Final Project: ECE 57000 Implementation Paper
An Attention-based Selection Network for Light Field Disparity Estimation

Anonymous Authors[1]

Abstract

Light field (LF) cameras are used in many real-world applications nowadays. Therefore, the need for a cost-effective and time-efficient disparity estimation model or algorithm to calculate the depth of the scene image is highly in demand. In the following paper, we reviewed general approaches and papers to estimate the depth of LF images. Furthermore, we discussed the implementation of an attention-based view selection residual neural network (RNN) and compared the implementation metric results to the original proposal. The discussion argued the reliability of the model through testing different LF scene data sets. In addition to describing the adjustment of the programming code to accept training models and utilize running machine physical resources. The paper addressed the intended goals and contribution of the implementation and described a few implementation limitations and trad-offs along with future suggestions.

1. Critical Review and Approach Discussion

In the following, we will compare, study, test, and review multiple theoretical approaches and papers. Both the papers and estimation methods are related to the topic of depth estimation based for LF applications. The selected papers will be summarized and critically reviewed based upon their comprehension, clarity, and ease of implementation from an author's perspective. On the other hand, the approaches discussed will be presented to introduce and clarify the selection network and attention map used in the implementation paper. The order under this section will be as follow:

- **Papers:**
- Attention-Based View Selection Networks for Light-Field Disparity Estimation (Tsai et al., 2020) (Implementation paper)
- Attention-based Multi-Level Fusion Network for Light Field Depth Estimation (Chen et al., 2021)
- Learning Light Field Angular Super-Resolution via a Geometry-Aware Network (Jin et al., 2020)
- **Approaches:**
- Attention-Based View Selection Networks for Light-Field Disparity Estimation (Tsai et al., 2020) (Implementation paper)
- Light-field-depth-estimation network based on epipolar geometry and image segmentation(Wang et al., 2020)
- EPINET: A Fully-Convolutional Neural Network Using Epipolar Geometry for Depth From Light Field Images (Shin et al., 2018)

1.1. Summery and Critical Review

1.1.1. Attention-Based View Selection Networks for Light-Field Disparity Estimation

In this paper, the authors introduce a disparity estimation for depth applications based on attention map methodology. The paper discusses the advantages of converting a consumer-based LF small-base line images to multi-view images with different viewpoints (Tsai et al., 2020). Furthermore, the paper explains decoding methods that is implemented in many depth estimation applications which presents their accuracy to computation ratio to estimate the final result. As a solution, the paper proposes an attention-based model that can effectively utilize The repetitive structure of LF images and The redundancy among sub-aperture views to balance the accuracy and computation trade-off in the model (Figure 1). In addition, the created attention map should arrange the importance of the different views in the scene, which will lead to better disparate accuracy.

Unlike many other proposals this paper mainly focuses on analyzing the intrinsic geometry of LFs that contains the spatial and angular information, the creating of the attention map doesn't require lines calculation of Epipolar plane images (EPIs) with different slopes, which are formed by the projections of the same point from different viewpoints. The method of attention maps effectively reduced the cost of the training models by limiting the aggregation computation

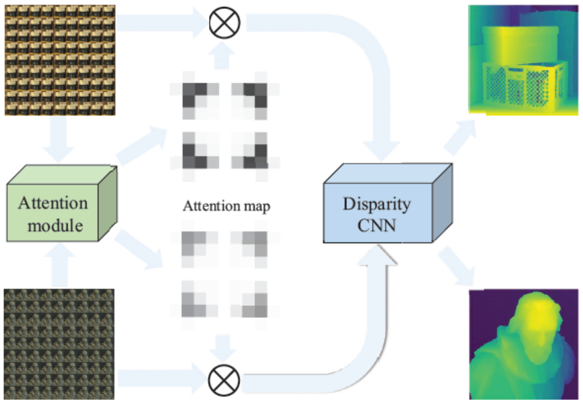

Figure 1. The figure illustrates the method of applying the attention module to generate different LF images that are to be directed into the disparity CNN

cost and limiting the integration process of the different combined models to estimate the disparity of LF image. Simply, the proposed solution takes a multi-view LF image as an input to produce a depth map. The deep learning mechanism behind it suggests passing the input into residual blocks with a dilation convolution added to achieve a larger receptive field. Afterward, the feature map is imported to the spatial pyramid pooling (SPP) model to extract information that contributes toward creating the attention model. Finally, the created attention models are fed into a disparity regression model to generate the attention map for a center view data in the LF image. The method shows high disparity accuracy outcome, along with efficient time and cost implementation. The code for the experiment is provided by the authors at the following Github link:

`github.com/LIAGM/LFattNet`

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

The second paper takes a narrow-base image as a crucial basis for LF inputs module and applies a novel attention-based multi-level fusion network to effectively fuse features for an input LF base and to accurately compute the depth estimation. The approach suggests considering four directions (0, 90, 45, and 135) of LFs, in degrees, and combining them with four branch structures to propose two different fusion methods (Chen et al., 2021) (Figure 2). The first fusion method is produced by combining the four branches with an intra-branch of a newly designed attention mechanism, fewer occlusions features, and richer textures selected between branches for efficiency and better estimation. The second fusion method chooses features of views that contain fewer occlusions selected within one branch. The depth estimation created in each experimental trial is the aggregation of cost and further extraction to recover a better evaluation attention map, especially in occlusion boundary cases.

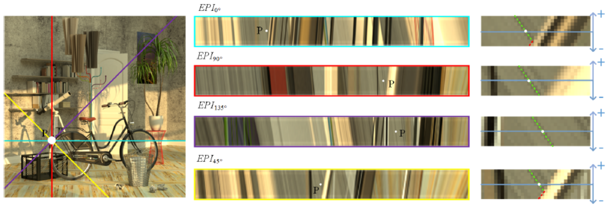

Figure 2. An example of four different direction LF applications, with an occlusion view change along the 45-degree angle

For depth estimation, the paper suggested implanting a novel network on small/narrow base LF images. The consideration of this method relies on the angular resolution of input LFs. Occlusion, noise, and texture don't affect the final estimation due to the network process of combining four-branch structures through the utilization of EPIs. The deep learning method used in this model suggests constructing a cost volume at first and then feature extraction in the four branches of the different resolution angles. Later, the features of views are fused in each branch based on the intra-branch feature fusion, then, merged through inter-branch fusion. In the end, the centered view depth map is generated through the aggregation module (Jin et al., 2020). (Figure 3 graphically explains the approach of multi-level Fusion Network for LFs)

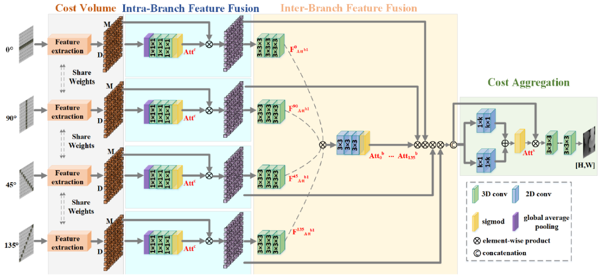

Figure 3. The figure explains the process of estimating the depth through the four different branches

1.1.3. Learning Light Field Angular Super-Resolution via a Geometry-Aware Network

The third paper explains constructing model for large base input views and large base applications. The method suggests an end-to-end train model that analysis the intrinsic geometry of the angular super-resolution of LF images. The method consists of reconstructing the scene image by applying three different modules. The first modules generates a high angular resolution LF to estimate a 4D depth photo. Afterward, the created depth is then utilized to synthesize all novel views by backward warping (Jin et al., 2020). Furthermore, the blending algorithm considers the complementary between images warped from different input views and the angular correlations between warped images at different novel views (Jin et al., 2020). The final estimation results demonstrate the efficiency of the working models to reconstruct a high angular resolution LF with higher quality estimation both qualitatively and quantitatively. Henceforth, the paper utilization of angular resolution on large base LFs contributes to the advanced depth view industry, which can positively contribute to enhancing the technology application in different sectors.For this paper, the authors' method out preforms other state-of-art approaches through exploring the angular relations and scene geometry,but with minimal deep learning techniques and high training costs.

1.2. Approach Discussion

Each of the following papers uses different technique to estimate the depth of LF scenes. Therefore, in the following, we will explain each approach and identify the pros and cons of each method. This step is essential as it clarifies our implementation plan and results through presenting used methods in Machine Learning (ML). Furthermore, we will further discuss the paper approach we chose, in addition, to the reasons behind achieving good metric results, and how the approach will score while training and evaluating disparate data sets.

1.2.1. Attention-Based View Selection Networks for Light-Field Disparity Estimation

In this proposal, the authors introduce a depth estimation solution based on attention map methodology and CNN. The paper discusses the advantages of converting a consumer-based LF small-base line images to multi-view images with different viewpoints (Tsai et al., 2020). As a solution, the paper proposes an attention-based map that can effectively utilize The repetitive structure of LF images and the redundancy among sub-aperture views.The created attention map should arrange the importance of the different views in the scene, which will lead to better disparate accuracy. From there, the the disparity CNN can predict the disparity map using the input views for better characteristics adaption. Figure 1 (Tsai et al., 2020) creates a map to explain the paper's approach.

1.2.2. Attention-Based View Selection Networks for Light-Field Disparity Estimation

In this proposal, the authors introduce a depth estimation solution based on attention map methodology and CNN. The paper discusses the advantages of converting a consumer-based LF small-base line images to multi-view images with different viewpoints (Tsai et al., 2020). As a solution, the paper proposes an attention-based map that can effectively utilize The repetitive structure of LF images and the redundancy among sub-aperture views.The created attention map should arrange the importance of the different views in the scene, which will lead to better disparate accuracy. From there, the the disparity CNN can predict the disparity map using the input views for better characteristics adaption. Figure 1 (Tsai et al., 2020) creates a map to explain the paper's approach.

1.2.3. Light-field-depth-estimation network based on epipolar geometry and image segmentation

The second suggested proposal applies a convolutional method to estimate the depth. Every LF image has an epipolar plane image (EPIs) structure, which contains the spatial and angular information of 2D slices of the LF images. The EPIs include lines with different slopes, which are formed by the projections of different viewpoints. By calculating the slope of such a line in the EPIs, we can obtain the disparity of the pixels in the images (Tsai et al., 2020). The paper suggests adopting different directions for epipolar images based on the image disparity. Furthermore, image segmentation utilizes the accuracy and speed of the function by obtaining the edge information of the central sub-aperture image.

1.2.4. EPINET: A Fully-Convolutional Neural Network Using Epipolar Geometry for Depth From Light Field Images

The last paper explains the method of applying deep learning to estimate the depth in an LF image. The paper intends to overcome the limitation of narrow baseline LF images by introducing a Fully Convolutional Neural Network (Fully-CNN). The paper proposes a learning method to establish an end-to-end mapping between LF images and higher-order regularization variables, which are used to refine the network. The proposal considers the image geometry and the balance between the trade-off of accuracy and speed of implementation.

2. Implementation Review

The previous discussions provided a preliminary overview of different approaches that are used for applications to estimate the depth of LF images. The approaches (Deep Learning, Convolution and Geometry, and CNN) introduced, in the above papers, generate different metric results based on the input HCI data sets. Each research presents trade-offs and limitations that we prioritized to configure our implementation goal and steps. The disparity estimation using an attention-based selection neural network stood out as it presented a better performance in the 4D Light Field Benchmark. In addition, the approach introduced the fastest and most accurate results in comparison to the other methods.

The strategy of implementation consisted of rewriting lines of the program code to accommodate newer programming tools versions. In addition to better utilizing different running machines resources, accept training over different input data sets, and remove all unwanted warnings and errors. The original code runs over a specific library version as noted on GitHub:

Python 3.5.2
Tensorflow-gpu 1.10
CUDA 9.0.176
Cudnn 7.1.4

Therefore, the first step was to replace and rewrite multiple functions and packages to run over the following specifications, while no unwanted errors and/or warnings:

Python 3.6.0 and above
Tensorflow 2.x.x
Tensorflow-addons 0.15
CUDA 10.x.x 11.x.x
Cudnn 8.1

Furthermore, new variables have been introduced to better adjust the training phase for different running machine resources. For instance, to adjust the number of epochs line 248 in (`LFattNet train.py`) can be modified. To adjust the display status of each epoch, the variable (display status ratio) can be adjusted. Consequently, the final model script can be trained and evaluated over different LF scene images from any 4D LF image data sets that are accepted by the residual neural network. To do training and validation modify `LFattNet train.py` (Lines 173,174, 184, 187, 199, 200), and uncomment `func model 81.py` (Line 250). For evaluation, modify `LFattNet evalution.py` (Line 74), and uncomment `func model 81.py` (Line 253).

3. Implementation Results and Discussion

3.1. Evaluation

In the implementation, we worked on implementing and running the paper's code (Attention-Based View Selection Networks for Light-Field Disparity Estimation) with the adjustment made above. The training, validation, and evaluation LF images were chosen from the HCI Light Field data set:

`lightfield-analysis.uni-konstanz.de/`

The process involved trade-offs and limitations due to the limited resources. Therefore, the training model ran 3 epochs and one display status ratio for the preliminary and final implementations. The losses on both implementations over different scene images were relatively low (Table 2). Also, the implementations' models were able to produce close metrics performance to that of the original author's paper (Table 1).

	MSE*100	Badpix 0.07
Original	1.904	3.356
Preliminary	1.750563	3.532567
Final	0.849560	2.390412

Table 1. Comparison between the preliminary and final implementations with the original paper metric results (Tsai et al., 2020). The numerical values illustrate the averages for the research and the achieved results by the two implementations.

Training Losses/epoch	Preliminary	Final
1	2.5640	2.2745
2	1.0254	1.3915
3	0.8037	0.5732

Table 2. Comparison between the losses on the preliminary and final implementations.

The preliminary implementation involved ('additional/museum','additional/kitchen') LF scene images for training, (`'stratified/backgammon'`, `'stratified/dots'`) LF images for validation, and (`'LFattNet output/backgammon'`, `'LFattNet output/dots'`, `'LFattNet output/boxes'`, `'LFattNet output/cotton'`) for testing. The previously mentioned training, validation, and testing data images were chosen by the original paper. Thus, were tested preliminary for performance comparison. Additionally, The final implementation involved ('additional/museum', 'additional/greek') LF scene images for training, (`'stratified/boxes'`, `'stratified/stripes'`) LF images for validation, and (`'LFattNet output/pillows'`, `'LFattNet output/rosemary'`) for testing. The mentioned LF images were chosen randomly to test the algorithm performance over different data sets. The training and evaluation full printing details (Parameters, checkpoints, and outputs) and all the other outputs can be found in the repository GitHub link:

`github.com/malwake-git/`
`ECE570-Implementation-Paper`

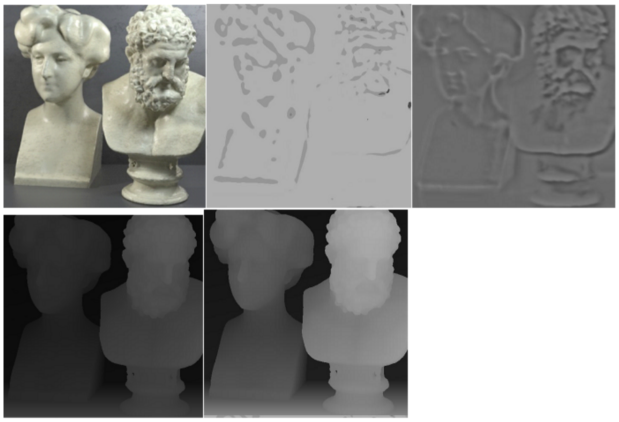

Figure 4. The figure shows the original Greek LF image on the top left and the tested generated image at the end. The phases of training epochs are shown between them

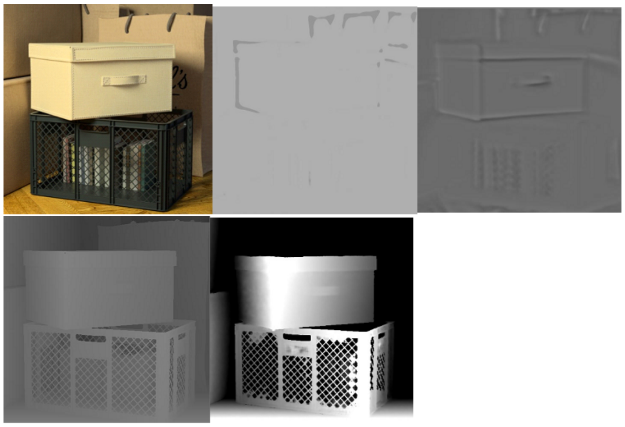

Figure 5. The figure shows the original Boxes LF image on the top left and the tested generated image at the end. The phases of training epochs are shown between them

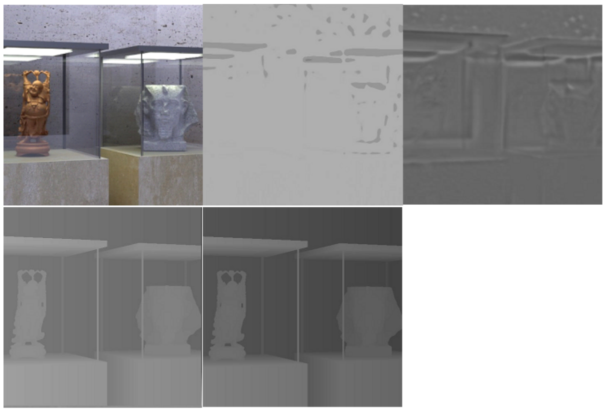

Figure 6. The figure shows the original museum LF image on the top left and the tested generated image at the end. The phases of training epochs are shown between them

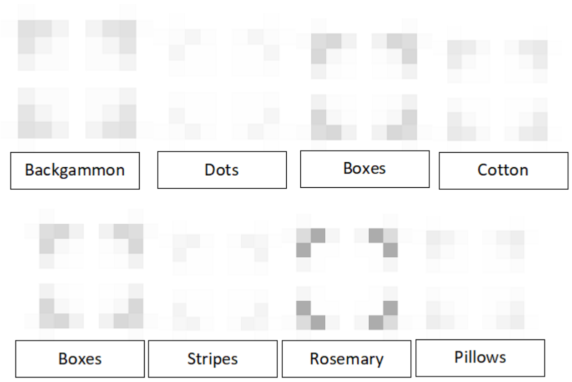

Figure 7. Attention maps extracted from the validation and testing LF images

3.2. Discussion

Although the final implementation model was able to achieve precise metric results, The lack of training and low parameters weights produced make the achieved results less reliable and realistic. The long training time and limited RAM and Disk sizes provided by Google Research Colab made the task harder to implement. Nevertheless, The idea of an attention-based map for the selection neural network obtained more accurate, less expensive outcomes, with relatively less time consumption compared to other algorithms and modules. The produced attention map used for validation and testing (Figure 4). In addition to the final disparity LF images are presented and visually compared to the original paper's implementation outcomes (Figures 5, 6, 7).

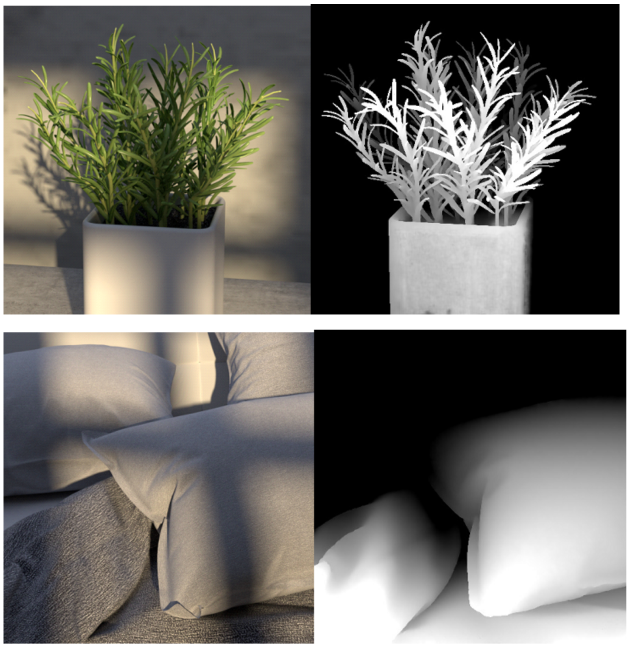

Figure 8. The figure shows the original LF image on the left and the tested generated image next to it, after applying the attention map and the CNN on the data set

References

Chen, J., Zhang, S., and Lin, Y. Attention-based multi-level fusion network for light field depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1009–1017, 2021. URL https://www.aaai.org/AAAI21Papers/AAAI-5186.ChenJiaxin.pdf.

Jin, J., Hou, J., Yuan, H., and Kwong, S. Learning light field angular super-resolution via a geometry-aware network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11141–11148, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6771/6625.

Shin, C., Jeon, H.-G., Yoon, Y., Kweon, I. S., and Kim, S. J. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. URL https://ieeexplore.ieee.org/document/8578597.

Tsai, Y.-J., Liu, Y.-L., Ouhyoung, M., and Chuang, Y.-Y. Attention-based view selection networks for light-field disparity estimation. In *Proceedings of the 34th Conference on Artificial Intelligence (AAAI)*, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6888/6742.

Wang, X., Tao, C., Wu, R., Tao, X., Sun, P., Li, Y., and Zheng, Z. Light-field-depth-estimation network based on epipolar geometry and image segmentation. *J. Opt. Soc. Am. A*, 37(7):1236–1243, Jul 2020. doi: 10.1364/JOSAA.388555. URL http://www.osapublishing.org/josaa/abstract.cfm?URI=josaa-37-7-1236.