Supplementary Material for: Deep Depth from Focus with Differential Focus Volume

Fengting Yang Xiaolei Huang Zihan Zhou
The Pennsylvania State University
{fuy34, suh972, zuz22}@psu.edu

1. Network Architecture

Figure 1 presents our architecture design. The overall framework is similar to [9], but we optimize the 2D encoder and the 3D CNN decoders for the DFF task. Given B focal stacks with N frames in each, we first reshape them into a $B \cdot N \times 3 \times H \times W$ tensor and pass it to the 2D CNN to extract features in four different scales. Four differential focus volumes (DFVs) are then built based on the features, which later are used to produce the focus probability distributions volume in the corresponding scale as outputs. In the end, all these outputs are upsampled with linear interpolation to the full resolution ($B \times N \times H \times W$) followed by depth probability regression for deep supervision at training time. At test time, only the largest (Level_1 in Figure 1) scale output is upsampled for the depth regression. Here, B, N, H and W denote the batch dimension, the frame dimension, the height dimension, and the width dimension, respectively.

For all the convolution layers in the figure, the three numbers in the in-box parenthesis indicate the in-feature channel number, the out-feature channel number, and the convolution stride, respectively. The parenthesis below the box presents the output size. All convolution layers use kernel size 3, except "Conv2d_1" , "Conv2d_proj" and the last convolution in "upConv3d_Blk" and "Conv3d_proj_Blk". The former one uses kernel size 7, and the latter three use size 1. We use batch normalization followed with ReLU for all convolution operations, except the final convolution in "Conv3d_proj_Blk" where softmax is applied to the N dimension. No activation function is applied to "Conv2d_proj" or the last convolution in "upConv3d_Blk". For 3D spatial pyramid pooling (SPP) module, we use four pooling scales m_a linearly sampled from 1 to $\left\lfloor \frac{min(N,H,W)}{2} \right\rfloor$, where $a = 1, ..., 4$, and the corresponding pooling kernel size $k_a = \left\lfloor \frac{N_a}{m_a}, \frac{H_a}{m_a}, \frac{W_a}{m_a} \right\rfloor$. For example, given an input in the shape $(10, 14, 14)$, the 4 scales will be $\{1, 2, 3, 5\}$, and the corresponding pooling kernel sizes are $\{(10, 14, 14), (5, 7, 7), (3, 4, 4), (2, 2, 2)\}$. We do not include 3D SPP in the last two levels for the speed and accuracy trade-off. The 2D SPP kernel sizes are the same as the 3D SPP, except that the N dimension is excluded.

2. Focus Probability Visualization

As Figure 1 illustrates, the direct output of our network in the scale level s is the focus probability distribution P^s (batch $\times$ 1 $\times$ frame_ID $\times$ height $\times$ width), where $p^s_{(b,0,i,u,v)}$ indicates the probability of the pixel at coordinate (u, v) in the frame i sample b to be the best-focused pixel. From this perspective, the whole network can be viewed as a deep focus measure.

To empirically prove this, we visualize the focus probability distribution of our full method (Ours-DFV) in Figure 2. The first three rows are from FoD500 dataset [3], the next three rows are from DDFF-12 dataset [1], and the last three rows are from Mobile depth dataset [8]. The corresponding depth prediction results are available in Figures 3, 4, and 5, respectively. From Figure 2, we can observe, in all the samples, the peak of the best-focused pixel distribution move from the closest objects toward the farthest objects as frame ID increases. This aligns with our input frame order (ascending focal distances).

3. Additional Qualitative Results

Figure 3 and Figure 4 show additional qualitative results on FoD500 and DDFF-12 datasets. Compared to DDFF [1] and DefocusNet [3], our methods, especially Ours-DFV, better preserve object boundaries and provide more smooth depth estimations. Some examples are highlighted with the red boxes. For uncertainty maps, the network turns to be more confident in the closer objects and less confident in objects that are farther or have weak textures. The high uncertainty is also frequently observed in object boundaries.

Figure 5 illustrates the rest of the results on the aligned scenes of the Mobile depth dataset. Differing from the other samples, the focal stacks in rows 6-8 are taken from the same scenes with different camera motions (zero, small, and large), therefore have slightly different frame alignment. We refer readers to [8] for more details of this dataset.

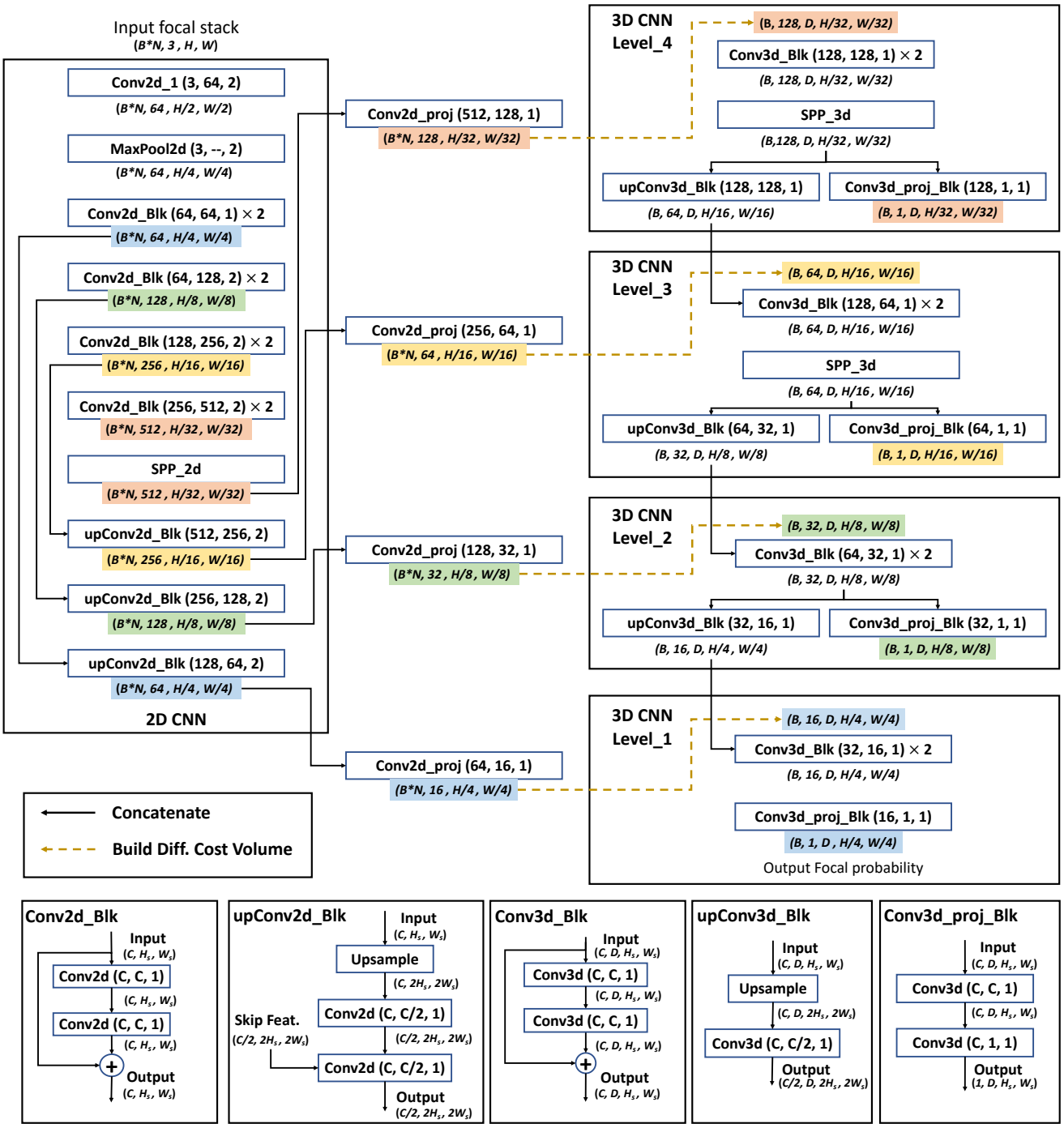

Figure 1. Our differential focus volume network architecture. The B, N, H, and W are the batch size, the input frame number the height, and the width, respectively. The subscript s indicates the scale level.

From the results, we can see that all deep methods generalize well in these scenes without any fine-tuning. In rows 6-8, even though some bumpy predictions can be observed in the books regions, all deep methods present reasonable and consistent depth estimation regardless of the alignment difference, respectively, which shows a certain degree of robustness to the alignment. Compared to other deep methods, Ours-DFV consistently provides more smooth estimations with better boundary preservation, such as the front objects in the first five rows.

Table 1. Ours-DFV results on DDFF-12 validation set with various number of input frames.

Method	#Frm	MSE ↓	RMS↓	log RMS ↓	Abs. rel.↓	Sqr. rel.↓	δ ↑	δ^2 ↑	δ^3 ↑	Bump.↓	avgUnc.↓	Time(ms)↓
Ours-DFV	2	$9.68e^{-4}$	$2.89e^{-2}$	$2.87e^{-1}$	$2.47e^{-1}$	$10.26e^{-3}$	60.79	88.53	96.41	$4.35e^{-1}$	$10.73e^{-2}$	**19.9**
	3	$6.66e^{-4}$	$2.35e^{-2}$	$2.36e^{-1}$	$2.00e^{-1}$	$7.22e^{-3}$	71.13	93.28	97.91	$4.26e^{-1}$	$7.80e^{-2}$	22.8
	4	$6.45e^{-4}$	$2.29e^{-2}$	$2.34e^{-1}$	$1.90e^{-1}$	$6.92e^{-3}$	72.66	93.46	97.84	**4.10**e^{-1}	$6.05e^{-2}$	27.8
	5	$6.63e^{-4}$	$2.35e^{-2}$	$2.39e^{-1}$	$1.86e^{-1}$	$6.92e^{-3}$	70.17	92.94	97.96	$4.21e^{-1}$	$5.39e^{-2}$	33.3
	6	$6.01e^{-4}$	$2.20e^{-2}$	$2.25e^{-1}$	$1.73e^{-1}$	$6.50e^{-3}$	75.65	93.53	97.80	$4.16e^{-1}$	$4.58e^{-2}$	38.6
	7	$6.19e^{-4}$	$2.22e^{-2}$	$2.27e^{-1}$	$1.77e^{-1}$	$6.43e^{-3}$	74.77	93.45	97.80	$4.19e^{-1}$	$3.90e^{-2}$	44.5
	8	$5.92e^{-4}$	**2.16**e^{-2}	**2.14**e^{-1}	**1.68**e^{-1}	$6.00e^{-3}$	**77.92**	94.67	98.03	$4.22e^{-1}$	$3.96e^{-2}$	48.4
	9	**5.90**e^{-4}	$2.18e^{-2}$	$2.23e^{-1}$	$1.86e^{-1}$	$6.48e^{-3}$	74.02	94.58	**98.20**	$4.20e^{-1}$	$3.18e^{-2}$	55.1
	10	$6.45e^{-4}$	$2.24e^{-2}$	$2.18e^{-1}$	$1.68e^{-1}$	**5.57**e^{-3}	75.52	**94.77**	98.08	$4.17e^{-1}$	**2.05**e^{-2}	59.6
DDFF [1]	5	$11.84e^{-4}$	$3.05e^{-2}$	$2.85e^{-1}$	$2.19e^{-1}$	$8.36e^{-3}$	56.00	87.60	97.11	$4.37e^{-1}$	–	191.7
DefocusNet [3]	5	$8.57e^{-4}$	$2.56e^{-2}$	$2.48e^{-1}$	$1.80e^{-1}$	$6.94e^{-3}$	73.16	92.04	96.86	$4.45e^{-1}$	–	34.4

4. Effect of Focal Stack Size

Most traditional DFF methods [2, 4–7] focus on finding the best-focused pixels in the given focal stack and are restricted to the frame-level accuracy for focus analysis. The input focal stack size can greatly affect their depth estimation accuracy. To maintain good accuracy, those methods usually take 10 - 30 frames per stack as input. In contrast, our methods estimate the best-focus distribution and can achieve sub-frame accuracy. This characteristic helps our model to deliver accurate depth estimation with fewer frames.

To study the effect of focal stack size on our model, we train Ours-DFV model on DDFF-12 training set and test it on its validation set with different stack sizes, $N = 2, ..., 10$. We also retrain DDFF [1] and DefocusNet [3] from scratch on the same training set as references. The reason that we exclude FoD500 dataset in this experiment is because it only has 5-frame focal stacks.

The evaluation metrics are the same as the experiment metrics in the main text, which are adopted from [1]. They are MSE, RMS, log RMS, absolute relative (Abs. rel.), squared relative (Sqr. rel.), three accuracy percentages (δ, δ^2, δ^3), bumpiness (Bump.), and average uncertainty (avgUnc.). The first 8 metrics reflect the estimation accuracy from absolute and relative perspectives, the Bump. metric evaluates the smoothness of results, and the avgUnc. is proposed by us to compare the prediction confidence between Ours-CV and Ours-DCV. $avgUnc. = \frac{1}{M}\sum_{j=1}^{M}\phi_j$, where ϕ_j is the uncertainty of the depth estimation of pixel x_j. The lower the value, the higher the confidence.

Table 1 shows the evaluation result. Ours-DFV is able to provide fairly accurate results with only 3-frame input stacks. The MSE error of 3-frame inputs is only 12.9% higher than the best case (9-frame inputs), and already outperforms DDFF and DefocusNet which take 5 frames as input. In general, the model delivers more accurate estimations as the input stack size increases. It is evident by that the best performances in terms of the first eight accuracy metrics are all achieved by the models with $N = \{8, 9, 10\}$ frames. We believe the fluctuation is due to the random process at the training time. As the frame number increases, the model also turns to be more confident in terms of the average uncertainty, in the cost of the runtime. Moreover, for all the cases, the model provides smooth depth predictions in terms of bumpiness.

However, we do notice that the impact of additional frames is diminishing as the frame number increases. DDFF [1] also reports that their network performance on DDFF-12 dataset stops improving after the frame number reaches 10, which is the stack size they finally released to the public. Further studies with a new dataset containing more frames per stack are required to find the actual reason, and we leave it for future work.

References

[1] Caner Hazirbas, Sebastian Georg Soyer, Maximilian Christian Staab, Laura Leal-Taixé, and Daniel Cremers. Deep depth from focus. In *ACCV*, pages 525–541, 2018. 1, 3

[2] Hae-Gon Jeon, Jaeheung Surh, Sunghoon Im, and In So Kweon. Ring difference filter for fast and noise robust depth from focus. *TIP*, 29:1045–1060, 2019. 3

[3] Maxim Maximov, Kevin Galim, and Laura Leal-Taixé. Focus on defocus: bridging the synthetic to real domain gap for depth estimation. In *CVPR*, pages 1071–1080, 2020. 1, 3

[4] Michael Moeller, Martin Benning, Carola Schönlieb, and Daniel Cremers. Variational depth from focus reconstruction. *TIP*, 24(12):5369–5378, 2015. 3

[5] Shree K Nayar and Yasuo Nakagawa. Shape from focus. *PAMI*, 16(8):824–831, 1994. 3

[6] Said Pertuz, Domenec Puig, and Miguel Angel Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013. 3

[7] Jaeheung Surh, Hae-Gon Jeon, Yunwon Park, Sunghoon Im, Hyowon Ha, and In So Kweon. Noise robust depth from focus using a ring difference filter. In *CVPR*, pages 6328–6337, 2017. 3

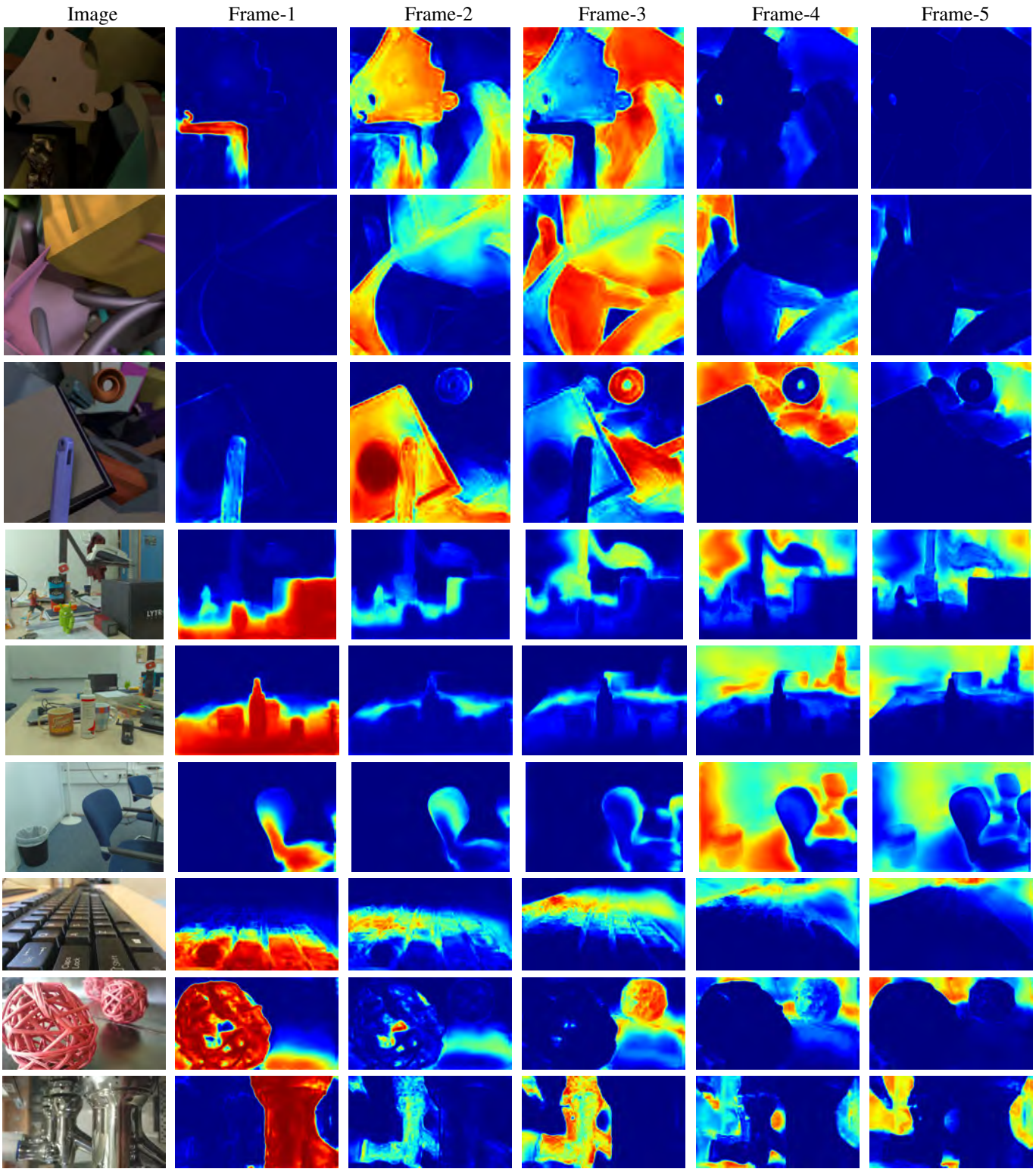

Figure 2. Focus probability visualization. Rows 1-3, 4-6, and 7-9 are from FoD500, DDFF-12, and Mobile depth dataset, respectively. The redder the color, the higher the probability.

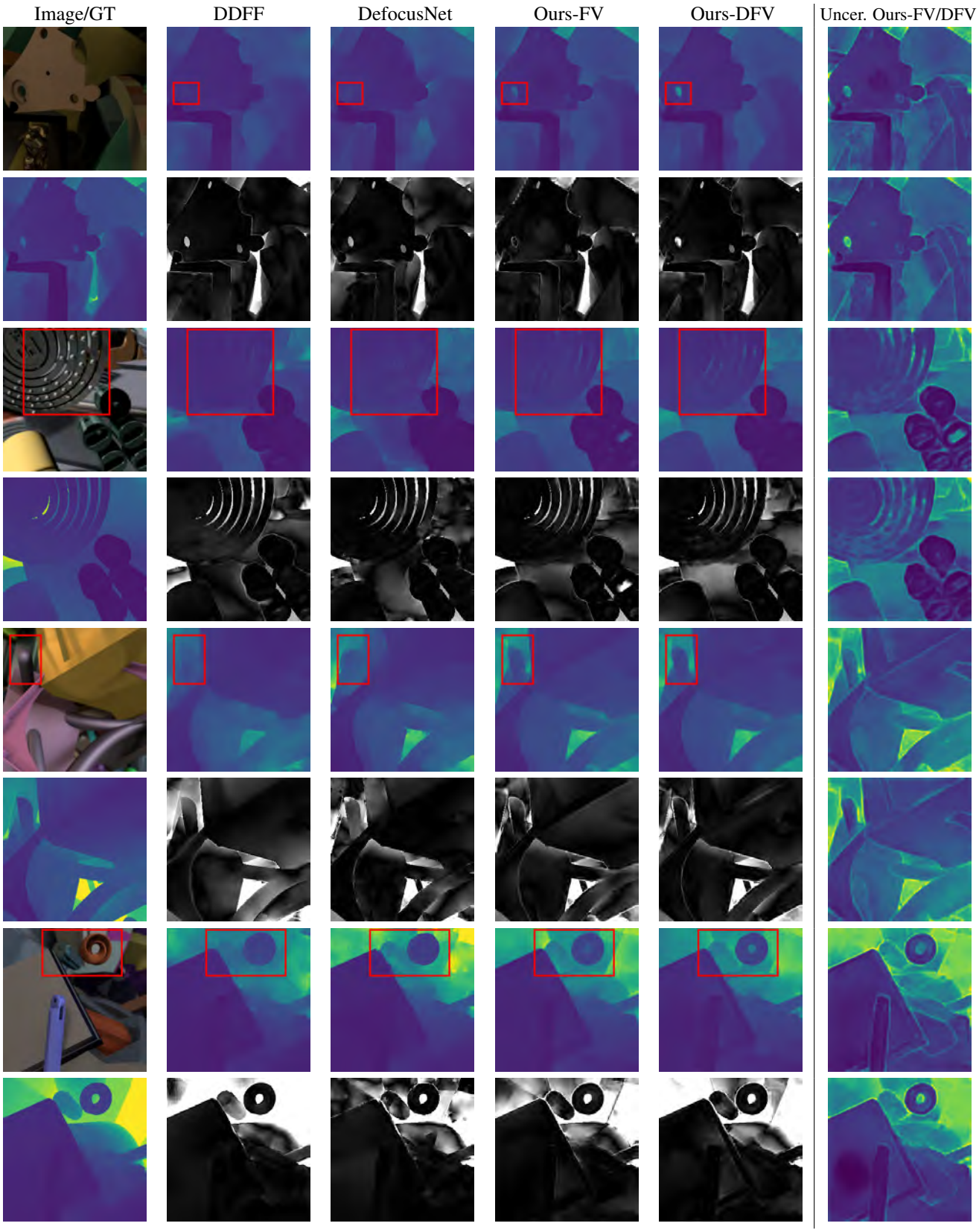

Figure 3. Qualitative results on FoD500 test set. The first column shows the first image in the input focal stack and the corresponding ground truth. The next 4 columns show the depth predictions (rows 1, 3, 5, and 7) and the corresponding error map (rows 2, 4, 6, and 8). The last column presents the corresponding uncertainty maps of Ours-FV (rows 1, 3, 5, and 7) and Ours-DFV (rows 2, 4, 6, and 8). The warmer the color, the higher the value.

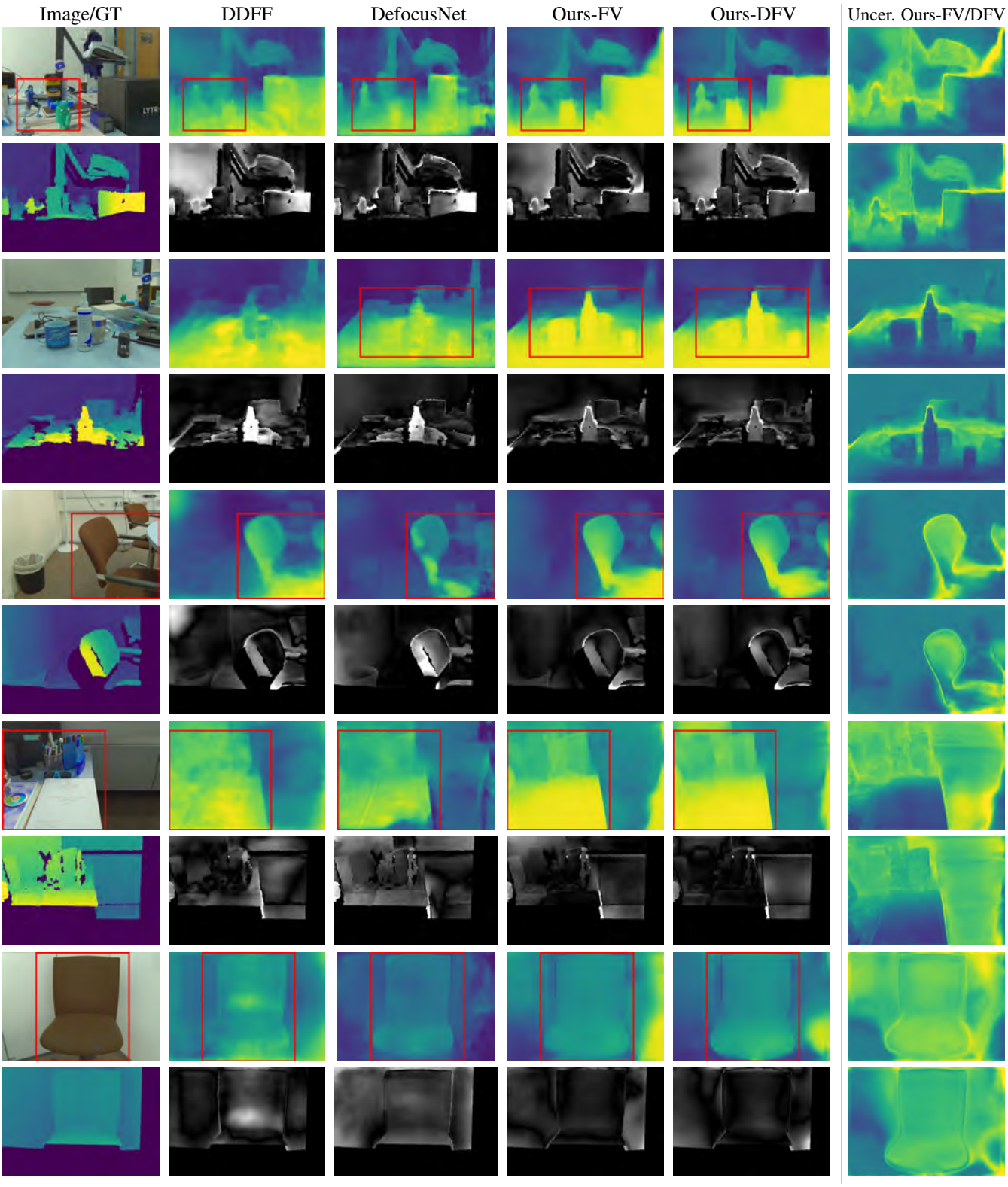

Figure 4. Qualitative results on DDFF12 validation set. The first column shows the first image in the input focal stack and the corresponding ground truth. The next 4 columns show the disparity predictions (rows 1, 3, 5, and 7) and the corresponding error map (rows 2, 4, 6, and 8). The last column presents the corresponding uncertainty maps of Ours-FV (rows 1, 3, 5, and 7) and Ours-DFV (rows 2, 4, 6, and 8). The warmer the color, the higher the value.

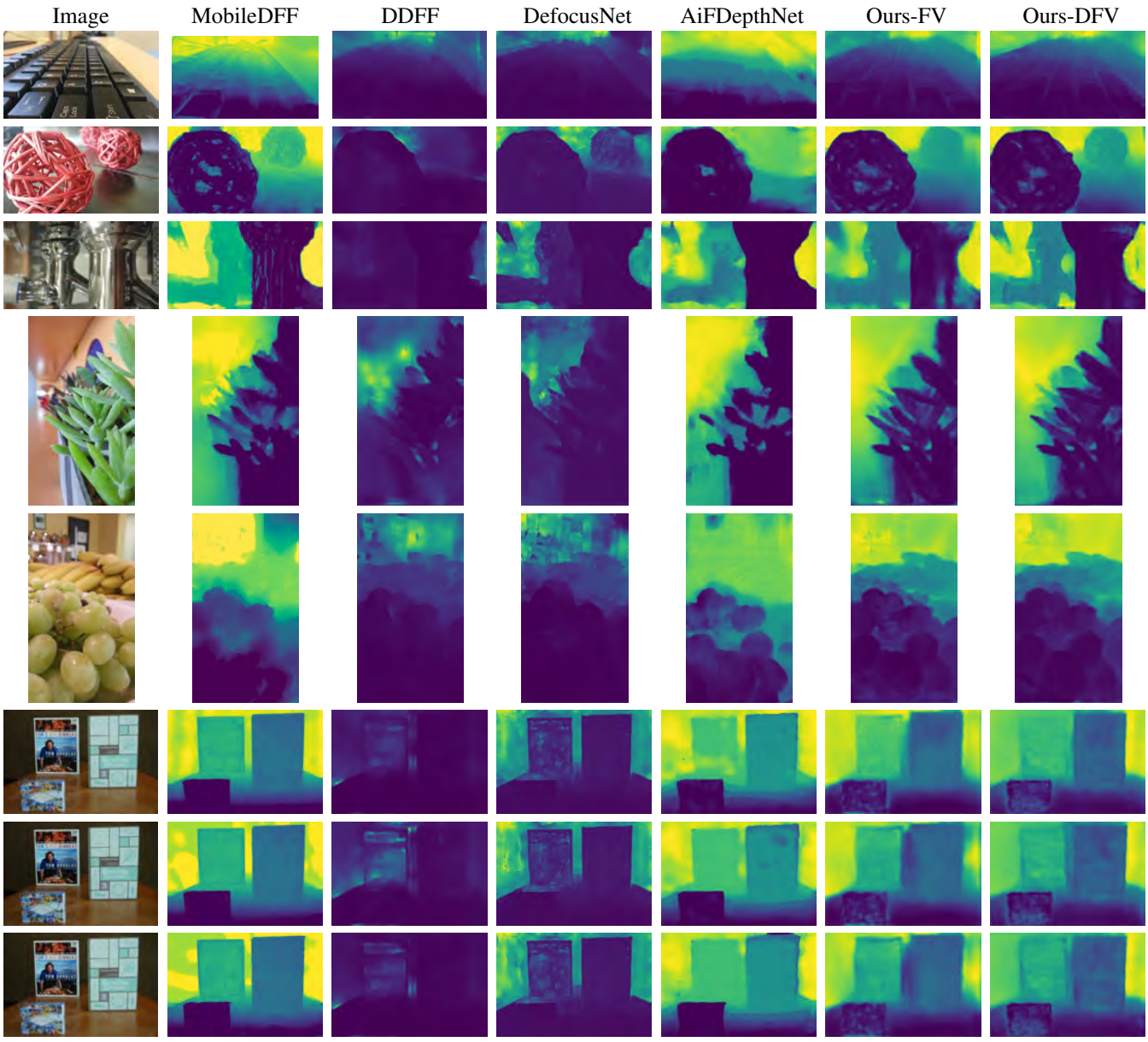

Figure 5. Qualitative results on Mobile depth dataset. The warmer the color, the larger the depth value.

[8] Supasorn Suwajanakorn, Carlos Hernandez, and Steven M Seitz. Depth from focus with your mobile phone. In *CVPR*, pages 3497–3506, 2015. 1

[9] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *CVPR*, 2019. 1