval, G.; Horowitz, M.; and Hanrahan, P. 2005a. Light Field Photography with a Hand-Held Plenoptic Camera. *Computer Science Technical Report CSTR* 2(11).

Ng, R.; Levoy, M.; Brédif, M.; Duval, G.; Horowitz, M.; and Hanrahan, P. 2005b. Light Field Photography with a Hand-held Plenoptic Camera. Research Report CSTR 2005-02, Stanford university. URL https://hal.archives-ouvertes.fr/hal-02551481.

Perwaß, C.; and Wietzke, L. 2018. Raytrix: Light Filed Technology. http://www.raytrix.de. Accessed: Oct. 22, 2018.

Sheng, H.; Zhang, S.; Cao, X.; Fang, y.; and Xiong, Z. 2017. Geometric Occlusion Analysis in Depth Estimation Using Integral Guided Filter for Light-Field Image. *IEEE Transactions on Image Processing* 26(12): 5758–5771. doi: 10.1109/TIP.2017.2745100.

Shi, J.; Jiang, X.; and Guillemot, C. 2019. A Framework for Learning Depth From a Flexible Subset of Dense and Sparse Light Field Views. *IEEE Transactions on Image Processing* PP. doi:10.1109/TIP.2019.2923323.

Shin, C.; Jeon, H.-G.; Yoon, Y.; Kweon, I. S.; and Kim, S. J. 2018. 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)*, 4748–4757.

Tao, M. W.; Hadap, S.; Malik, J.; and Ramamoorthi, R. 2013. Depth from Combining Defocus and Correspondence Using Light-Field Cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 673–680.

Wang, T.-C.; Efros, A. A.; and Ramamoorthi, R. 2015. Occlusion-aware Depth Estimation Using Light-field Cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3487–3495.

Wanner, S.; Straehle, C.; and Goldluecke, B. 2013. Globally Consistent Multi-label Assignment on the Ray Space of 4D

Light Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wanner, S.; Straehle, C.; and Goldluecke, B. 2014. Variational Light Field Analysis for Disparity Estimation and Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.

Williem, W.; and Park, I. K. 2016. Robust Light Field Depth Estimation for Noisy Scene with Occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4404.

Williem, W.; Park, I. K.; and Lee, K. M. 2017. Robust Light Field Depth Estimation Using Occlusion-Noise Aware Data Costs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Yu-Ju, T.; Liu, Y.-L.; Ouhyoung, M.; and Chuang, Y.-Y. 2020. Attention-Based View Selection Networks for Light-Field Disparity Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34: 12095–12103. doi: 10.1609/aaai.v34i07.6888.

Zhang, S.; Lin, Y.; and Sheng, H. 2019. Residual Networks for Light Field Image Super-Resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11038–11047. doi:10.1109/CVPR.2019.01130.

Zhang, S.; Sheng, H.; Li, C.; Zhang, J.; and Xiong, Z. 2016. Robust Depth Estimation for Light Field via Spinning Parallelogram Operator. *Computer Vision and Image Understanding* 145: 148–159.

Zhang, Y.; Lv, H.; Liu, Y.; Wang, H.; Wang, X.; Qian, H.; Xiang, X.; and Dai, Q. 2017. Light-Field Depth Estimation via Epipolar Plane Image Analysis and Locally Linear Embedding. *IEEE Transactions on Circuits and Systems for Video Technology* 27(4): 739–747.

Zhao, T.; and Wu, X. 2019. Pyramid Feature Attention Network for Saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.