Supplementary Material for "Du^{2}Net: Learning Depth Estimation from Dual-Cameras and Dual-Pixels"

Yinda Zhang, Neal Wadhwa, Sergio Orts-Escolano, Christian Häne,
Sean Fanello, and Rahul Garg

Google Research

In this supplementary material, we provide more information for data collection, implementation details, and quantitative and qualitative results. Some images are also shown on our project webpage[1] for better visualization. For convenience, a copy of the webpage and the video is included in this supplementary material.

1 Data Collection

In this section, we provide information about our data capture rig and how we obtain the ground truth disparity, confidence, and the occlusion mask.

1.1 Data Capture

Fig. 1: Example capture from our data collection rig. Top: Views from the main cameras of the five phones on the rig. Bottom: Views from the telephoto cameras. All ten views are used to compute ground truth depth using multi-view stereo techniques.

As shown in Fig. 4(a) in the main paper, our capture rig consists of five Google Pixel 4 phones. Each phone captures a stereo pair (and dual-pixel data),

[1] https://augmentedperception.github.io/du2net/

giving us ten views of the same scene (Fig. 1). The five phones are synchronized using [1] allowing us to capture dynamic scenes, e.g., plants moving in the wind. Our dataset is captured both indoors and outdoors, and contains both man made and natural scenes.

More examples of the captured dual-camera and dual-pixel images can be found in the project webpage. On each row, we provide a pair of rectified dual-camera (DC) images, I_l and I_r corresponding to the left and the right cameras, and a pair of dual-pixel (DP) images from the right camera sensor, I_t^{DP} and I_b^{DP} corresponding to the top and bottom half-pixels on the sensor. You may toggle between the views by clicking on the images to get a sense of the amount of parallax from the two sources of inputs.

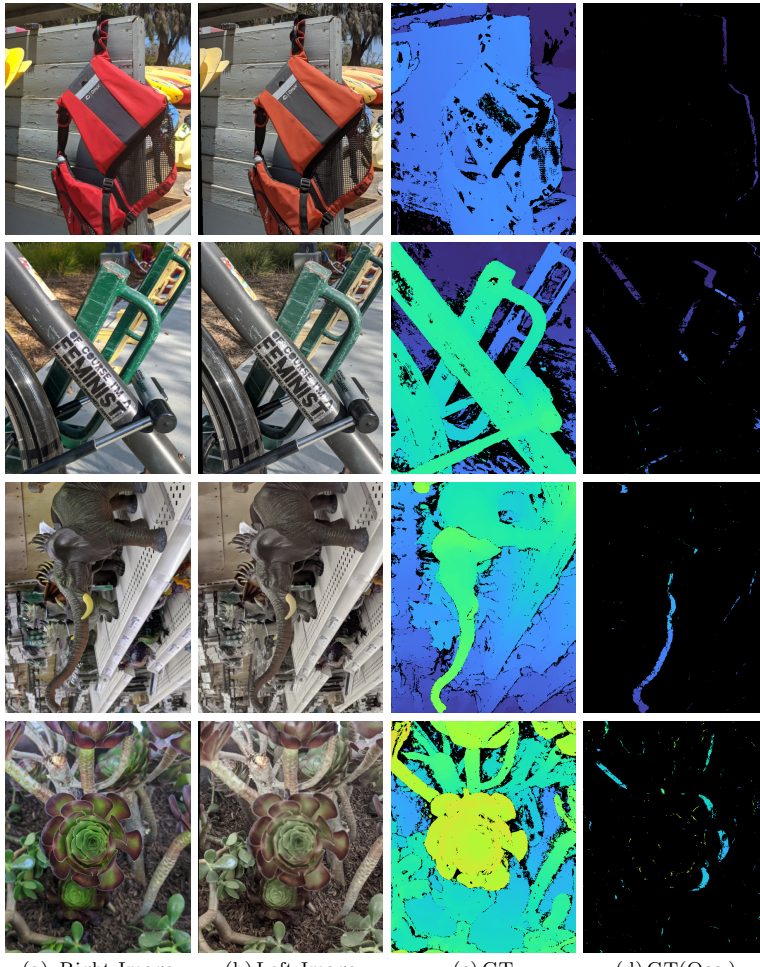
1.2 Computing Ground Truth Disparity D^{gt} and Confidence C^{gt}

We now describe how we compute D^{gt} and C^{gt} given a capture from the rig. Since the rig may not be perfectly rigid, we first compute camera poses, i.e., intrinsics and extrinsics, using structure from motion [7]. For computing ground truth depth using multi-view stereo, we use a method similar to [6] that is designed to give accurate depth for fine structures while avoiding edge fattening artifacts. We describe it in more detail below.

All ten RGB images are resized to 756×1008. For each view, we use a plane sweep algorithm, with 256 planes sampled using inverse perspective sampling between 0.2m and 100m, and take the minimum of a filtered cost volume as each pixel's depth. To compute the cost volume, for each pixel, we compute the sum of absolute differences for each of the warped neighbors and then bilaterally filter the cost volume using the grayscale reference image as the guide image thus avoiding edge fattening artifacts [11]. We use a spatial sigma of 3 pixels and a range sigma of 12.5 for the bilateral filter.

Following [6], we also estimate per-pixel confidence for depth, i.e., a scalar in the range $[0, 1]$. Specifically, we check for depth coherence across views by checking for left / right consistency [3]. We first compute consistency with each of the 9 neighboring images using the consistency measure in [6]. Then, under the assumption that a pixel must be visible in at least two other cameras for its depth to be reliable, we take the product of the largest two consistency values for each pixel to compute our final confidence.

Even though we capture data from all five phones, we only use the data from the center camera for training and testing since it's likely to have the most accurate depth. We use the estimated camera poses for the center phone to rectify the stereo pair [5]. Specifically, we compute $\mathbf{W_l}$ and $\mathbf{W_r}$, i.e., warp maps corresponding to the left and and the right cameras that are applied to the RGB images, depth maps and confidences for the left and the right images. The camera poses are also used to convert depth into disparity between the recitifed pair. See Fig. 2 for examples of rectified RGB images and the corresponding ground truth disparity.



(a) Right Image (b) Left Image (c) GT (d) GT(Occ.)

Fig. 2: Examples of collected data. The right view and left view (a,b) of the binocular stereo pair, the ground truth disparity (c), and the ground truth disparity for occluded pixels (d). Low confidence disparity is rendered in black in (c) and (d).

Further, since the telephoto camera has a smaller field of view than the main camera, we apply a center crop of size 448×560 to all rectified images to restrict ourselves to the area of overlap.

1.3 Computing Occlusion Confidence C^{occ}

As mentioned in the main paper, we also compute C^{occ}, i.e., a per-pixel confidence where the ground truth disparity is accurate but the pixel is occluded in the other camera. This allows us to evaluate and compare the methods in regions that are occluded in one of the stereo views.

To compute it, we first estimate the set of pixels in the right image that are in field of view of the left image but are occluded by an occluder. This can be estimated as:

$$Occ_r = \left\{ (x,y) \quad s.t. \quad \begin{array}{l} 0 \le x' < W, \\ |D_r^{gt}(x,y) + D_l^{gt}(x',y)| > \Delta \\ |D_l^{gt}(x',y) + D_r^{gt}(x' + D_l^{gt}(x',y), y)| \le \Delta \\ |D_l^{gt}(x',y)| > |D_r^{gt}(x,y)| \end{array} \right\} \tag{1}$$

where $x' = x + D_r^{gt}(x,y)$, W is the width of the rectified image, and Δ is set to 1-pixel disparity. The first condition enforces that the pixel is in field of view of the other (left) camera; the second condition ensures that the pixel is not visible in the other camera by checking for failed left-right consistency check; the final two conditions check that the pixel is occluded by an object that is visible in both the views (consistency check succeeds) and is in front of the occluded pixel. Finally, for pixels that are in the set Occ_r we set $C_r^{occ}(x,y)$ to be the product of the confidences of the pixel and the occluder, i.e.,

$$C_r^{occ}(x,y) = \begin{cases} C_r^{gt}(x,y) \cdot C_l^{gt}(x',y), & \text{if} \quad (x,y) \in Occ_r \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

A few examples are shown in Fig. 2. Our conservative criterion for occlusion confidence ensures that we have few false positives.

2 Implementation Details

In this section, we provide more details about confidence volume fusion, network architecture, and evaluation with affine fitting.

2.1 Cost Volume and Confidence Volume

In Sec. 4.2 of the main paper, we fuse confidence volume instead of the cost volume. Here we give more explanation and motivation.

The commonly used cost volume [9] is a 3D volume with two dimensions for the image space (H, W) and one dimension for the disparity space $R = [0, 1, \cdots, d_{max}]$. Each voxel (x, y, d) is a floating number indicating the feature distance if the pixel (x, y) in one view is matched under the given disparity d with the other view. The distance can be computed using various distance metrics, such as ℓ_1 or ℓ_2.

A soft–arg min, which is introduced in Eq. 1 in [9], is used to convert the cost volume into a disparity map. Specifically, a confidence volume is calculated as the softmax on the negative cost volume along the disparity dimension, and output disparity is the sum of disparity hypotheses weighted by the confidence. The operator to convert cost volume to confidence volume can also be written as:

$$\text{Confidence}(x, y, d) = \frac{e^{-\text{Cost}(x,y,d)/t}}{\sum_{d \in R} e^{-\text{Cost}(x,y,d)/t}}, \tag{3}$$

where t controls the sharpness of the softmax and is set to 0.5 in our implementation. The voxels along the disparity dimension for each (x, y) forms a probability distribution, i.e., $\sum_{d \in R} \text{Confidence}(x, y, d) = 1$, indicating the likelihood of each disparity proposal in R being correct (i.e. confidence). Therefore, the output disparity is

$$D(x, y) = \sum_{d \in R} \text{Confidence}(x, y, d) \cdot d. \tag{4}$$

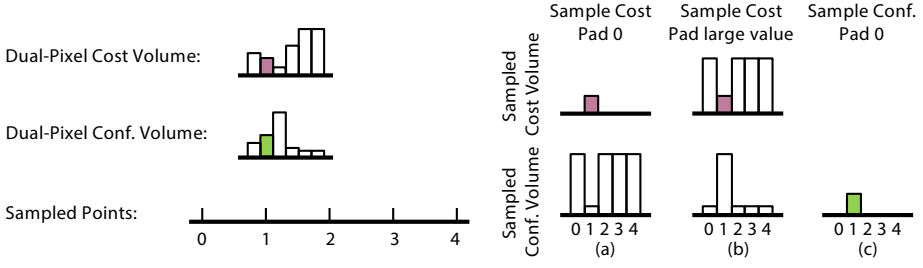


Fig. 3: Explanation of Volume Sampling. See text for details.

We now explain why it is better to sample confidence volume (Eq. 3) instead of the cost volume. Besides normalizing the scales of the two cost volumes, they may produce dramatically different results in some cases.

Fig. 3 illustrates one such case. For simplicity, we drop the image dimension (H, W) and only visualize the disparity dimension R for one pixel. On the left, we show an example of cost volume learned from DP inputs and the corresponding confidence volume. Note that the confidence is inversely related with the cost. Since the DC inputs covers larger range of disparity compared to DP, the warping process in Eq. 3 of the main paper usually samples many points out of the DP disparity range. In Fig. 3, we demonstrate the case where only one sample falls

in the valid range of the DP disparity – sampling point with disparity 1. If we sample cost volume and pad 0 for samples out of range, we obtain a cost volume shown in (a) on the right. The corresponding confidence volume (according to Eq. 3) indicates that 1 is a bad disparity hypothesis and all the others are equally good, which is very inconsistent with the information provided in the original DP cost volume. In the second case (b), we sample cost volume but pad with a large value. Now, the disparity hypothesis 1 becomes the best hypothesis (unlike (a)) but the confidence in 1 is much higher than the original confidence in DP confidence volume. In contrast, if we sample the confidence volume and pad 0 as shown in (c), the produced confidence volume maintains exactly the same confidence from DP volume for disparity 1, while the others are set to zero.

2.2 Network Architecture

We provide detailed network architecture in Fig. 4.

2.3 Evaluation with Affine Fitting

In the Sec. 5.4 of the main paper, we compared to dual-pixel based depth estimation solution DPNet [6]. Since depth from DP can only be predicted up to an unknown affine transformation, Garg et al. [6] first estimate the affine transformation by solving a weighted least squares problem using the ground truth:

$$\hat{\alpha}, \hat{\beta} = \underset{\alpha,\beta}{\operatorname{argmin}} \left\| C^{gt} \cdot ((\alpha + \beta \cdot D_{raw}) - D^{gt}) \right\|^2, \tag{5}$$

where D_{raw} is the network output. $D_{fit} = \hat{\alpha} + \hat{\beta} \cdot D_{raw}$ is then used for computing the metrics. Even though Du^{2}Net produces disparity in absolute scale and is free from this ambiguity, we apply the same post-processing when comparing to the DPNet for fairness (Du^{2}Net* in Tab. 2 of the main paper).

3 More Experiment Results

In this section, we provide more quantitative and qualitative evaluations.

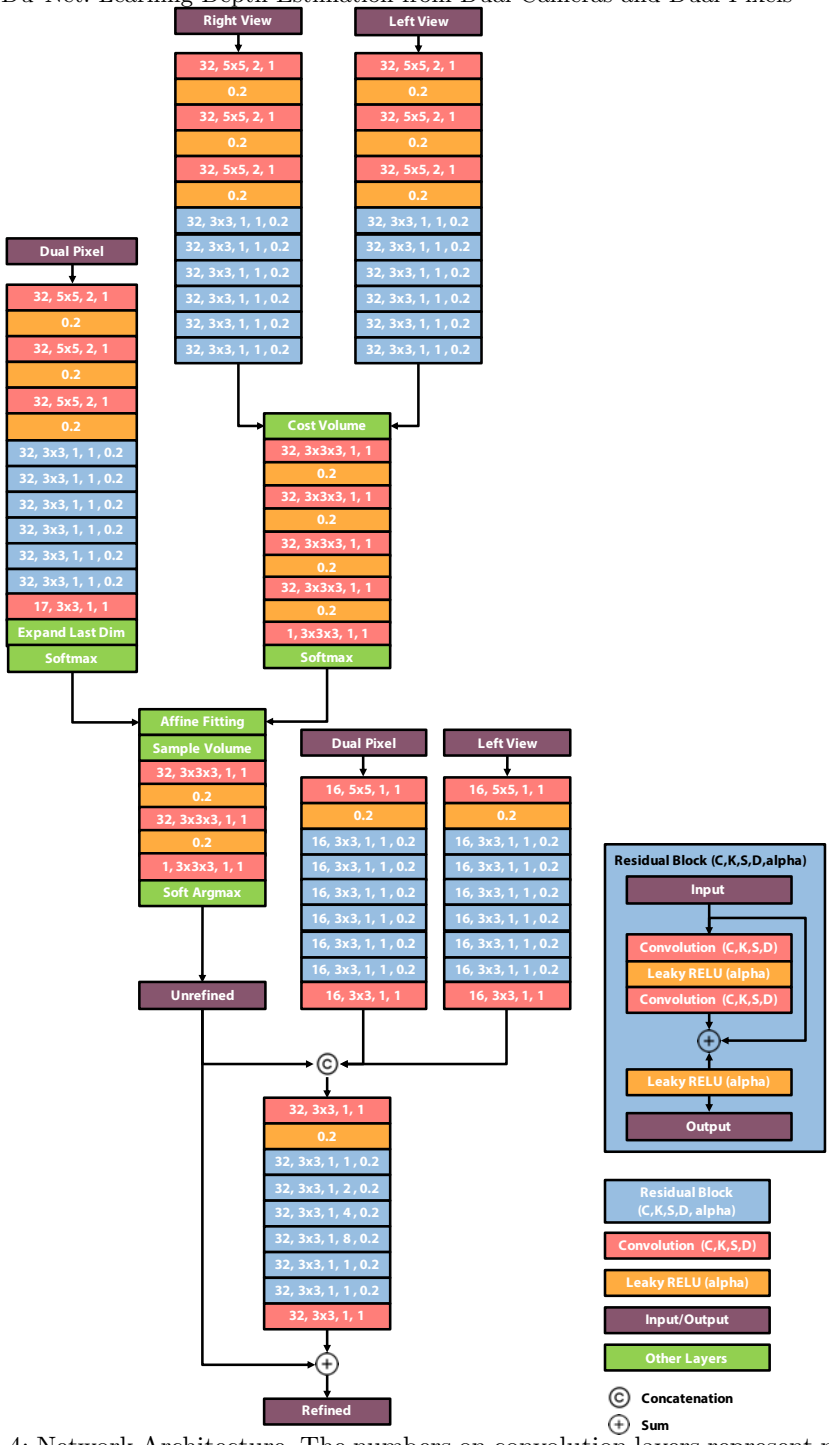


Fig. 4: Network Architecture. The numbers on convolution layers represent number of channels, size of filter, stride, and dilation respectively. The number on the leaky ReLU layer represents the slope for the negative input.

Conf. Volume	All Pixels					Occluded Pixels				
	MAE	RMSE	$\delta>1.25$	$\delta>2$	$\delta>3$	MAE	RMSE	$\delta>1.25$	$\delta>2$	$\delta>3$
DC	1.023	2.502	18.74	10.65	6.32	3.956	6.230	66.33	52.82	40.16
DP+DC (2D)	0.969	2.423	17.37	9.72	5.79	3.718	5.834	65.74	51.55	38.51
DP+DC (C)	0.964	2.372	17.51	9.79	5.80	3.671	5.768	65.49	51.63	38.62
DP+DC (Ours)	**0.902**	**2.252**	**16.16**	**8.96**	**5.26**	**3.526**	**5.523**	**64.98**	**50.82**	**37.41**

Table 1: Ablation study on volume fusion. We compare different ways of fusing DP with the DC confidence volume. '(2D)' indicates fusion of the 2D disparity maps extracted from the two confidence volumes. '(C)' indicates fusing cost volumes instead of confidence volumes.

3.1 Weighted Metrics

Our ground truth comes with a confidence mask (Sec. 1.2), and we use it to calculate weighted evaluation metrics.

$$\text{MAE} = \frac{\sum_p |D(p) - D^{gt}(p)| \cdot C^{gt}(p)}{\sum_p C^{gt}(p)}, \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{\sum_p (D(p) - D^{gt}(p))^2 \cdot C^{gt}(p)}{\sum_p C^{gt}(p)}}, \tag{7}$$

$$\delta > \epsilon = \frac{\sum_p \mathbb{1}(|D(p) - D^{gt}(p)| > \epsilon) \cdot C^{gt}(p)}{\sum_p C^{gt}(p)}, \tag{8}$$

where D is the predicted disparity, D^{gt} is the ground truth disparity, C^{gt} is the confidence map, $\mathbb{1}$ is an indicator function which equals 1 if the condition is true and 0 otherwise, and p is a pixel in the image.

3.2 More Quantitative Ablation Study

In the Sec. 5.3 of the main paper, we showed the quantitative evaluation of our method under different ablations on all pixels C^{gt}. We perform the same comparison on the occluded pixels using C^{occ} (Sec. 1.3) to show the performance in occluded regions.

Tab. 1 shows the evaluation of the unrefined disparity (i.e. the output of the fused volume) under different fusion strategies. Consistent with the conclusion drawn from all pixels, our method outperforms all the others on the occluded regions.

Tab. 2 shows the evaluation of the refined disparity (i.e. the output of the refinement) under different settings. Refinement with DP consistently outperform the case without DP on all the metrics. Using only DP is better than using both under some metrics, which is reasonable since RGB may not be very helpful to recover details in the occluded region and may even hurt the valuable information encoded in DP.

Refinement	All Pixels					Occluded Pixels				
	MAE	RMSE	$\delta>1.25$	$\delta>2$	$\delta>3$	MAE	RMSE	$\delta>1.25$	$\delta>2$	$\delta>3$
RGB	0.838	2.197	14.17	7.74	4.55	2.627	4.866	46.17	33.70	24.18
DP (I)	0.835	2.173	14.25	7.75	4.54	2.518	4.600	46.07	33.25	23.47
DP	0.829	2.184	13.84	7.51	4.45	2.481	4.619	**44.87**	**31.99**	22.54
RGB+DP (Ours)	**0.817**	**2.141**	**13.64**	**7.33**	**4.35**	**2.469**	**4.564**	44.94	32.09	**22.47**

Table 2: Ablation study on refinement. We compare the different ways of using DP to refine the best unrefined disparity from the left and show evaluation on the final disparity. '(I)' indicates that input DP images are warped before computing features for refinement.

3.3 More Qualitative Ablation Study

We show more qualitative comparison in Fig. 5. Our method fusing DP into the cost volume (d) significantly improves the quality of the unrefined disparity compared to the case using only DC (c). Based on this improved unrefined disparity with less error (d), refinement using both DP and DC (f) can further improve the object boundary and thin structures compared to the case using only DC (e).

3.4 Loss Functions

Our loss function is defined on three intermediate disparities in low resolution and the final disparity in high resolution (Eq. 5 in the main paper). The three intermediate disparities are the D_{DP} from dual-pixel confidence volume, D_{DC} from dual-camera confidence volume, and D_{unref} from the fused confidence volume, as explained in Fig. 3 and Sec. 4.2 of the main paper. These low resolution disparities are bilinearly upsampled to the full resolution to compute the loss with the ground truth. The final disparity is the output of the refinement block directly in full resolution. The losses on each disparity are weighted by λs. We tried different combination and found it is important to set λ_{DC} larger. The major reason is that the correctness of D_{DC} is important to ensure correct affine transformations such that the DP channels can learn to produce disparity. The training procedure does not converge if λ_{DC} is too small.

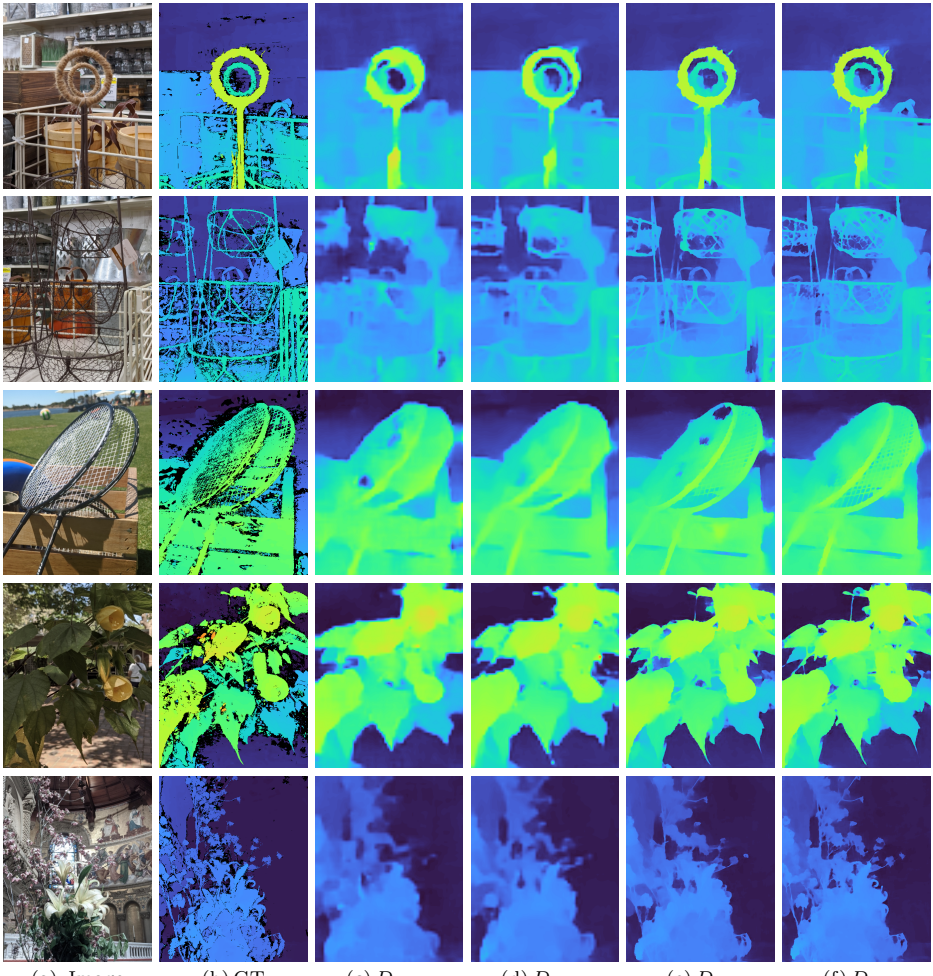
We also train our model with a weighted Charbonnier loss [2]:

$$L(D) = \frac{\sum_{\mathbf{p}}(\sqrt{((D(\mathbf{p})-D^{gt}(\mathbf{p}))/c)^2+1}-1)\cdot C^{gt}(\mathbf{p})}{\sum_{\mathbf{p}} C^{gt}(\mathbf{p})}, \tag{9}$$

where D^{gt} is the ground truth disparity, C^{gt} is the per-pixel confidence of the ground truth and c is a hyper-parameter (scale factor) we set to 2 for disparity in range $[0, 128]$. The performance is shown in Tab. 3. Our model achieves better metrics with Huber loss except the RMSE than using Charbonnier loss [2].

3.5 More Qualitative Comparison to SOTA

We show more qualitative comparison to state-of-the-art stereo and DP based approaches in Fig. 6, 7, and 8. Compared to other stereo based approaches



(a) Image (b) GT (c) D_{unref} DC, RGB+DP (d) D_{unref} Du^{2}Net (e) D_{ref} DP+DC,RGB (f) D_{ref} Du^{2}Net

Fig. 5: Ablations of our method. The right camera image (a), ground truth disparity (b) with low confidence disparity in black, D_{unref} (c) from an ablation where only the DC input is used for the confidence volume, D_{unref} (d) from Du^{2}Net, D_{ref} (e) from an ablation where only the RGB image is used for refinement, and D_{ref} (f) from Du^{2}Net. DP input is useful for both the confidence volume and refinement stages to recover accurate depth for fine structures and occluded regions.

Method	Loss	All Pixels					Occluded Pixels				
		MAE	RMSE	$\delta>1.25$	$\delta>2$	$\delta>3$	MAE	RMSE	$\delta>1.25$	$\delta>2$	$\delta>3$
PSM-Net [4]	Huber	0.815	2.289	13.48	7.98	4.97	2.799	5.188	44.59	34.73	26.78
Du^{2}Net	Charbonnier	0.817	**2.141**	13.64	7.33	4.35	2.469	4.564	44.94	32.09	22.47
Du^{2}Net	Huber	**0.802**	2.147	**13.21**	**7.17**	**4.25**	**2.396**	**4.543**	**42.37**	**30.62**	**21.91**

Table 3: Performance of our model with different loss functions. Our model achieves better metrics with Huber loss.

[4, 10] that only take DC as the input, our method performs better at object boundary and thin structures. Compared to dual-pixel only approach [6], our method produces significantly better depth for distant areas in the background while maintaining the foreground details (Fig. 8).

3.6 Analysis of Best and Worst Cases

In Fig. 9 we show representative images from the best (top 3 rows) and the worst (bottom 3 rows) results for our method as ranked by the MAE metric. As expected, the method performs very accurately on images with high frequency details and textured scenes whereas it does worse (along with other methods) in textureless areas.

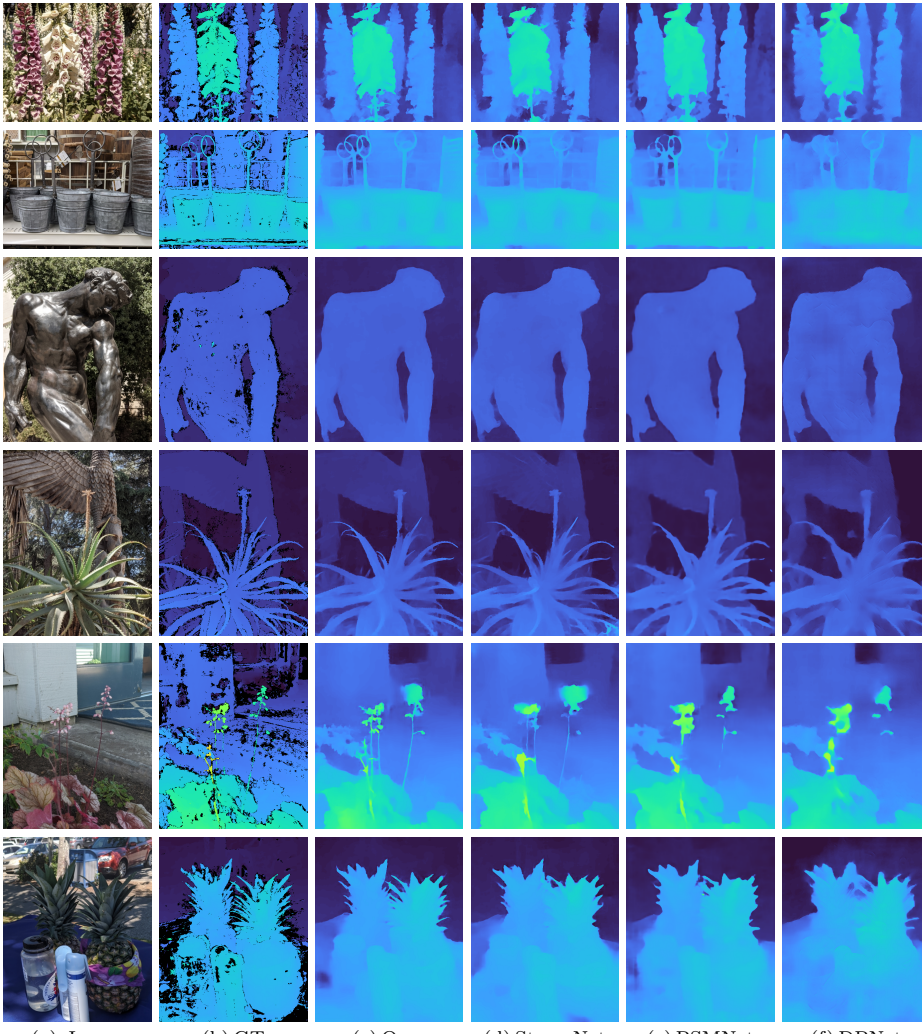
3.7 More Results for Applications

We show more examples of computational photography applications. Fig. 10 shows results of synthetic shallow depth-of-field effect using disparity from different models. Our method produces better details for object boundary and thin structure, which prevents artifacts near the subject boundary.

We also provide more comparisons on the 3D photo [8] in supplementary video in the project webpage. Again, our more accurate depth minimizes visual artifacts like unnatural distortion of rigid scene structures and bleeding between foreground and background.

References

1. Ansari, S., Wadhwa, N., Garg, R., Chen, J.: Wireless software synchronization of multiple distributed cameras. In: ICCP (2019)
2. Barron, J.T.: A general and adaptive robust loss function. In: CVPR (2019)
3. Bleyer, M., Rhemann, C., Rother, C.: Patchmatch stereo-stereo matching with slanted support windows. In: BMVC (2011)
4. Chang, J., Chen, Y.: Pyramid stereo matching network. In: CVPR (2018)
5. Fusiello, A., Trucco, E., Verri, A.: A compact algorithm for rectification of stereo pairs. Machine Vision and Applications (2000)
6. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: ICCV (2019)



(a) Image (b) GT (c) Ours (d) StereoNet (e) PSMNet (f) DPNet

Fig. 6: Qualitative comparison to state-of-the-art stereo and DP based methods.

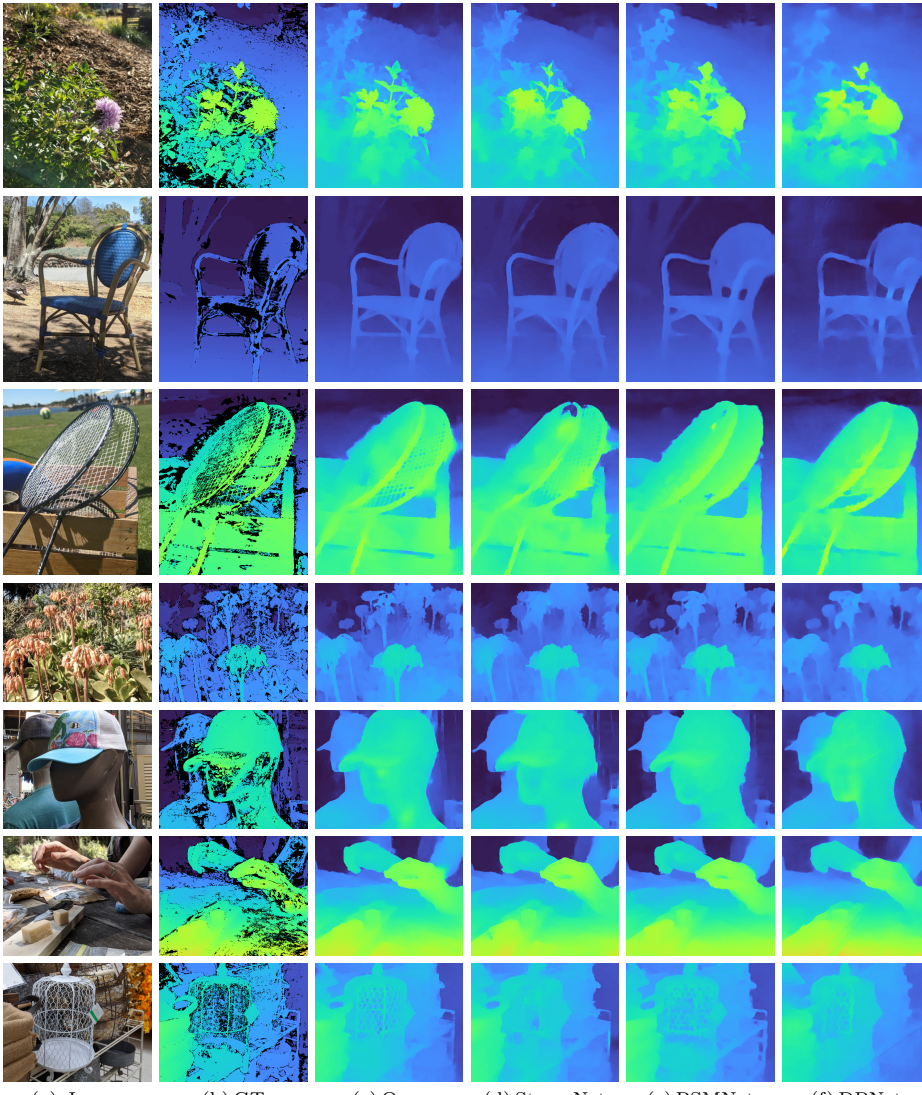
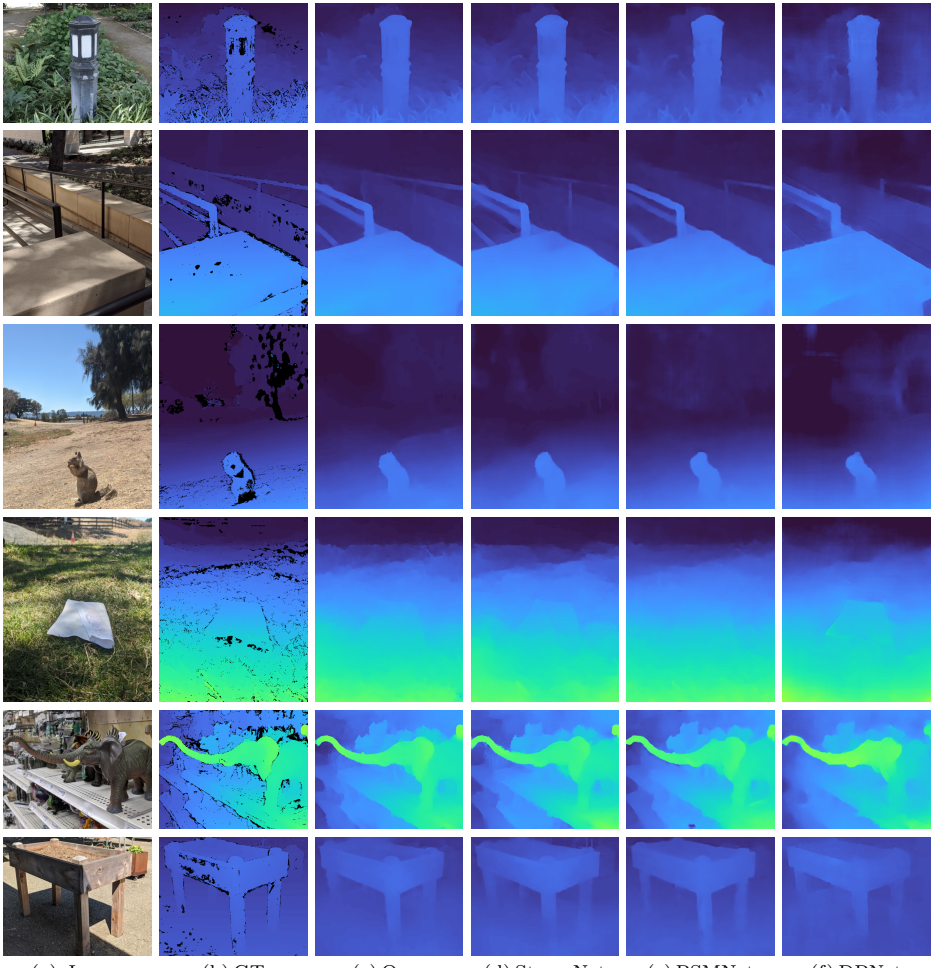
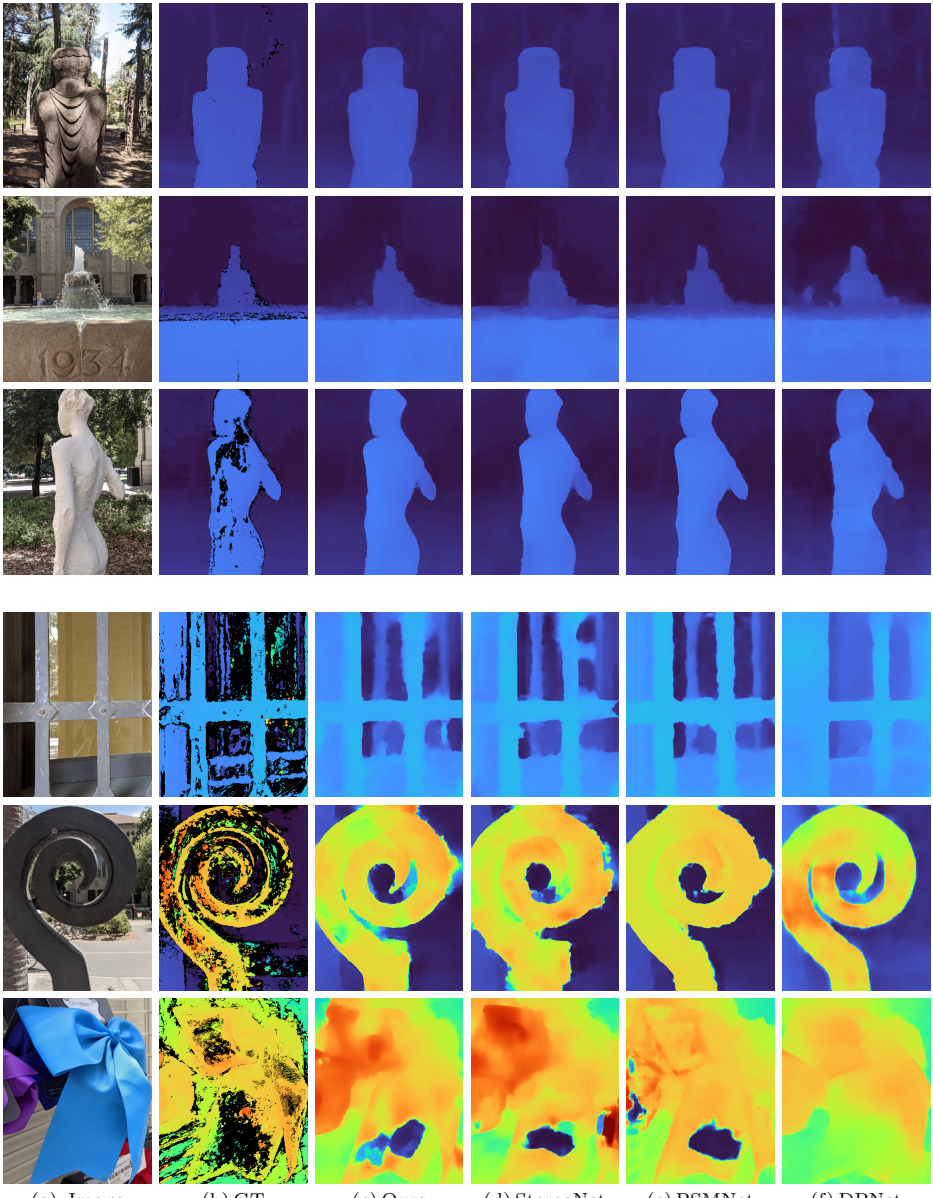


Fig. 7: Qualitative comparison to state-of-the-art stereo and DP based methods.



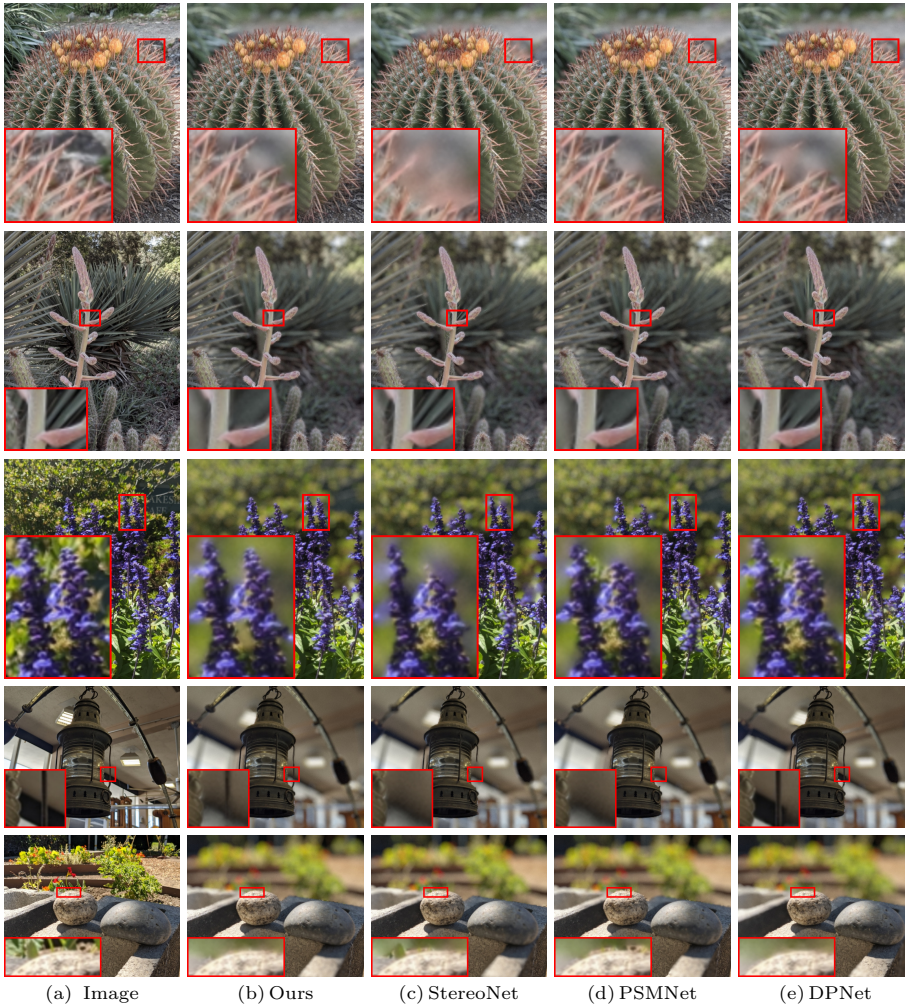
(a) Image (b) GT (c) Ours (d) StereoNet (e) PSMNet (f) DPNet

Fig. 8: DPNet (f) performs worse in distant areas compared to methods that take DC as input ((c), (d), (e)) due to the small baseline between the two dual-pixel images.



(a) Image (b) GT (c) Ours (d) StereoNet (e) PSMNet (f) DPNet

Fig. 9: Representative images from our best (top 3 rows) and worst (bottom 3 rows) results, as rated by MAE metric.



(a) Image (b) Ours (c) StereoNet (d) PSMNet (e) DPNet

Fig. 10: Synthetic shallow depth-of-field results for different methods. Accurate depth near occlusion boundaries is critical for avoiding artifacts near the subject boundary.

7. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2003)
8. Hedman, P., Kopf, J.: Instant 3D Photography. SIGGRAPH (2018)
9. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: CVPR (2017)
10. Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In: ECCV (2018)
11. Richardt, C., Orr, D., Davies, I., Criminisi, A., Dodgson, N.A.: Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In: ECCV (2010)