Bridging Unsupervised and Supervised Depth from Focus via All-in-Focus Supervision

Ning-Hsu Wang1,2 Ren Wang1 Yu-Lun Liu1 Yu-Hao Huang1
Yu-Lin Chang1 Chia-Ping Chen1 Kevin Jou1
1MediaTek Inc. 2National Tsing Hua University
https://github.com/albert100121/AiFDepthNet

Abstract

Depth estimation is a long-lasting yet important task in computer vision. Most of the previous works try to estimate depth from input images and assume images are all-in-focus (AiF), which is less common in real-world applications. On the other hand, a few works take defocus blur into account and consider it as another cue for depth estimation. In this paper, we propose a method to estimate not only a depth map but an AiF image from a set of images with different focus positions (known as a focal stack). We design a shared architecture to exploit the relationship between depth and AiF estimation. As a result, the proposed method can be trained either supervisedly with ground truth depth, or unsupervisedly with AiF images as supervisory signals. We show in various experiments that our method outperforms the state-of-the-art methods both quantitatively and qualitatively, and also has higher efficiency in inference time.

1. Introduction

Depth estimation has been one of the most fundamental computer vision problems in decades. Many downstream tasks, such as augmented reality (AR), virtual reality (VR) and autonomous driving, highly rely on this research topic. Recently, it also enabled an increasing number of applications for smartphone photography, such as depth-of-field adjustment, background substitution, and changing focus after the picture is taken.

Consequently, depth sensing has become a fundamental component for capturing devices. Active depth sensing solutions such as Time-of-Flight (ToF) and structured light are often expensive and power-consuming due to the need for specialized hardware. Passive techniques, such as binocular or multi-view stereo, are more cost and power-efficient but prone to errors in textureless regions.

Deep learning based stereo matching methods tackle this problem in a data-driven way by learning depth estimation directly from input images. However, they require a large amount of high-quality paired training data, which are time-consuming and expensive to acquire. They also suffer when the training data are imperfect: synthesized and unrealistic input images, or inaccurately registered depth maps.

Some unsupervised learning approaches [9, 42] were proposed to address this problem. They usually use image reconstruction loss and consistency loss without the need for ground truth depth data. They can also mitigate domain gaps by training directly with real-world stereo images without corresponding registered depth maps.

Another relatively under-explored cue for depth estimation is defocus blur. The task of depth-from-focus (or defocus) aims to estimate the depth of a scene from a focal stack, i.e., images taken at different focal positions by the same camera. Conventional optimization based depth-from-focus approaches [37, 36, 26] estimate the level of sharpness for each pixel and often suffer from textureless objects or aperture problems. Deep learning techniques [11, 23] help to overcome these issues but need ground truth depth data for supervised learning. It’s difficult and time-consuming to retrieve focal stacks with registered depth maps, let alone the imperfect depth data obtained by hardware solutions such as ToF sensors [11]. One could synthesize defocus blur on a synthetic dataset with synthetic depth maps [23]. However, it is still questionable whether the thin lens synthesis model could represent real-world optics well.

In this paper, we propose a novel method to estimate depth and an all-in-focus (AiF) image jointly from an input focal stack. We exploit the relationship between these two tasks and design a shared common network. Moreover, the proposed network can be trained either supervisedly with ground truth depth maps or unsupervisedly with only ground truth AiF images. Compared to high-quality labeled depth, acquiring AiF images is relatively easier because AiF images can be captured with smaller apertures along with longer
exposures. However, collecting the corresponding focal stack of an AiF image might be difficult because of the focus breathing phenomenon, where a camera’s field of view changes as the lens moves. To address this problem, we use synthetic data without such effects during training, and apply a calibration process on real data during testing.

Our contribution is three-fold:

- When trained supervisedly, our method outperforms the state-of-the-art methods in various comparisons, while our method also runs faster.
- To our knowledge, the proposed method is the first that can learn depth estimation unsupervisedly from only AiF images and performs favorably against the state-of-the-art methods.
- Domain gaps can be mitigated by our method with test-time optimization on real-world data, especially when ground truth depth data are not available.

2. Related Work

2.1. Depth from Focus

Depth estimation is a fundamental computer vision task that aims to use different cues such as color, semantics, stereo, or the difference in image sequences to predict or fuse depth maps [5, 40]. Most of the previous works assume that input images are all-in-focus, whereas in real-world scenarios, images are usually considered to be defocused in the background or with a shallow depth-of-field (DoF). Nevertheless, some approaches elaborate on depth estimation with defocused images. Conventional optimization based approaches [37, 36] proposed to directly estimate depth from focal stacks, and a variant approach [26] tries to generate an index map in which every pixel is assigned to the focus position leading to the maximal sharpness. Chen et al. [6] found the relationship between relative blur and disparity, and make use of it to enhance the robustness of matching. Depth from focus sweep video [18] targets estimating depth from images with successive focus positions. Recently, deep learning based approaches [11, 24] could model the blurriness more precisely and achieve much better depth quality. On the other hand, some works [2, 4, 35] use deep learning techniques to remove the defocus blur for single images.

2.2. Multi-focus Image Fusion

Although real-world images usually have defocus blur, most computer vision applications are supposed to cooperate with all-in-focus (AiF) images. Therefore, more and more attention is given on shallow-DoF image deblurring and multi-focus image fusion. A multi-focus image fusion approach [41] applies Laplacian to different scales of images. Zhan et al. [43] proposed a guided filter to help edge preserving during multi-focus image fusion. Nejati et al. [27] learn a sparse representation of relative sharpness measurement and produce a pixel-level score map for decision through pooling. An unsupervised learning approach [32] was proposed to fuse either multi-exposure or multi-focus images to generate a high dynamic range (HDR) or an AiF image. Liu et al. [22] proposed to learn a focus map and a segmentation map through deep neural networks, and fuse images by integrating these maps.

2.3. Light Field

A light field camera captures spatially distributed light rays. By re-rendering through digitized multi-view images, variable aperture pictures can be generated after capturing. With this characteristic, a post-capture refocusing can also be accomplished by properly manipulating those light field images [28]. As light field camera provides multi-view information from different poses, it could help many computational photography applications. As the number of views grows, more information could be retrieved.

Light field images often suffer from low spatial resolution as the sampling resources are limited. Cheng et al. [7] and Jin et al. [17] proposed to apply deep learning on constructing super-resolution images from light field images to cover this situation. Depth estimation could also be achieved by leveraging the abundant information from light field images [20, 12, 13, 8, 38, 21, 45, 29, 31]. Levoy et al. [20] proposed to use the epipolar plane images (EPIs), which contain both spatial and angular information, for depth estimation from light field images. Heber et al. [12] proposed to learn the end-to-end mapping between the 4D light field and its corresponding 4D depth field accompanied with a high-order regularization refinement. By following [12], they then designed an encoder-decoder architecture to extract the geometric information from light field images [13]. Some works [15, 33] focus on providing light field datasets with ground truth depth or AiF images for further use.

2.4. Realistic Data Synthesis

Deep learning significantly improves the quality of computer vision tasks with a large amount of training data. However, collecting real-world data is often costly and time-consuming. Therefore, many works instead target synthesizing realistic data to provide sufficient data for training. Barron et al. [3] proposed to synthesize defocus images by a layer-wised rendering scheme with a z-buffer. Wadhwa et al. [39] combine a person segmentation mask with a depth map calculated from a dual-pixel camera to generated shallow depth-of-field images. Gur et al. [10] use a differentiable point spread function, namely the PSF convolution layer, to synthesize realistic defocus images and train a depth from defocus network to generate depth maps. Herrmann et al. [14] proposed a learning-based approach for autofocus and recently provided a real dataset for training. However, the
3. Method

We describe the proposed method in this section. Sec. 3.1 gives an overview of our method. Then, a shared network used for both of depth estimation and AiF image reconstruction is illustrated in Sec. 3.2. The attention mechanism that bridges the two tasks is introduced in Sec. 3.3. We then depict the core concept to turn AiF images into supervisory signals for unsupervised depth estimation in Sec. 3.4. Finally, Sec. 3.5 shows the loss functions for training with depth supervision and AiF supervision.

3.1. Overview

Depth from focus aims to recover the depth from a focal stack through defocus cues. As shown in Fig. 1, given a focal stack $S \in \mathbb{R}^{H \times W \times 3 \times F}$ of $F$ images with gradually varying focus positions $P \in \mathbb{R}^{H \times W \times 1 \times F}$, our method produces several attention representations through a shared network. Then a depth map $D \in \mathbb{R}^{H \times W \times 1}$ and an all-in-focus (AiF) image $I \in \mathbb{R}^{H \times W \times 3}$ of this scene can be generated with these attention maps.

3.2. Network Architecture

DDFF [11] uses 2D ConvNets to address the sharpness measurement problem, which is one of the main challenges in depth from focus. On the other hand, DefocusNet [24] applies a global pooling layer as a communication tool between several weights-sharing 2D ConvNets. Another objective of this architecture is to allow focal stacks with arbitrary sizes. However, some important information across the stack dimension might not be effectively captured due to the limitation of 2D convolution and simple global pooling.

For this reason, we adopt the Inception3D [1] as the backbone of our model. As shown in Fig. 1, our model is an encoder-decoder network consisting of 3D convolutions. With the 3D convolution, defocus cues could be better captured across frames and thus facilitate the tasks of depth estimation and AiF image reconstruction. Moreover, our model can also handle focal stacks with arbitrary sizes attributed to the nature of 3D convolution.

3.3. Attention Mechanism

The output of our network is an intermediate attention $M \in \mathbb{R}^{H \times W \times 1 \times F}$. The underlying expectation of the intermediate attention $M$ is that it should reflect the probability of each focus position leading to the maximal sharpness. Then it can benefit both of depth estimation and AiF image reconstruction.

For depth estimation, we propose to normalize the intermediate attention $M$ into a depth attention $M_{\text{depth}}$ via a softplus normalization:

$$M_{\text{depth}} = \varsigma(M),$$

where

$$M_{\text{depth}} = \frac{\ln (1 + \exp (M_{i,j,1,t}))}{\sum_{n=1}^{F} \ln (1 + \exp (M_{i,j,1,n}))}.$$  \hspace{1cm} (2)

The softplus function is a smooth version of ReLU.

The depth attention $M_{\text{depth}}$ can also be interpreted as the probability distribution because the softplus function ensures non-negativeness and normalizes $M_{\text{depth}}$ into a valid probability distribution. Then the expected depth value of each pixel can be derived via:

$$D_{i,j,1} = \sum_{t=1}^{F} (M_{\text{depth}} \cdot P)_{i,j,1,t}.$$  \hspace{1cm} (3)

For AiF image reconstruction, we perform similar procedures except that the normalization function changes from softplus to softmax. That is, the AiF attention $M_{\text{AiF}}$ is obtained as:

$$M_{\text{AiF}} = \sigma(M),$$

where

$$M_{\text{AiF}} = \frac{\exp (M_{i,j,1,t})}{\sum_{n=1}^{F} \exp (M_{i,j,1,n})}.$$  \hspace{1cm} (5)

The AiF attention $M_{\text{AiF}}$ can then be used for AiF image reconstruction. The expected AiF image of each pixel is expressed as:

$$I_{i,j,k} = \sum_{t=1}^{F} (M_{\text{AiF}} \cdot S)_{i,j,k,t}.$$  \hspace{1cm} (6)

The reason why we adopt the softmax and softplus normalizations separately for depth and AiF image estimation is to tackle the problem of sparse focal stacks. By sparse, we mean that the stack size is small, and the focus positions inside the focal stack are not dense.

For AiF image reconstruction, it is best to select the sharpest pixel along the stack dimension. Blending multiple pixels inside a sparse stack usually does not help. Therefore, we leverage the softmax function to pursue the peaking phenomenon to extract the clearest pixels.

For depth estimation, the softmax normalization results in severe quantization for sparse focal stacks because it simply selects the nearest focal position of maximal sharpness. On the other hand, the softplus normalization leads to a flatter distribution so that more accurate depth can be predicted by interpolating among sparse focal positions. Fig. 2 shows the effects of this design choice.
Figure 1: **An overview of the proposed method.** Given a stack of images with varying focus positions, i.e., focal stack, our model first produces an intermediate attention map $M$. The intermediate attention map can be shared between depth estimation and all-in-focus (AiF) image reconstruction. With different normalization functions, the attention map can be further manipulated to generate either depth or AiF results.

Figure 2: **The effect of normalization on attention maps.** The first row shows that if we adopt the same softmax normalization for depth and AiF attention, both of them tend to be a flatter distribution in supervised learning and face a peaking phenomenon in unsupervised learning (AiF supervision). The second row shows that with separate normalization as described in Sec. 3.3, each of depth and AiF attention becomes a proper distribution in both supervised and unsupervised learning.

### 3.4. Towards Unsupervised Depth Estimation

Similar to our method, DefocusNet [24] also combines the tasks of depth estimation and AiF image reconstruction via *intermediate defocus maps*. Their intermediate defocus maps necessitate supervisory signals, which can be derived by calculating the circle of confusions from the ground truth depth. Furthermore, they also propose to use a DepthNet following the intermediate defocus maps to predict the output depth. Therefore, their method can only be trained supervisedly when ground truth depth data are available.

Instead, we propose to use the intermediate attention $M$ to bridge the two tasks, which does not necessitate intermediate supervisory signals from ground truth depth data. Moreover, there are no learnable parameters after $M$. We only use fixed normalization functions to generate the output depth and AiF images. Therefore, even when ground truth depth data are not available, we can still train the shared network to generate the intermediate attention $M$ only via supervisory signals from ground truth AiF images. That is, our method can be trained supervisedly or unsupervisedly with or without ground truth depth data.

### 3.5. Training Loss

For supervised depth estimation, our model is trained with a simple $L_1$ loss:

$$L_{supervised} = L_{depth} = \mathbb{E}[||D - D_{gt}||_1],$$

where $D_{gt}$ stands for the ground truth depth.

For unsupervised depth estimation, our model can also be trained by a $L_1$ loss with AiF supervision:

$$L_{AiF} = \mathbb{E}[||I - I_{gt}||_1],$$

where $I_{gt}$ denotes the ground truth AiF image. Furthermore, we also encourage our depth map to be locally smooth using an edge-aware weighting as in [9]. The smoothness loss is defined as:

$$L_{smooth} = \mathbb{E}[W_x \left| \frac{\partial D_{i,j,1}}{\partial x} \right| + W_y \left| \frac{\partial D_{i,j,1}}{\partial y} \right|],$$

where

$$W_x = \exp \left( -\frac{\lambda}{3} \sum_k \left| \frac{\partial I_{i,j,k}}{\partial x} \right| \right)$$

$$W_y = \exp \left( -\frac{\lambda}{3} \sum_k \left| \frac{\partial I_{i,j,k}}{\partial y} \right| \right),$$
and $\lambda$ is a hyper-parameter for the edge weighting based on the ground truth AiF image. The total loss of our unsupervised depth estimation is then:

$$L_{\text{unsupervised}} = L_{\text{AiF}} + \alpha L_{\text{smooth}},$$

(11)

where $\alpha$ indicates the importance of the smoothness loss.

4. Evaluation

In this section, we describe the datasets used in the experiments, and report ablation studies and comparisons with the state-of-the-art methods on depth from focus. In the supplementary material, we present additional results for AiF image reconstruction, and visual comparisons on DDFF 12-Scene [11], Middlebury Stereo Datasets [34], and the DefocusNet dataset [24].

4.1. Datasets

Totally five datasets are used in the quantitative experiments and visual comparisons. Their descriptions and statistics are summarized in Table 1. For more details, please see the supplementary material.

4.2. Evaluation Metrics

In this paper, we evaluate quantitative results with the following metrics: mean-absolute error (MAE), mean-squared error (MSE), root-mean-squared error (RMSE), log root-mean-squared error (logRMS), relative-absolute error (Abs. rel.), relative-squared error (Sqr. rel.), bumpiness (Bump), accuracy with $\delta = 1.25$, per-image inference time (Secs.).

4.3. Implementation Details

We implement our method in PyTorch [30]. Our model is trained from scratch using the Adam optimizer [19] ($\beta_1 = 0.9$, $\beta_2 = 0.999$), with a learning rate of $10^{-4}$. For DDFF 12-Scene [11], 4D Light Field Dataset [15], FlyingThings3D [25], and the DefocusNet dataset [24], we use 10, 10, 15, and 5 as the input stack sizes, respectively. In unsupervised learning, the weight of smoothness loss $\alpha$ is set to 0.002 in all experiments. We apply random spatial transformations (flipping, cropping, rotation) and random color jittering (brightness, contrast, and gamma) for data augmentation during training. The training patch size is $256 \times 256$ after random cropping except for DDFF 12-Scene [11]. Because the image size of DDFF 12-Scene is $224 \times 224$, which is larger than $256 \times 256$, we do not apply random cropping on this dataset. All the experiments are conducted with a single NVIDIA GTX 1080 GPU.

4.4. Ablation Studies

To understand the performance of each proposed components in our method, we conduct various ablation studies on 4D Light Field Dataset [16]. All models are supervised trained with disparity maps if not specified.

Architecture. Table 2 shows the comparison between 2D and 3D convolutions and with/without the proposed attention mechanism. The 3D convolution performs better than 2D convolution as the 3D convolution is able to capture and aggregate features across both spatial and stack dimensions. The design of attention bridges the tasks of depth estimation and AiF image reconstruction, and thus enables the possibility of unsupervised depth estimation as described in Sec. 3.4 while maintaining roughly the same prediction quality.

Focal stack size. Our network can handle arbitrary input stack sizes due to the nature of 3D convolution. Fig. 3 shows a comparison on different input stack sizes. We train the models in three different ways: 1) 10: training with a fixed input stack size of 10. 2) Arbitrary: training with arbitrary input stack sizes randomly sampled from 2 to 10. 3) Same: training with the same size as test stacks.

The model trained with a fixed stack size performs poorly while testing with different stack sizes. As expected, the model trained with the same stack size as test data performs the best as the input setting for training and testing are consistent. However, the model trained with arbitrary input stack sizes performs favorably against the ones trained with the same stack size as test data. This demonstrates the robustness of our method across different input stack sizes.

Fig. 4 shows the visual comparison of different input stack sizes, generated with the model trained with arbitrary input stack sizes. It is obvious that the quality of the estimated disparity map improves as the stack size increases.

Unsupervised learning using all-in-focus images. On the basis of the attention mechanism described in Sec. 3.3, our method is able to be trained with either ground truth disparity maps (supervised) or ground truth AiF images (unsupervised). Table 3 and Fig. 5 respectively show the quantitative and qualitative results for both of our supervised and unsu-
Table 1: Summary of the evaluation datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image source</th>
<th>Cause of defocus</th>
<th>Disparity/Depth GT</th>
<th>AiF GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DefocusNet [24]</td>
<td>Synthetic</td>
<td>Blender rendering</td>
<td>Depth</td>
<td></td>
</tr>
<tr>
<td>4D Light Field Dataset [16]</td>
<td>Synthetic</td>
<td>Light-field composition</td>
<td>Disparity</td>
<td>✓</td>
</tr>
<tr>
<td>Middlebury Stereo Datasets [34]</td>
<td>Real</td>
<td>Disparity rendering</td>
<td>Disparity</td>
<td>✓</td>
</tr>
<tr>
<td>Mobile Depth [37]</td>
<td>Real</td>
<td>Real</td>
<td>Stitching by MRF</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Ablation study on network architecture. The 3D convolution leads to better performance on disparity estimation because of its ability to capture features for both spatial and stack dimensions. Although the design of attention does not increase the performance explicitly, it bridges the tasks of depth estimation and AiF image reconstruction and thus enables the unsupervised learning for depth estimation.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Attention</th>
<th>MAE↓</th>
<th>MSE↓</th>
<th>RMSE↓</th>
<th>Bump↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>✓</td>
<td>0.0788</td>
<td>0.0472</td>
<td>0.2014</td>
<td>1.5776</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0851</td>
<td>0.0461</td>
<td>0.1984</td>
<td>2.3168</td>
</tr>
<tr>
<td>2D</td>
<td>✓</td>
<td>0.1070</td>
<td>0.0577</td>
<td>0.2259</td>
<td>2.1435</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1179</td>
<td>0.0576</td>
<td>0.2291</td>
<td>2.3168</td>
</tr>
</tbody>
</table>

Table 3: Ablation study on supervision. The results of supervised learning perform better than the ones from unsupervised learning (AiF supervision). Unsupervised learning often generates disparity maps that suffer from the quantization effect and lead to poor results. After adding the smoothness loss, the output disparity maps become locally smooth and perform better quantitatively.

<table>
<thead>
<tr>
<th>Supervised</th>
<th>Stack size</th>
<th>L_smooth</th>
<th>MAE↓</th>
<th>MSE↓</th>
<th>RMSE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>10</td>
<td></td>
<td>0.0788</td>
<td>0.0472</td>
<td>0.2014</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td></td>
<td>0.2425</td>
<td>0.1174</td>
<td>0.3401</td>
</tr>
<tr>
<td>No</td>
<td>37</td>
<td></td>
<td>0.2099</td>
<td>0.1039</td>
<td>0.3202</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td>✓</td>
<td>0.1671</td>
<td>0.0746</td>
<td>0.2698</td>
</tr>
<tr>
<td>No</td>
<td>37</td>
<td>✓</td>
<td>0.1116</td>
<td>0.0584</td>
<td>0.2311</td>
</tr>
</tbody>
</table>

Figure 4: Visual comparison on different focal stack sizes. The quality of the estimated disparity map improves as the stack size increases.

RGB stack size 3 stack size 5 stack size 9 GT

Figure 5: Visual comparison on different supervision settings. The results of supervised learning are better than the ones from unsupervised learning (AiF supervision). The output disparity maps from unsupervised learning suffer from the quantization effect. By adding the smoothness loss, the disparity results become locally smooth and perform favorably against the ones from supervised learning qualitatively.

4.5. Comparisons to the State-of-the-art Methods

After ablation on different components of the proposed method, we conduct comparisons with the state-of-the-art methods on various datasets in this section.

DFFF 12-Scene. Table 4 shows the quantitative comparison on DFFF 12-Scene [11]. All methods are supervisedly trained on this dataset with ground truth depth. The proposed method performs favorably against the state-of-the-art supervised models. As shown in Table 3, with a large input stack size of 37 and the smoothness loss, the performance of unsupervised learning is able to approximate that of supervised learning with a smaller stack size of 10.

However, the results of unsupervised learning with a smaller stack size suffer from the quantization problem as described in Sec. 3.4. As shown in Fig. 5, after adding the smoothness loss, the output disparity map becomes locally smooth and perform better qualitatively.

4D Light Field Dataset. Compared to DFFF 12-Scene [11], 4D Light Field Dataset [16] provides the ability to simulate shallower DoF images as its larger baseline leads to a larger synthetic aperture. Table 5 shows the quantitative comparison with state-of-the-art methods. All of them are trained with this dataset. Our supervised model achieves the best MSE (0.0472), which outperforms the state-of-the-art method DefocusNet [24] (0.0593). Furthermore, our unsupervised model even performs better than most of the other methods under all metrics. Furthermore, our method runs faster than another deep learning based method DFFF [11].
Table 4: **Quantitative comparison on DDFF 12-Scene.** Note that RMSE is not presented due to the mismatched results from the benchmark website and the paper of DDFF [11]. For DefocusNet [24], we only show the original metric reported in their paper, which is MSE. Red text indicates the best, and blue text indicates the second-best performing method.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE ↓</th>
<th>log RMS ↓</th>
<th>Abs. rel. ↓</th>
<th>Sqr. rel. ↓</th>
<th>Bump. ↓</th>
<th>δ = 1.25↑</th>
<th>δ = 1.25²↑</th>
<th>δ = 1.25³↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>8.6e⁻⁴</td>
<td>0.29</td>
<td>0.25</td>
<td>0.01</td>
<td>0.63</td>
<td>68.33</td>
<td>87.40</td>
<td>93.96</td>
</tr>
<tr>
<td>DefocusNet [24]</td>
<td>9.1e⁻⁴</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DDFF [11]</td>
<td>9.7e⁻³</td>
<td>0.32</td>
<td>0.29</td>
<td>0.01</td>
<td>0.59</td>
<td>61.95</td>
<td>85.14</td>
<td>92.98</td>
</tr>
<tr>
<td>PSPNet [44]</td>
<td>9.4e⁻⁴</td>
<td>0.29</td>
<td>0.27</td>
<td>0.01</td>
<td>0.55</td>
<td>62.66</td>
<td>85.90</td>
<td>94.42</td>
</tr>
<tr>
<td>Lytro</td>
<td>2.1e⁻³</td>
<td>0.31</td>
<td>0.26</td>
<td>0.01</td>
<td>1.02</td>
<td>55.65</td>
<td>82.00</td>
<td>93.09</td>
</tr>
<tr>
<td>PSP-LF [44]</td>
<td>2.7e⁻³</td>
<td>0.45</td>
<td>0.46</td>
<td>0.03</td>
<td>0.54</td>
<td>39.70</td>
<td>65.56</td>
<td>82.46</td>
</tr>
<tr>
<td>DFLF [11]</td>
<td>4.8e⁻³</td>
<td>0.59</td>
<td>0.72</td>
<td>0.07</td>
<td>0.65</td>
<td>28.64</td>
<td>53.55</td>
<td>71.61</td>
</tr>
<tr>
<td>VDFF [26]</td>
<td>7.3e⁻³</td>
<td>1.39</td>
<td>0.62</td>
<td>0.05</td>
<td>0.79</td>
<td>8.42</td>
<td>19.95</td>
<td>32.68</td>
</tr>
</tbody>
</table>

Fig. 6: **Visual comparison on 4D Light Field Dataset.** S: supervised. US: unsupervised (AiF-supervision).

Table 5: **Quantitative comparison on 4D Light Field Dataset.** Our supervised model outperforms the state-of-the-art method DefocusNet. Furthermore, our unsupervised model (with AiF supervision) even performs better than most of the other supervised methods. Please see Fig. 6 for the visual comparison. (∗ represents that the model is pre-trained on DDFF 12-Scene.)

<table>
<thead>
<tr>
<th>Method</th>
<th>Supervised</th>
<th>MSE ↓</th>
<th>RMSE ↓</th>
<th>Bump. ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>Yes</td>
<td>0.0472</td>
<td>0.2014</td>
<td>1.58</td>
</tr>
<tr>
<td>DefocusNet [24]</td>
<td>Yes</td>
<td>0.0593</td>
<td>0.2355</td>
<td>2.69</td>
</tr>
<tr>
<td>*DDFF [11]</td>
<td>Yes</td>
<td>0.19</td>
<td>0.42</td>
<td>1.92</td>
</tr>
<tr>
<td>*PSPNet [44]</td>
<td>Yes</td>
<td>0.37</td>
<td>0.53</td>
<td>1.21</td>
</tr>
<tr>
<td>VDFF [26]</td>
<td>Yes</td>
<td>1.3</td>
<td>1.15</td>
<td>1.58</td>
</tr>
<tr>
<td>Ours</td>
<td>No</td>
<td>0.0746</td>
<td>0.2398</td>
<td>2.58</td>
</tr>
</tbody>
</table>

supervised methods, except for DefocusNet.

Fig. 6 shows the qualitative results of 4D Light Field Dataset [16]. Our supervised model delivers sharper depth boundaries and less noise in textureless regions. Meanwhile, our unsupervised model also achieves visually comparable results with the supervised DefocusNet [24].

**DefocusNet.** The DefocusNet dataset [24] is a synthetic dataset using physically based rendering (PBR) shaders. Since only a subset of this dataset has been released and there are no AiF images, we can only conduct supervised learning on the provided subset. The quantitative results of our method and DefocusNet [24] are shown in Table 6. Note that the results of DefocusNet are retrained on the provided subset with the released code and setting.

**Mobile Depth.** The Mobile Depth dataset [37] is a real-world dataset of focal stacks captured with mobile phones and cameras. Due to the lack of training data in this dataset, all compared models are trained on FlyingThings3D [25], and the input focal stacks are synthesized by the rendering technique used in Barron et al. [3].

As shown in Fig. 7, the output depth maps of our super-
Table 7: Analysis on generalization ability across different datasets. We train our models and DefocusNet [24] on FlyingThings3D [25] with synthesized focal stacks by the rendering technique used in Barron et al. [3], and test these models on Middlebury Stereo Datasets [34] and the DefocusNet dataset. The results show that both of our supervised and unsupervised (with AiF supervision) models have better generalization ability than DefocusNet.

<table>
<thead>
<tr>
<th>Method</th>
<th>Supervised</th>
<th>Test Dataset</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>absRel</th>
<th>sqrRel</th>
<th>Sec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>Yes</td>
<td>Middlebury</td>
<td>3.8249</td>
<td>58.5698</td>
<td>5.9355</td>
<td>2.4776</td>
<td>45.5886</td>
<td>0.0287</td>
</tr>
<tr>
<td>Ours</td>
<td>No</td>
<td>Middlebury</td>
<td>5.4499</td>
<td>99.6029</td>
<td>8.2606</td>
<td>3.8954</td>
<td>80.7120</td>
<td>0.0283</td>
</tr>
<tr>
<td>Ours</td>
<td>Yes</td>
<td>DefocusNet</td>
<td>0.1827</td>
<td>0.0795</td>
<td>0.2607</td>
<td>72.4664</td>
<td>0.0241</td>
<td>0.0283</td>
</tr>
<tr>
<td>Ours</td>
<td>No</td>
<td>DefocusNet</td>
<td>0.1816</td>
<td>0.0627</td>
<td>0.2380</td>
<td>59.4459</td>
<td>0.0195</td>
<td>0.0195</td>
</tr>
<tr>
<td>DefocusNet [24]</td>
<td>Yes</td>
<td>DefocusNet</td>
<td>0.3200</td>
<td>0.1478</td>
<td>0.3722</td>
<td>138.2917</td>
<td>0.0527</td>
<td>0.0527</td>
</tr>
</tbody>
</table>

**Generalization ability analysis.** Middlebury Stereo Datasets[34] is a real-world dataset of stereo images along with ground truth disparity maps. We use this dataset to analyze the generalization ability of our method. To this end, we firstly let our models and DefocusNet [24] be trained on FlyingThings3D [25] with synthesized focal stacks by the rendering technique used in Barron et al. [3], and then test these models on Middlebury Stereo Datasets as well as the DefocusNet dataset. The quantitative results are shown in Table 7. One can see that both of our supervised and unsupervised (with AiF supervision) models achieve better generalization than DefocusNet on these two datasets.

**Running time.** As indicated in Table 7, our method is faster than the state-of-the-art method DefocusNet [24]. The main reason might be that our model does not have a bottleneck like the global pooling layer in DefocusNet.

5. Conclusion

We have proposed a method to jointly estimate the depth map and the all-in-focus (AiF) image from an input focal stack with a shared network. By the design of the proposed attention mechanism, the shared network can be trained either supervisedly, or unsupervisedly with AiF images. Our method can further mitigate domain gaps even in the absence of ground truth depth. Experimental results show that our method outperforms the state-of-the-art methods while also having higher efficiency in inference time.

Figure 7: Visual comparison on the Mobile Depth dataset. With test-time optimization, our models perform better than DefocusNet [24] and favorably against Mobile Depth [37] qualitatively. (S: supervised. US: unsupervised (AiF supervision). *: test-time optimization.)
References


