Frequency-Aware Deep Depth from Focus

Tao Yan , Yingying Wang , Jiangfeng Zhang , Yuhua Qian* , Jieru Jia , Lu Chen , Feijiang Li
Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China
hongyanyutian@sxu.edu.cn, 202222404025@email.sxu.edu.cn, zjf_8099@163.com,
jinchengqyh@126.com, jierujia@sxu.edu.cn, chenlu@sxu.edu.cn, fjli@sxu.edu.cn

Abstract

In large aperture imaging, the shallow depth of field (DoF) phenomenon requires capturing multiple images at different focal levels, allowing us to infer depth information using depth from focus (DFF) techniques. However, most previous works design convolutional neural networks from a time domain perspective, often leading to blurred fine details in depth estimation. In this work, we propose a frequency-aware deep DFF network (FAD) that couples multi-scale spatial domain local features with frequency domain global structural features. Our main innovations include two key points: First, we introduce a frequency domain feature extraction module that uses the Fourier transform to transfer latent focus features into the frequency domain. This module adaptively captures essential frequency information for focus changes through element-wise multiplication, enhancing fine details in depth results while preserving global structural integrity. Second, the time-frequency joint module of FAD improves the consistency of depth information in sparse texture regions and the continuity in transition areas from both local and global complementary perspectives. Comprehensive experiments demonstrate that our model achieves compelling generalization and state-of-the-art depth prediction across various datasets. Additionally, it can be quickly adapted to real-world applications as a pre-trained model.

1 Introduction

Obtaining three-dimensional (3D) shape information of a scene using visual cues is an important problem in the field of computer vision. Current methods for 3D shape reconstruction are primarily based on optical imaging and can be broadly categorized into active optical reconstruction and passive optical reconstruction. Active optical methods require additional equipment to assist in obtaining the 3D structure of the scene. Typical methods include laser confocal microscopy and structured light 3D measurement. However, the former suffers from low scanning efficiency and a limited measurement range, making it unsuitable for scenes with large depth variations and high real-time constraints. On the other hand, the latter is easily affected by high reflectivity and lighting conditions during the reconstruction process, which can lead to decreased reconstruction accuracy [Yan *et al.*, 2020b]. Passive optical reconstruction primarily recovers the 3D structure of a scene from 2D images. Due to the low cost and convenience of acquiring scene images, many different types of methods have emerged, such as multiview stereo vision [Wei *et al.*, 2023] and monocular depth estimation [Yang *et al.*, 2024] methods. However, multiview stereo vision heavily relies on the accuracy of left-right view matching, monocular depth estimation requires all-in-focus images as a prerequisite for accurate depth estimation. For scenes requiring high-resolution imaging, a large aperture means shallow depth of field (DoF), making it difficult to obtain all-in-focus images in a single shot.

To address the aforementioned challenges, Depth from Focus (DFF) technology [Nayar and Nakagawa, 1994] holds promise for efficiently and cost-effectively obtaining high-precision 3D shape information of a scene. DFF works by adjusting the distance between the camera and the scene at regular intervals, capturing a sequence of images that cover the entire depth range of the scene. Depth information is then obtained by analyzing the focus changes within the multi-focus image sequence. This is particularly important for scenes requiring high-resolution reconstruction. For instance, DFF technology is used to achieve high-precision 3D shape reconstruction of industrial microscopic scenes [Yan *et al.*, 2020a]. Additionally, some Nikon commercial cameras also feature focal length variation shooting capabilities [Mandl *et al.*, 2024], which can subsequently be utilized to achieve high quality DoF synthesis using DFF technology. Particularly with the emergence of deep neural network models, their powerful ability to represent focus features has shown strong performance in addressing DFF problems [Hazirbas *et al.*, 2019]. Nevertheless, due to the spatial domain operation limitations of convolutional neural networks, existing deep learning-based DFF methods still have room for improvement in terms of preserving fine-grained details and robustness of the scene.

In this work, aiming for finer depth details, we introduce the FAD model from the perspective of frequency domain feature extraction. Unlike existing state-of-the-art DFF methods, our model demonstrates well generalization ability and

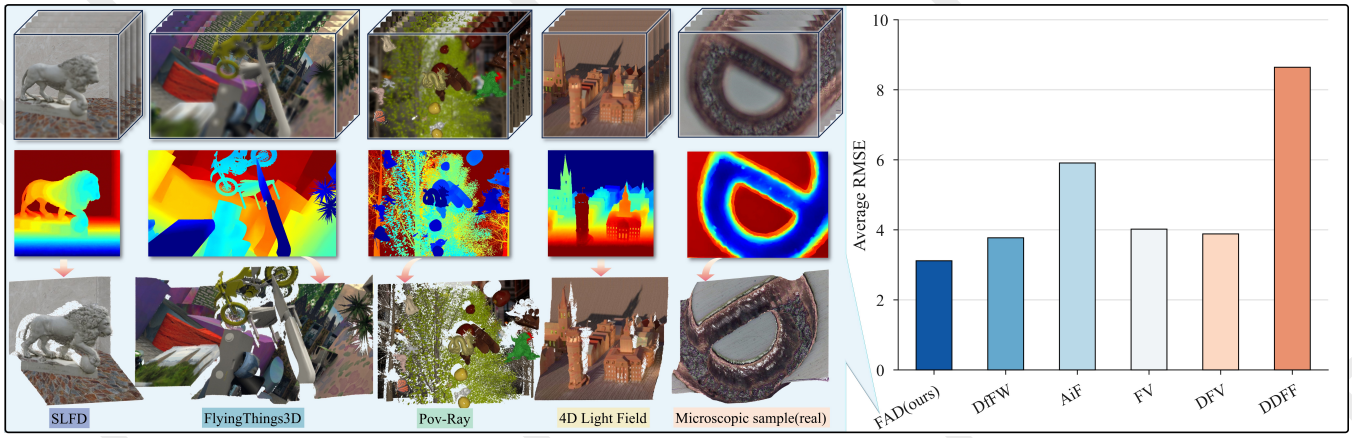

Figure 1: Our proposed FAD model provides state-of-the-art DFF results in various types of synthesis and real scenarios. Detailed data can be found in the experimental section of this paper.

achieves high-precision scene reconstruction directly from multi-focus image sequences. Our contribution is three-fold:

- A frequency domain feature extraction module is introduced, which effectively preserves the global structural information of the scene, thereby enabling the reconstruction results to achieve more accurate fine-grained details.
- The time-frequency joint module in our FAD model not only enhances the robustness of datasets with varying resolutions but also effectively ensures the consistency of depth information in homogeneous regions and the continuity of depth information in heterogeneous regions within the reconstruction results.
- Compared to state-of-the-art DFF methods, the FAD model exhibits outstanding performance across eight different types of existing datasets. Notably, it can serve as a pre-trained model for rapid adaptation and application in real microscopic scenarios. An experimental evidence is provided in Figure 1.

2 Related Work

2.1 Depth from Focus (DFF)

DFF originated in the field of microscopic imaging and is now widely used in the macro imaging process of commercial cameras [Suwajanakorn *et al.*, 2015]. The commonality between these two applications is the need to shorten the distance between the camera and the scene being measured, which results in a shallower DoF. Additionally, shortening the distance means less light enters the camera, necessitating a larger aperture to allow more light in, which further exacerbates the shallowness of the DoF. These settings lead to only part of the scene being in focus in a single image, with the rest becoming blurred. Therefore, DFF first captures a sequence of images covering the entire DoF by varying the distance between the camera and the scene. It then aggregates the indices of the sharpest pixels from this sequence to derive the depth information of the scene.

2.2 State-of-the-art DFF Methods

Early DFF methods focused on model design, particularly the development of image focus measure operators and depth map refinement. Focus measure operators aim to enhance the signal-to-noise ratio between sharp and defocused pixels, improving the accuracy of focus determination. Representative operators include Laplacian-based measures [Yan *et al.*, 2020a], ring difference filters [Jeon *et al.*, 2020] and transformation-based measures [Yan *et al.*, 2020b]. Depth map refinement uses prior knowledge to correct focus measurement errors, with notable methods such as cost aggregation [Jeon *et al.*, 2020] and regularization [Ali and Mahmood, 2021]. However, these methods often struggle in scenarios with sparse textures and noise interference. In addition, these methods cannot accurately infer complex samples that satisfy similar distributions. To overcome the above issues, deep learning-based DFF methods have emerged [Yang *et al.*, 2022; Maximov *et al.*, 2020; Wang *et al.*, 2021; Fujimura *et al.*, 2024]. These methods design end-to-end convolutional neural networks to directly learn depth information from multi-focus image sequences. However, a common characteristic of these networks is that they fit the focus measure features of different datasets through convolutional neural networks. Essentially, this approach belongs to the fitting and learning of temporal focus measure operators in model design-based DFF methods. As a result, these methods tend to overlook the global structural information of the scene, leading to blurred depth details in the reconstruction results.

2.3 Applications of Frequency-domain in Vision

Recently, an increasing number of studies have begun incorporating frequency domain methods into the field of deep learning to optimize the performance of deep neural networks [Yang and Soatto, 2020; Xu *et al.*, 2019]. Some studies transform images into the frequency domain to leverage frequency information and enhance the performance of specific tasks [Lee *et al.*, 2018; Rao *et al.*, 2021], others use the convolution theorem to accelerate the computation of convolutional neural networks (CNNs) via fast Fourier transform (FFT) [Li *et*

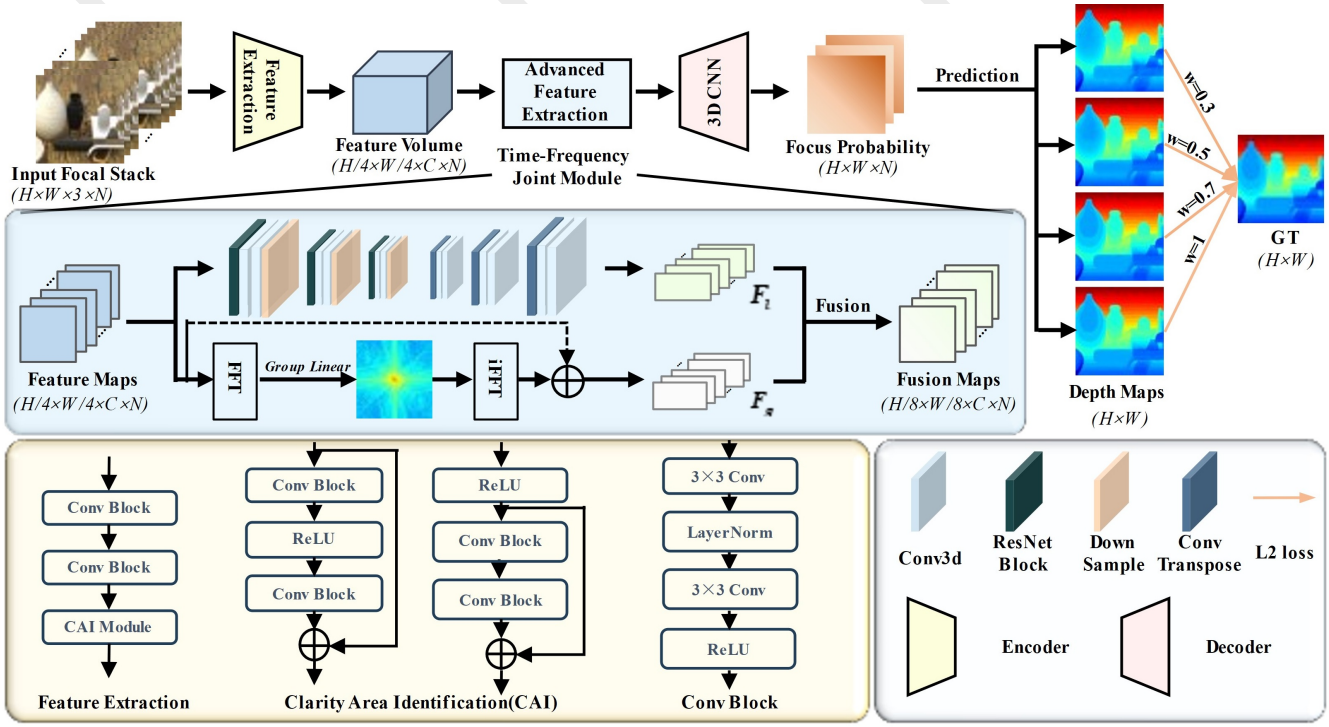

Figure 2: The whole framework of FAD. The time-frequency joint module achieves high-quality depth prediction in different scenarios by complementing multi-scale local features with global structural features.

al., 2020; Ding *et al.*, 2017; Huang *et al.*, 2023]. Inspired by the above works, we believe that introducing a frequency domain processing module in DFF will help enhance the global feature representation of the network, thereby capturing more subtle focal changes.

3 Methodology

Our approach combines spatial and frequency domain information, employing a U-Net architecture to enhance depth estimation accuracy by integrating features from both domains. First, we provide an overview of the proposed FAD method, detailing the overall network process. Next, we describe the extraction of spatial and frequency domain features, including how to effectively integrate these features into the U-Net architecture. Finally, we discuss the implementation details of the network, covering parameter settings, optimization strategies and loss function selection for training to ensure model stability and efficiency.

3.1 Overview

Figure 2 presents an overview of the proposed FAD method, which is based on an encoder-decoder architecture utilizing 3D convolutions instead of 2D convolutions. Unlike 2D convolutions, which can only capture features from a single image frame, 3D convolutions are capable of capturing focus change trends across different frames in a multi-focus image sequence. Feature fusion within the network is achieved through skip connections, which effectively transmit feature information across various levels, thereby enhancing the network's representational capacity. This approach also facilitates gradient propagation and accelerates model convergence.

In the encoder phase, the input multi-focus image sequence undergoes initial feature extraction through a series of 3D convolutional layers. Each layer includes convolution operations, batch normalization and ReLU activation functions to ensure that the extracted features are highly representative and capable of nonlinear expression. Following this, these features are meticulously processed by the clarity area identification (CAI) module, yielding a feature volume that serves as the input for the time-frequency joint module. As the network deepens, the encoder compresses image information into higher-level feature representations while preserving the spatiotemporal relationships within the multi-focus image sequences. Max pooling is employed for downsampling, and to mitigate the loss of depth information, a 3×3 convolution is applied to the output of the downsampling process. In the middle phase of the encoder-decoder, separate spatial and frequency domain feature extraction modules are employed to maintain precise focus information in both dimensions. This approach enables the network to better adapt to varying scales and resolutions, thoroughly learning both global structures and local details, thereby improving the accuracy of depth estimation.

In the decoder phase, features from the encoder are decoded layer by layer using a series of transposed convolutions to gradually restore the spatial resolution of the feature maps. These transposed convolutions, combined with symmetrical 3D convolutional layers from the encoder, progressively re-

turn the feature maps to the dimensions of the input images. Each layer in the decoder incorporates skip connections from the encoder to enhance information fusion during decoding. This method not only leverages the layer-by-layer recovery capability of the decoder but also effectively integrates the multi-level features extracted during the encoder phase, ensuring the final output image has a higher quality depth estimation.

3.2 Clarity Area Identification

Our initial feature extraction phase consists of 3D convolution and the CIA module. The 3D convolution, with its unique capability, allows for feature exchange with adjacent focal slices, enabling it to detect and identify subtle defocus variations that are often hidden in weak areas with little discernible texture. This ensures high precision and sensitivity in detection. In particular, previous methods utilized global average pooling to simplify the multi-dimensional information of focal slices into a single value, which inevitably led to a significant loss of deep representation information. To overcome this obstacle, we innovatively adopted the CIA module, which uses 3D convolution and ReLU to compute a focal attention score, allowing for the capture of more refined deep representation information.

3.3 Spatial Domain Feature Extraction

The multi-scale feature extraction method can effectively capture details and structural information of different scales in an image. This characteristic is particularly important for reconstructing sparse textures and low-contrast areas in DFF tasks. Therefore, our FAD method utilizes multi-scale feature extraction to enable the model to extract spatial domain features from the scene.

Specifically, given a feature map X, it is first subjected to multi-scale average pooling operations to obtain feature representations X_i at different scales. Let the pooling scales be s_1, s_2, s_3. The multi-scale pooling process can then be represented as:

$$X_i = \mathrm{AvgPool}_{s_i}(X), i \in \{1, 2, 3\}\,. \tag{1}$$

Next, convolution operations are applied to the pooled feature maps X_i to further extract spatial domain features at multiple scales. Let the convolution kernel sizes at different scales be k_1, k_2, k_3. The convolution process can then be represented as:

$$F_i = \mathrm{Conv}_{k_i}(X_i), i \in \{1, 2, 3\}\,. \tag{2}$$

The convolution results at different scales are concatenated and fused to obtain the final multi-scale feature representation F_l.

This multi-scale feature extraction method first utilizes average pooling operations to reduce computational load and redundant information while preserving important spatial information. Subsequent convolution operations can further extract and refine features, enhancing the model's ability to perceive spatial structures at different scales. Finally, by fusing feature representations from multiple scales, the performance of the model in handling complex DFF scenes and tasks at different resolutions is improved.

3.4 Frequency Domain Feature Extraction

Although the spatial domain feature extraction module can capture some global information, the local nature of convolution operations limits its ability to effectively represent global structural information. In contrast, the Fourier transform can process global information in the frequency domain by performing a weighted sum of all image pixels. Therefore, we introduce a frequency domain feature extraction module in the FAD model to more accurately capture subtle focus changes that are difficult to detect in the spatial domain, thereby effectively enhancing the model's sensitivity and accuracy in the depth information extraction process.

Given a feature X as the output of the encoder, we first transform the feature map X into the frequency domain by performing a Fast Fourier Transform (FFT), obtaining its frequency representation X_f. Next, to adaptively select the required frequency information, we learn adaptive weights for X_f to weight the spectrum, obtaining the weighted representation X_ω. This process can be expressed as:

$$X_\omega = M \odot \mathrm{FFT}(X), \tag{3}$$

where M is a learnable adaptive weight matrix, and $\odot$ denotes element-wise matrix multiplication. To implement M, we use a 1×1 convolution followed by a ReLU activation function and another linear layer. With respect to the output of the linear layer, the result of this soft thresholding is multiplied with the output of the Fourier transform, yielding the final data representation X_s. This step is designed to simultaneously enhance the sparsity of the output and improve the interpretability of the data, thereby enhancing the accuracy and robustness of deep information extraction.

$$X_s = \mathrm{SoftShink}(X_\omega) \odot \mathrm{FFT}(X). \tag{4}$$

Finally, we use the inverse Fast Fourier Transform (iFFT) to convert the processed features from the frequency domain back to the spatial domain, obtaining the spatial representation $X_{f^{-1}}$:

$$X_{f^{-1}} = \mathrm{iFFT}(X_s). \tag{5}$$

To prevent information loss, we combine the features before and after the Fourier transform, X and $X_{f^{-1}}$, resulting in a comprehensive feature F_g that ensures the integrity of information during the transformation from the spatial domain to the frequency domain.

$$F_g = X + X_{f^{-1}}. \tag{6}$$

By employing a parallel extraction method for both spatial domain features F_l and frequency domain features F_g, the final feature map F is obtained.

$$F = F_l + F_g. \tag{7}$$

The FAD model utilizes frequency domain information to compensate for multi-scale spatial domain features, effectively enhancing the model's performance in fine depth structure estimation.

3.5 Prediction

The feature map F, extracted by the time-frequency joint module, undergoes refinement through 3D CNN to generate four feature maps F_i at different scales. Subsequently, these feature maps undergo an upsampling process sequentially, and a Softplus function is applied for nonlinear normalization, resulting in refined depth attention maps. Immediately following this, the depth attention maps are element-wise multiplied with equally spaced focal positions P, precisely integrating information from each focal position to obtain the final scene depth map D_i.

$$D_i = (log(1 + exp(F_i))) \odot P, i \in \{1, 2, 3, 4\}. \quad (8)$$

3.6 Implementation Details

We build our network model using the PyTorch framework and train and test it on a single NVIDIA GeForce RTX 4090. During training, we employ the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.99$) with an initial learning rate of 10^{-3} and a batch size of 4. Additionally, we apply data augmentation techniques such as image flipping, cropping, rotation, and gamma correction. Images are randomly cropped into 256×256 patches and fed into the network.

For our depth estimation method, we optimize the entire model by comparing predicted pixel depths to ground truth depths with a multi-scale weighted loss function. The specific loss function is defined as follows:

$$L_{depth} = \sum_{i=1}^{4} \omega_i \, || \, D_i - D_{gt} \, ||_2, \quad (9)$$

where $||\bullet||_2$ denotes the $L2$ loss, and D_{gt} represents the ground truth depth map. $i \in \{1, 2, 3, 4\}$ indicates the predicted depth maps at different pyramid-like scales. In this model, ω_i is set to 0.3, 0.5, 0.7 and 1, respectively.

4 Experiments

In this section, we detail the evaluation metrics and datasets used in our experiments and compare the proposed FAD method with state-of-the-art DFF methods. Furthermore, we conduct ablation experiments to assess the effectiveness of each component of the proposed network. Finally, we evaluate the model's generalization ability across various synthetic and real microscopic datasets.

4.1 Metrics

For quantitative evaluation, we use the following metrics to evaluate the quantitative results: mean-squared error (MSE), root-mean-squared error (RMSE), log root-mean-squared error (log RMSE), relative-absolute error (AbsRel), relative-squared error (SqRel), accuracy with $\delta_i = 1.25^i, i = 1, 2, 3$ and inference time (Secs.).

4.2 Comparison of Algorithms and Datasets

The comparison includes seven deep learning-based DFF methods (DDFS [Fujimura *et al.*, 2024], DfFW [Won and Jeon, 2022], FV [Yang *et al.*, 2022], DFV [Yang *et al.*, 2022], AiF [Wang *et al.*, 2021], DefocusNet [Maximov *et al.*, 2020], DDFF [Hazirbas *et al.*, 2019]) and two model design-based DFF methods (RDF [Jeon *et al.*, 2020] and RR [Ali and Mahmood, 2021]). The datasets consist of five synthetic datasets (SLFD [Shi *et al.*, 2019], Pov-Ray [Heber and Pock, 2016], DefocusNet[Maximov *et al.*, 2020], 4D Light Field [Honauer *et al.*, 2017], FlyingThings3D [Mayer *et al.*, 2016]) and four real datasets (NYU Depth V2 [Carvalho *et al.*, 2018], DDFF 12-Scene [Hazirbas *et al.*, 2019], Middlebury [Scharstein *et al.*, 2014], Microscopic), among which the microscopic dataset is unlabeled.

4.3 Main Results

The best results among all experimental results are highlighted in bold, and the second-best results are indicated with underline.

Results on DefocusNet Dataset. Table 1 presents the quantitative comparison results for DefocusNet dataset. It can be seen that, except for the AbsRel metric, which is slightly inferior to the DfFW method, our FAD method outperforms other methods in all other metrics. As shown in Figure 3, our FAD method significantly outperforms the DDFS and DfFW methods in inferring the elliptical cavities and spiral details of the objects.

Table 1: Quantitative evaluation on DefocusNet dataset.

Method(Pub.)	AbsRel ↓	MSE ↓	RMSE ↓
DefocusNet(CVPR2020)	0.1386	0.0127	0.1043
AiF(ICCV2021)	0.1115	0.1542	0.3642
FV(CVPR2022)	0.1356	0.0133	0.1076
DFV(CVPR2022)	0.1255	0.0140	0.1086
DfFW(ECCV2022)	**0.0809**	<u>0.0087</u>	<u>0.0859</u>
DDFS(IJCV2024)	0.1963	0.0369	0.1749
FAD(ours)	<u>0.0929</u>	**0.0084**	**0.0847**

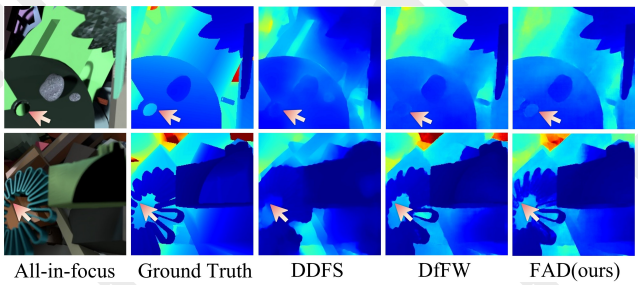

Figure 3: Visual comparison on DefocusNet dataset.

Results on SLFD Dataset. Table 2 presents the quantitative comparison results on the SLFD dataset. Our method has significant improvements in all metrics, although it is slightly slower than DfFW. As shown in Figure 4, our method excels at maintaining the integrity of object edges and fine structures within the scene, including details such as the spokes of a bicycle and the switch of a lamp.

Results on 4D Light Field Dataset. As shown in Table 3, our FAD method significantly outperforms advanced deep learning-based DFF methods across three evaluation metrics.

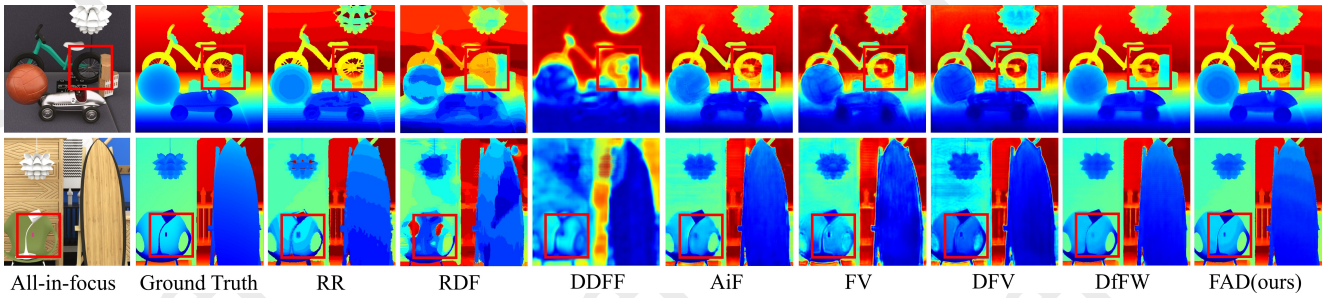

Figure 4: Visual comparison on SLFD dataset.

Table 2: Quantitative evaluation on SLFD dataset.

Method(Pub.)	MSE ↓	log RMSE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	δ=1.25 ↑	δ=1.25^2 ↑	δ=1.25^3 ↑	Secs.(S)
DDFF(ACCV2018)	14.0337	0.4349	3.6381	0.7372	4.0416	58.15	70.18	77.28	4.0683
AiF(ICCV2021)	6.1070	0.3466	1.9236	0.3365	2.3541	77.15	86.55	92.36	0.1354
FV(CVPR2022)	4.5492	0.3752	2.0222	0.3703	1.3285	69.98	82.46	88.81	0.1336
DFV(CVPR2022)	2.5957	0.2259	1.5696	0.1559	0.5152	82.51	94.90	97.76	0.1344
DfFW(ECCV2022)	0.8939	0.1548	0.8846	0.0773	0.2085	92.72	97.38	99.14	**0.1116**
FAD(ours)	**0.3460**	**0.1059**	**0.5547**	**0.0471**	**0.0775**	**95.79**	**98.63**	**99.58**	0.1259

Due to the introduction of frequency domain features, it can be seen from Figure 5 that our FAD significantly outperforms other methods in terms of the continuity of the light string and the details of the holes in the boxes.

Table 3: Quantitative evaluation on 4D Light Field dataset.

Method(Pub.)	AbsRel ↓	MSE ↓	RMSE ↓
DefocusNet(CVPR2020)	-	0.0593	0.2355
DDFF(ACCV2018)	0.3296	0.1146	0.3310
AiF(ICCV2021)	0.1685	0.0472	0.2014
FV(CVPR2022)	0.1900	0.0301	0.1537
DFV(CVPR2022)	0.1915	0.0317	0.1549
DfFW(ECCV2022)	0.1670	0.0230	0.1288
FAD(ours)	**0.1527**	**0.0218**	**0.1248**

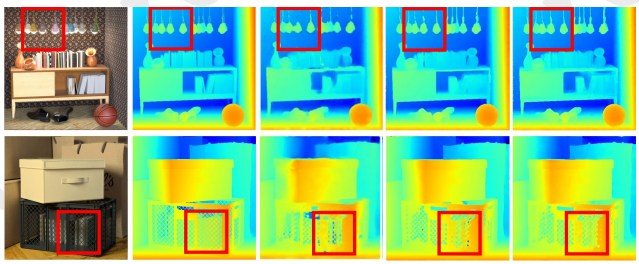

Figure 5: Visual comparison on 4D Light Field dataset.

Results on DDFF 12-Scene Dataset. Table 4 presents the quantitative comparison results for DDFF 12-Scene. Compared to other methods, our FAD method achieves the best results across all metrics. As shown in Figure 6, our FAD method significantly outperforms the DFV and DfFW methods in preserving edge details for statues and chairs.

Table 4: Quantitative evaluation on DDFF 12-Scene.

Method(Pub.)	AbsRel ↓	MSE ↓	RMSE ↓
DDFF(ACCV2018)	0.2362	$4.53e^{-4}$	0.2362
FV(CVPR2022)	0.0971	$1.85e^{-4}$	0.0119
DFV(CVPR2022)	0.1001	$2.05e^{-4}$	0.0124
DfFW(ECCV2022)	0.1549	$2.23e^{-4}$	0.0135
FAD(ours)	**0.0505**	**0.50**e^{-4}	**0.0062**

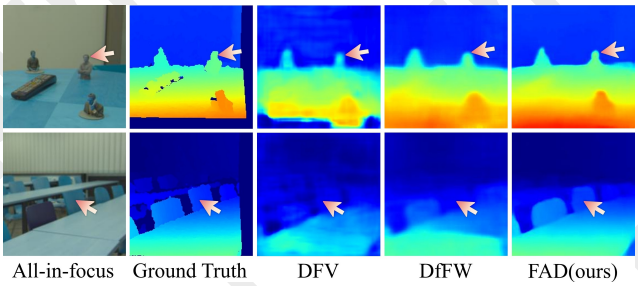

Figure 6: Visual comparison on the DDFF 12-Scene.

4.4 Ablation Study

To further validate the rationale behind our FAD method and the effectiveness of its components, we conduct ablation experiments on the 4D Light Field dataset and analyze the results based on model performance. In Ablation Experiment 1 (U), we train the model using only the U-shaped backbone network structure. In Ablation Experiment 2 (U + S), we incorporate the spatial domain module to extract finer features. In Ablation Experiment 3 (U + F), we utilize the frequency domain module to capture frequency information and extract global features. In Ablation Experiment 4 (U + S + F), we combine both the spatial and frequency domain modules. As

shown in Figure 7, the experimental results indicate that simultaneously utilizing spatial and frequency domain information can complement each other and greatly enhance the model's performance.

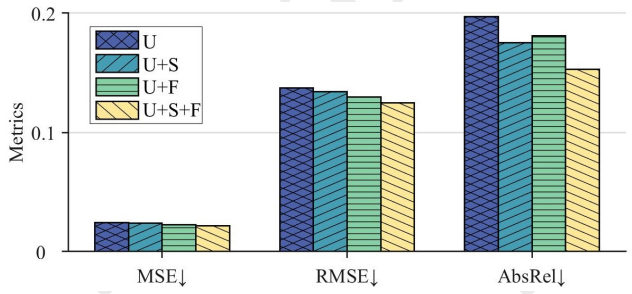

Figure 7: Ablation study of FAD on the 4D Light Field dataset.

4.5 Generalization

Generalization among Different Datasets. To validate the generality of the proposed FAD method, we train it on the FlyingThings3D dataset and test it on three additional datasets: Middlebury, SLFD and Pov-Ray. Comparisons with three state-of-the-art DFF methods reveal that both the quantitative and visual results, presented in Table 5 and Figure 8, demonstrate that our FAD method achieves impressive performance across all tested datasets. Especially in terms of the edges of pipes in the Middlebury sample, the contours of green plants in the SLFD sample and the shapes of small leaves in the Pov-Ray sample, our FAD method shows significant advantages.

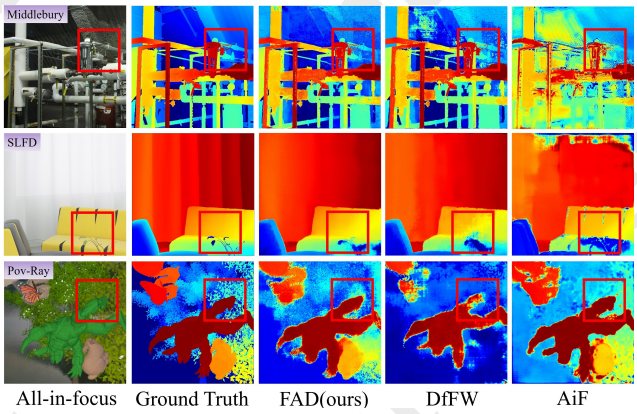

Figure 8: The generalization results of the FAD and the state-of-the-art DFF method across different datasets.

Generalization to Real Microscopic Scenes. To validate the FAD method's real-world applicability, we first trained the FAD on the synthetic Pov-Ray dataset. Subsequently, we applied 3D TFT [Yan. *et al.*, 2023] pseudo-labeling to 100 unlabeled microscopic samples. A pre-trained bias correction model [Zaken *et al.*, 2022] was then employed to optimize the network's bias parameters using this pseudo-labeled data, and then directly perform inference on the microscopic data. As shown in Figure 9, due to the data gap between synthetic and real data, the inference results are not ideal. However, after incorporating a small amount of pseudo-labels, the fine-tuned FAD is able to infer more refined depth details. Especially in Sample 1, the given label cannot capture the subtle scratches, while the FAD method perfectly restores the depth information of those scratches.

Table 5: Generalization on FlyingThings3D.

Method	Train	Test	MSE ↓	RMSE ↓	SqRel ↓
DefocusNet	Flying Things3D	Middlebury	157.440	9.079	4.245
AiF			58.570	5.936	3.039
DfFW			9.178	2.930	0.376
FAD(ours)			**8.970**	**2.930**	**0.345**
DefocusNet	Flying Things3D	SLFD	-	-	-
AiF			15.311	3.548	1.114
DfFW			9.830	2.845	0.825
FAD(ours)			**8.975**	**2.723**	**0.706**
DefocusNet	Flying Things3D	Pov-Ray	-	-	-
AiF			79.825	8.725	28.015
DfFW			47.704	6.898	**11.667**
FAD(ours)			**38.087**	**6.128**	11.879

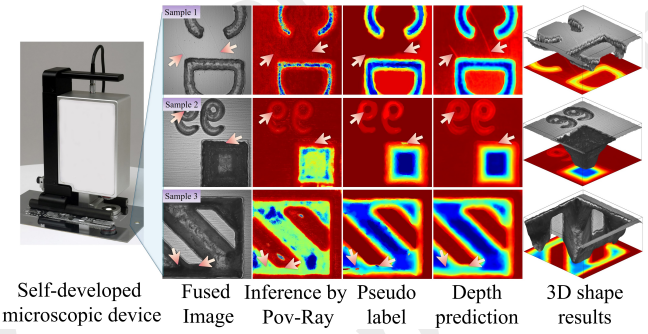

Figure 9: Using our self-developed microscopic data acquisition device, the application of pseudo-label fine-tuning effectively improves the depth prediction results of FAD model on microscopic data.

5 Conclusion

In this work, we develop a novel frequency-aware deep DFF architecture, referred to as FAD. This network introduces a new frequency domain feature extraction module and achieves a more robust DFF task through a time-frequency feature joint architecture. Experimental results indicate that the proposed FAD network not only achieves optimal performance on public DFF datasets but also generalizes quickly to real-world scenarios through fine-tuning, particularly demonstrating significant advantages in preserving fine details in depth prediction results. In the future, an interesting direction for research is how to achieve mixed training on different types of DFF datasets through multi-task learning, thereby improving the model's generalization capabilities.

Acknowledgments

This work was supported by the Major Program of National Natural Science Foundation of China (T2495250), the Key Program of the National Natural Science Foundation of China (62136005), the National Natural Science Foundation of China (62472268, 62373233, 62476160, 62441239), the Funds for central government-guided local science and technology development (YDZJSX20231C001) and the Fundamental Research Program of Shanxi Province (202403021211226).

References

[Ali and Mahmood, 2021] Usman Ali and Muhammad Tariq Mahmood. Robust focus volume regularization in shape from focus. *IEEE Transactions on Image Processing*, 30:7215–7227, 2021.

[Carvalho *et al.*, 2018] Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. Deep depth from defocus: How can defocus blur improve 3d estimation using dense neural networks? In *European Conference on Computer Vision(ECCV)*, pages 307–323, 2018.

[Ding *et al.*, 2017] Caiwen Ding, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang, Xuehai Qian, Yu Bai, Geng Yuan, et al. Circnn: accelerating and compressing deep neural networks using block-circulant weight matrices. In *IEEE/ACM International Symposium on Microarchitecture(MICRO)*, pages 395–408, 2017.

[Fujimura *et al.*, 2024] Yuki Fujimura, Masaaki Iiyama, Takuya Funatomi, and Yasuhiro. Mukaigawa. Deep depth from focal stack with defocus model for camera-setting invariance. *International Journal of Computer Vision*, 132(6):1970–1985, 2024.

[Hazirbas *et al.*, 2019] Caner Hazirbas, Sebastian Georg Soyer, Maximilian Christian Staab, Laura Leal-Taixé, and Daniel Cremers. Deep depth from focus. In *Asian Conference on Computer Vision(ACCV)*, pages 525–541, 2019.

[Heber and Pock, 2016] Stefan Heber and Thomas Pock. Convolutional networks for shape from light field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3746–3754, 2016.

[Honauer *et al.*, 2017] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision(ACCV)*, pages 19–34, 2017.

[Huang *et al.*, 2023] Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Zheng-Jun Zha, Yan Lu, and Baining Guo. Adaptive frequency filters as efficient global token mixers. In *IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 6049–6059, 2023.

[Jeon *et al.*, 2020] Hae-Gon Jeon, Jaeheung Surh, Sunghoon Im, and In So Kweon. Ring difference filter for fast and noise robust depth from focus. *IEEE Transactions on Image Processing*, 29:1045–1060, 2020.

[Lee *et al.*, 2018] Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. Single-image depth estimation based on fourier domain analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 330–339, 2018.

[Li *et al.*, 2020] Shaohua Li, Kaiping Xue, Bin Zhu, Chenkai Ding, Xindi Gao, David Wei, and Tao Wan. Falcon: A fourier transform based approach for fast and secure convolutional neural network predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 8705–8714, 2020.

[Mandl *et al.*, 2024] David Mandl, Shohei Mori, Peter Mohr, Yifan Peng, Tobias Langlotz, Dieter Schmalstieg, and Denis Kalkofen. Neural bokeh: Learning lens blur for computational videography and out-of-focus mixed reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2024.

[Maximov *et al.*, 2020] Maxim Maximov, Kevin Galim, and Laura Leal-Taixé. Focus on defocus: Bridging the synthetic to real domain gap for depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 1068–1077, 2020.

[Mayer *et al.*, 2016] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 4040–4048, 2016.

[Nayar and Nakagawa, 1994] Shree K Nayar and Yasuo Nakagawa. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, 1994.

[Rao *et al.*, 2021] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021.

[Scharstein *et al.*, 2014] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR)*, pages 31–42, 2014.

[Shi *et al.*, 2019] Jinglei Shi, Xiaoran Jiang, and Christine Guillemot. A framework for learning depth from a flexible subset of dense and sparse light field views. *IEEE Transactions on Image Processing*, 28(12):5867–5880, 2019.

[Suwajanakorn *et al.*, 2015] Supasorn Suwajanakorn, Carlos Hernandez, and Steven M. Seitz. Depth from focus with your mobile phone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3497–3506, 2015.

[Wang *et al.*, 2021] Ning-Hsu Wang, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen, and Kevin Jou. Bridging unsupervised and supervised depth

from focus via all-in-focus supervision. In *IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 12621–12631, 2021.

[Wei *et al.*, 2023] Yi Wei, Shaohui Liu, Jie Zhou, and Jiwen Lu. Depth-guided optimization of neural radiance fields for indoor multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10835–10849, 2023.

[Won and Jeon, 2022] Changyeon Won and Hae-Gon Jeon. Learning depth from focus in the wild. In *European Conference on Computer Vision(ECCV)*, pages 1–18, 2022.

[Xu *et al.*, 2019] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing(ICONIP)*, pages 264–274, 2019.

[Yan *et al.*, 2020a] Tao Yan, Zhiguo Hu, Yuhua Qian, Zhiwei Qiao, and Linyuan Zhang. 3d shape reconstruction from multifocus image fusion using a multidirectional modified laplacian operator. *Pattern Recognition*, 98:107065, 2020.

[Yan *et al.*, 2020b] Tao Yan, Peng Wu, Yuhua Qian, Zhiguo Hu, and Fengxian Liu. Multiscale fusion and aggregation pcnn for 3d shape recovery. *Information Sciences*, 536:277–297, 2020.

[Yan. *et al.*, 2023] Tao Yan., Yuhua Qian., and Feijiang Li. Intelligent microscopic 3d shape reconstruction method based on 3d time-frequency transformation. *SCIENTIA SINICA Informationis*, 53(2):282–308, 2023.

[Yang and Soatto, 2020] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 4085–4095, 2020.

[Yang *et al.*, 2022] Fengting Yang, Xiaolei Huang, and Zihan Zhou. Deep depth from focus with differential focus volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 12642–12651, 2022.

[Yang *et al.*, 2024] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024.

[Zaken *et al.*, 2022] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–9, 2022.