Facial Depth and Normal Estimation using Single Dual-Pixel Camera[⋆]

Minjun Kang[1†] Jaesung Choe[1] Hyowon Ha[4] Hae-Gon Jeon[2]
Sunghoon Im[3] In So Kweon[1] Kuk-Jin Yoon[1]

[1]KAIST [2]GIST [3]DGIST [4]Meta Reality Labs
[†]kmmj2005@kaist.ac.kr

Abstract. Recently, Dual-Pixel (DP) sensors have been adopted in many imaging devices. However, despite their various advantages, DP sensors are used just for faster auto-focus and aesthetic image captures, and research on their usage for 3D facial understanding has been limited due to the lack of datasets and algorithmic designs that exploit parallax in DP images. In this paper, we introduce a DP-oriented Depth/Normal estimation network that reconstructs the 3D facial geometry. In addition, to train the network, we collect DP facial data with more than 135K images for 101 persons captured with our multi-camera structured light systems. It contains ground-truth 3D facial models including a depth map and surface normal in metric scale. Our dataset allows the proposed network to be generalized for 3D facial depth/normal estimation. The proposed network consists of two novel modules: Adaptive Sampling Module (ASM) and Adaptive Normal Module (ANM), which are specialized in handling the defocus blur in DP images. Finally, our proposed method achieves state-of-the-art performances over recent DP-based depth/normal estimation methods.

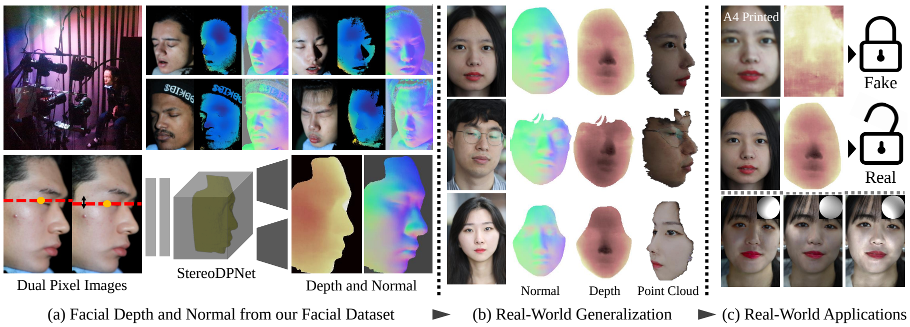
Keywords: Dual-Pixel · Depth/Normal estimation · Face Reconstruction

1 Introduction

A huge number of facial images are posted every day on social media [16, 17]. Accordingly, acquiring facial geometry from images has emerged as an interesting research topic, since 3D facial geometry can be used for various applications [22, 5, 52, 33, 54, 50]. 3D facial geometry can be obtained by either using multiple cameras [14, 4] or active sensing devices [28, 26]. However, these methods often suffer from uncontrolled lighting conditions or hardware synchronization.

Recently, **Dual-Pixel (DP)** sensors get noticed due to their popularity in being installed in many portable imaging devices such as the iPhone13 ProMax and Samsung Galaxy 22 and their strengths of capturing two images perfectly synchronized with the same exposure, white balance, and geometric rectification. Based on these properties, currently, there have been few studies that explore the possibility of DP sensors for scene depth estimation [15, 36, 53, 35, 49]. Usually,

[⋆] This paper is the short version of ECCV'22 and is NEVER considered an official publication.



(a) Facial Depth and Normal from our Facial Dataset (b) Real-World Generalization (c) Real-World Applications

Fig. 1. Our method aims at the generalized estimation of unmet facial geometry, which can be used for various applications, such as face spoofing or relighting.

these studies regard DP images as extremely narrow-baseline stereo images having different defocus-blur to infer depth maps. Although the DP sensors are actively used to take face pictures, there has been a limited study [48] that recovers facial geometry using a Dual-Pixel camera. Previous methods have difficulty in facial geometry estimation, which is due to the lack of a facial DP dataset with precise 3D geometry and an appropriate algorithm for generalized estimation.

To address the issue, we present a DP-oriented 3D facial dataset and a depth/normal estimation network toward high-quality facial geometry reconstruction with DP cameras. We represent the 3D facial geometry not only with the depth map but also with the normal map for various applications such as face relighting. Our dataset involves 135,744 face data for 101 persons consisting of DP images and their corresponding depth maps and surface normal maps, which are captured by our structured light camera system. Based on these data, we train our depth/normal estimation network, called stereoDPNet, to infer 3D facial information from DP images. In particular, our stereoDPNet is fully oriented from the properties of dual-pixel images that have an extremely small range of disparity with defocus-blur. Our network design carefully treats these distinctive properties through our Adaptive Sampling Module (ASM) and Adaptive Normal Module (ANM). Finally, the contributions are as follows:

- DP-oriented 3D facial dataset with more than 135K DP images and their corresponding high-quality 3D models.
- Novel depth/normal estimation network for facial 3D reconstruction from a DP image with better generalization.

2 Proposed Method

This paper covers dual-pixel based facial understanding: from data acquisition (Section 2.2) to general estimation by stereoDPNet (Section 2.3). Different from natural images from typical cameras, dual-pixel sensors capture images having an extremely small range of disparity as well as defocus-blur, as shown in Figure 2-(c). Through our carefully designed dataset and network, we design a well-generalized methodology that even can infer facial geometry from unmet DP facial images.

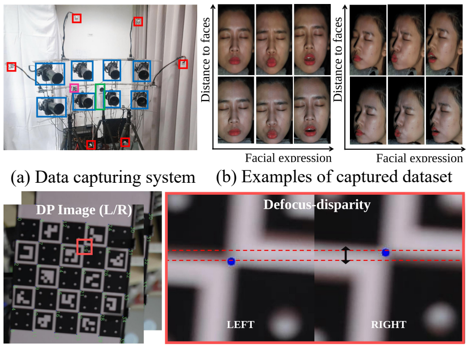
2.1 Preliminaries

Depth Estimation from Dual-Pixel. Dual-Pixel images can be considered as a pair of stereo images since the DP camera captures two sub-aperture images with small parallax. There exists an extremely small range of pixel discrepancy from the same scene point between these two images ($-4\text{px} \sim +4\text{px}$) and pioneer works [45, 15] introduce the affine relationship between metric depth and the defocus-disparity driven from the paraxial and thin-lens approximations.

$$\begin{aligned} d(x,y) &= \alpha \bar{b}(x,y) \\ &\approx \alpha \frac{Lf}{1-f/g}\left(\frac{1}{g} - \frac{1}{Z(x,y)}\right) \\ &\triangleq A(L,f,g) + \frac{B(L,f,g)}{Z(x,y)}, \end{aligned} \tag{1}$$

Based on this property, there have been several works that estimate scene depth from Dual-Pixel by adopting simple U-Net [15], using additional stereo camera [53], parametrized point spread functions [36, 35], and multiplane image representation [49]. Compared to these works, our proposed cost-volume-based network provides better geometry by capturing this narrow range of defocus-disparity and converting to metric depth by using Equation (1).

Monocular Face Reconstruction. In general, it is only available to reconstruct faces from monocular images with a limited assumption [46], many 3D face regression methods [39, 13, 18] rely on the prior knowledge of face morphable model [44, 7], facial keypoints/landmarks [13, 41, 6], and symmetric assumption [47]. Recently, generative model-based methods [9, 8] are also actively explored. However, these methods lead to failure with unmet conditions (*e.g.* extreme poses), and the biased result to the prior knowledge/training dataset.



(a) Data capturing system (b) Examples of captured dataset

(c) Captured defocus-disparity from Dual-Pixel

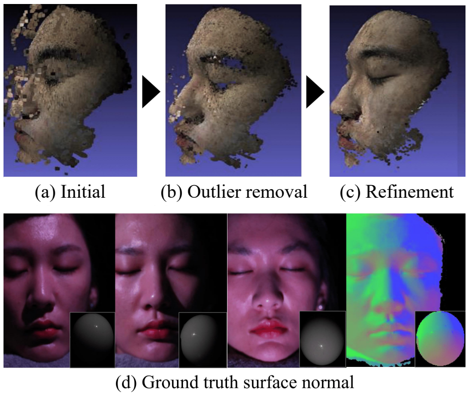
Fig. 2. Capturing face with Dual-Pixel. (a) (2×4 multi-camera array (blue), 6 LEDs (red), a projector (green), and a LED controller (magenta). (b) Examples of the captured facial dataset. (c) Example of captured defocus-disparity in DP images.

2.2 Dual-Pixel Facial Dataset

Dataset Configuration. Given an array of multiple DP cameras, we capture various human faces with different expressions and light conditions. The dataset consists of 135,744 photos, which are a combination of 101 people, eight cameras, seven different lighting condition, four facial heading directions (left, right, center and upward), three facial expressions (normal, open mouth and frown), and two fixed distances of subjects from the camera array, as illustrated in Figure 2-(b). The distances between the camera array and subjects range from 80 *cm* to 110 *cm*.

Since the focus distance is about 97 *cm*, our captured images contain both front focused and back focused cases. Our dataset includes 44,352 female photos as well as 91,392 male photos, ages range from 19 to 45. In main experiments, we use 76 people (76%) as a train set and the others (24%) as a test/validation set without any overlap with the train set.

Ground Truth Data Acquisition. Structured light systems are designed for high-quality 3D geometry acquisition under controlled environments by projecting pre-defined patterns on surfaces of objects [40, 20, 12] and by analyzing the projected patterns to measure 3D shapes of the objects. It is extensively used for ground-truth depth maps in stereo matching benchmarks [24, 1, 43] and shape from shading [21]. In this work, we tailor the structured light-based facial 3D reconstruction method [19] with our well-synchronized multi-camera system. Thanks to our capturing system and structured light-based reconstruction method, we obtain dense, high-quality facial 3D corresponding to high-resolution DP images in Figure 1-(a). Moreover, we calibrate point light directions by using a chrome ball and applying a photometric stereo in [34] to obtain accurate surface normal maps of subjects' faces in Figure 3(d). We utilize the RANSAC algorithm in obtaining both surface normal and albedo for robust estimation by excluding severe specular reflection. By using the surface normals, initial depth is refined by conforming the initial facial depth and the surface normal [34], as illustrated in Figure 3(a), (b), and (c).


(a) Initial (b) Outlier removal (c) Refinement

(d) Ground truth surface normal

Fig. 3. Ground-truth depth and surface normal acquisition. (a) Initial depth from the structured light. (b) Depth after removing outliers. (c) Depth via fusion of the initial depth and the surface normal obtained from the photometric stereo in (c).

2.3 Facial Depth and Normal Estimation

Overall Architecture. Given DP images with left I^L and right I^R, stereoDPNet is trained to infer a disparity map $\hat{d}$ and a surface normal map $\hat{n}$. To do so, first, the feature extraction layer infers DP image features F^L and F^R, respectively. Second, using F^L and F^R, the proposed ASM captures an amount of spatially varying blur in dynamic ranges, and then adaptively samples the features. Then, the sampled features G^L and G^R are stacked into a cost volume $\mathcal{V}$. Third, the cost volume is aggregated through three stacked hourglass modules to infer the aggregated cost volume C_A. Lastly, this aggregated volume C_A is used to regress a disparity map following the baseline and infer a surface normal map by ANM.

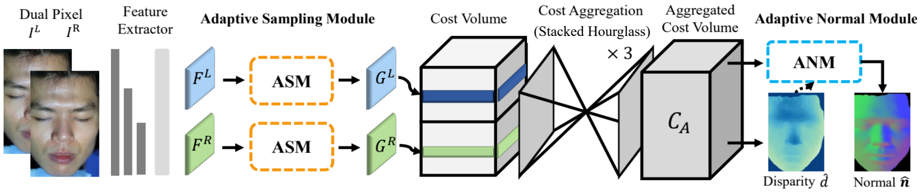


Fig. 4. Architecture of StereoDPNet. Given DP images, our network is trained to infer facial depth/normal maps. Our two key modules, Adaptive Sampling Module and Adaptive Normal Module overcome the extremely narrow baseline in DP images by capturing disparities in blurry regions. Note that pre-calibrated disparity to depth conversion using Equation (1) is used to get metric-scale depth and normal.

Adaptive Sampling Module. To cope with narrow disparity range and defocus-blur in DP, we design ASM in Figure 5 inspired by defocus blur matching method [11] and depth from light-field image [25].

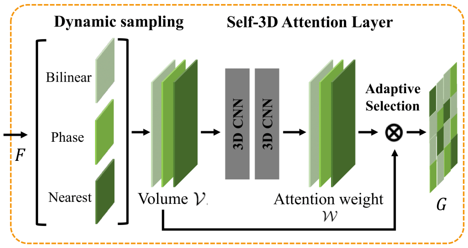


Fig. 5. Adaptive Sampling Module (ASM) consists of a dynamic sampling and a self-3D attention layer.

According to Jeon *et al.* [25], the sub-pixel shift from different sampling strategies to construct cost volume for matching provides varying results depending on the local scene configurations. To take advantage of various conventional sampling methods, we incorporate them into ASM. The dynamic sampling layer in ASM is designed with a combination of nearest-neighbor, bilinear, and phase-shift interpolation, which can have various receptive fields to find varying blur sizes and can obtain subpixel-level shifted features of F^L and F^R. To this end, the shifted features from the three different sampling strategies are concatenated into a volumetric feature $\mathcal{V}$.

To extract useful features from $\mathcal{V}$, we design a self-3D attention layer. Our self-3D attention layer adaptively selects sampling strategies using attention map $\mathcal{W}$ to include prominent texture information in an extracted feature map. Finally, the sampled features with the sub-pixel shift, G^L and G^R, are obtained by averaging the sampled volume $\mathcal{V}_S$. The matching cost volume, constructed from the selected feature maps (G^L, G^R), contains rich texture information with relative blur and performs effective matching in homogeneous regions as well [11].

Adaptive Normal Module. We design ANM in Figure 6 to produce a surface of human faces complementary to an estimated defocus-disparity map.

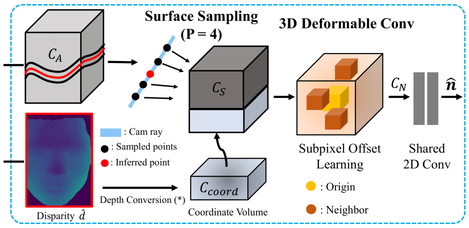


Fig. 6. Adaptive Normal Module (ANM) regresses surface normal by surface sampling and a 3D deformable CNN.

According to [29], an accurately aggregated cost volume contains an implicit function representation of underlying surfaces for depth estimation. Since the surface normal mainly depends on the shape of the local surface, it is redundant to use all voxel embeddings in C_A for facial normal estimation. We thus sample the P candidates of hypothesis planes among M planes from the aggregated volume C_A using the estimated disparity map (Equation (2)). Since the surface normal is defined with the metric scale depth, we convert disparity to a depth map using pre-calibrated Equation (1) and provide this volumetric information with our network denoted as coordinate volume C_{coord}.

Since a human face has a variety of curved local surfaces, we need to consider dynamic ranges of neighbors to extract a local surface from the sampled hypothesis planes C_S in the previous stage. To do this, we follow the assumption of local plane in [31, 37, 32] and form local planes by a small set of neighbor points. Since these local patches have arbitrary shapes and sizes composed with its sampled neighboring points, we use 3D deformable convolutions [51] to consider the neighboring points within the dynamic ranges. The learnable offsets of the deformable convolution in 3D space allow us to adaptively sample neighbors and find the best local plane. The final feature volume C_N is predicted after passing two 3D deformable convolution layers to extract surface normal $\hat{\mathbf{n}}$.

2.4 Loss Functions

The aggregated volume C_A passes through a classifier to produce a final matching cost $\mathcal{A}$, and the softmax function $\sigma(\cdot)$ is applied to regress the defocus-disparity $\hat{d}$. Accordingly, we compute the disparity as follows:

$$\hat{d}_{u,v} = \sum_{m=1}^{M} d^m \cdot \sigma\left(\mathcal{A}^m_{u,v}\right), \tag{2}$$

where $\hat{d}_{u,v}$ is the defocus-disparity and $\mathcal{A}_{u,v}$ is the final matching cost at a pixel (u, v). M and d^m are the range of defocus-disparity, and predefined discrete disparity levels, respectively. Following [10], we minimize a disparity loss $\mathcal{L}_{\text{disp}}$ using a smooth L_1 loss as follows:

$$\mathcal{L}_{\text{disp}} = \frac{1}{H \cdot W} \sum_{u=1}^{W} \sum_{v=1}^{H} \mathcal{M}_{u,v} \cdot \text{smooth}_{L_1}\left(d_{u,v} - \hat{d}_{u,v}\right), \tag{3}$$

where $d_{u,v}$ is a ground-truth defocus-disparity at a pixel (u, v) converted from the ground-truth metric scale depth and $\mathcal{M}_{u,v}$ is the facial mask in Section 2.2.

For the surface normal estimation, shared 2D convolutions are applied to the feature volume C_N to regress a surface normal. The final convolutional layers follow the same structure of the baseline architecture in [29]. Finally, we train ANM by minimizing a cosine similarity normal loss $\mathcal{L}_{\text{normal}}$ as:

$$\mathcal{L}_{\text{normal}} = \frac{1}{H \cdot W} \sum_{u=1}^{W} \sum_{v=1}^{H} \mathcal{M}_{u,v} \cdot \left(1 - \mathbf{n}_{u,v} \cdot \hat{\mathbf{n}}_{u,v}\right), \tag{4}$$

where $\mathbf{n}_{u,v}$ and $\hat{\mathbf{n}}_{u,v}$ are a ground-truth, and a predicted normal at a pixel (u, v). Finally, our StereoDPNet is fully supervised by our constructed dataset in Section 2.2 and minimizing the combination of Equation (3) and Equation (4).

Method	Task	Absolute error metric [mm] ↓					Affine error metric [px] ↓			Accuracy metric ↑	
		AbsRel	AbsDiff	SqRel	RMSE	RMSElog	WMAE	WRMSE	$1-\rho$	$\delta<1.01$	$\delta<1.01^2$
PSMNet [10]	ST	0.006	5.314	0.054	6.770	0.008	0.093	0.126	0.054	0.818	0.983
StereoNet [27]	ST	0.005	4.306	0.038	5.811	0.006	0.112	0.150	0.087	0.903	0.991
DPNet [15]	DP	0.008	7.175	0.092	8.833	0.010	0.110	0.148	0.086	0.688	0.959
MDD [36]	DP	-	-	-	-	-	1.830	2.348	0.575	-	-
BTS [30]	M	0.007	6.575	0.081	8.102	0.009	0.111	0.150	0.077	0.731	0.964
NNet [29]	DN	0.004	3.608	0.027	4.858	0.005	0.073	0.102	0.048	0.934	0.995
Ours	DN	**0.003**	**2.864**	**0.019**	**3.899**	**0.004**	**0.064**	**0.091**	**0.034**	**0.966**	**0.995**

Table 1. Depth Benchmark Results. Our proposed method outperforms the existing stereo matching methods [10], [27], DP-oriented state-of-the-art methods [15], [36], monocular depth estimation [30], and depth/normal network for stereo matching [29]. Note that MDD [36] cannot be measured by absolute metrics since it adopts different defocus-disparity geometry of Equation (1). ST, DP, M, and DN denotes "Stereo Matching", "DP-oriented method", "Monocular", and "Depth and Normal", respectively.

Method	ANM		Absolute [mm] ↓		Affine [px] ↓			Accuracy ↑		Normal [deg] ↓	
	SS	D3D	AbsDiff	RMSE	WMAE	WRMSE	$1-\rho$	$\delta<1.01$	$\delta<1.01^2$	MAE	RMSE
ASM Only			4.895	6.223	0.095	0.127	0.056	0.850	0.992	-	-
NNet [29]			3.608	4.858	0.073	0.102	0.048	0.934	0.995	9.634	11.877
ASM + NNet			3.271	4.434	0.064	0.090	**0.033**	0.947	**0.997**	9.072	11.045
ASM + NNet	✓		3.214	4.519	**0.062**	**0.089**	0.037	0.943	0.990	8.894	10.837
StereoDPNet	✓	✓	**2.864**	**3.899**	0.064	0.091	0.034	**0.966**	0.995	**7.479**	**9.386**

Table 2. Normal Benchmark Results with Ablation Study of ANM. NNet [29] is a baseline model of our overall architecture. We compare the performance of depth and surface normal estimation by adding each component. SS denotes "Surface Sampling" and D3D denotes "Deformable 3D convolution" of ANM respectively.

3 Experiments

To evaluate the effectiveness and the robustness of our work, we carry out various experiments on our dataset as well as DP images captured under real-world environments. We use evaluation metrics in a public benchmark suite[1] and affine invariant metrics [15] for the evaluation of estimated depth in Table 1 and Table 3. To measure the quality of normal map in Table 2, we use the metric following the DiLiGenT benchmark [42].

Depth Benchmark. We compare our method with recent DP-based depth estimation approaches [15, 36] as well as widely used stereo matching networks [10, 27], a depth/normal network for stereo matching [29] and a state-of-the-art monocular depth estimation network [30], whose results are reported in Table 1 and in Figure 7. While other methods fail to handle defocus blur or struggle to find correspondences in human faces, our method predicts the most outstanding and stable results not only with a test set but also with unmet wild examples. In addition to our facial benchmark, we conduct an additional experiment on another real-world DP dataset [36] in Table 3 to validate the generalization of our network. For a fair comparison, we augment our network using synthetic DP dataset [3] and don't apply any post-processing (*i.e.* bilateral or guided filter).

Surface Normal Benchmark. To the best of our knowledge, this is the first attempt to estimate both the surface normal and the defocus-disparity from

[1] http://www.cvlibs.net/datasets/kitti/

Metrics	Method					
	PSMNet [10]	StereoNet [27]	DPNet [15]	MDD [36]	NNet [29]	**Ours**
WMAE (↓)	0.102	0.111	0.132	0.107	0.103	**0.085**
WRMSE (↓)	0.154	0.214	0.192	0.168	0.143	**0.133**
$1-\rho$ (↓)	0.351	0.261	0.420	**0.187**	0.345	0.276

Table 3. Comparisons on the public dataset [36]. We provide quantitative comparison result of the methods in Table 1 on the public dataset [36].

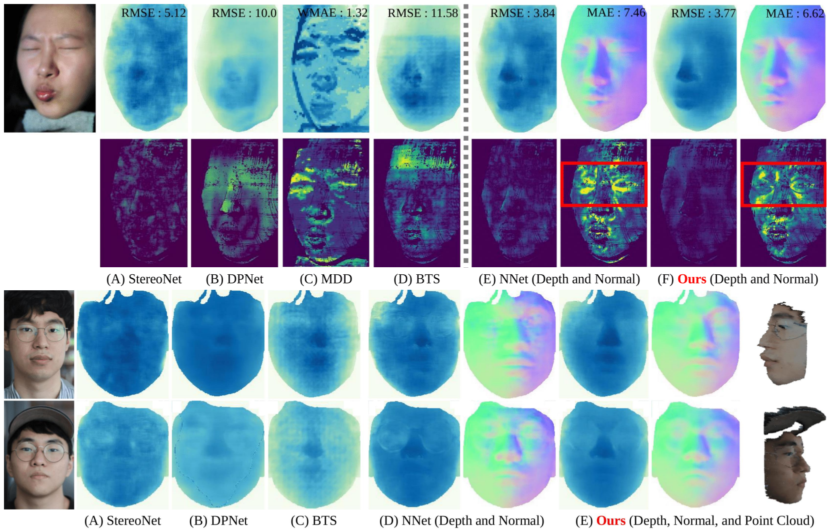


Fig. 7. Qualitative results. We show the qualitative results in the test set (**upper row**) and in the unmet real world (**lower row**). Compared to the other methods in Table 1. StereoDPNet clearly captures the surface and boundary depth of the face.

single DP images. Since the basic structure of ANM is derived from the recent depth and normal network [29] for multi-view stereo, we show the performance improvement of our ANM, compared to the baseline method [29] by adding each component in Table 2. We find that joint learning of disparity and surface normal leads to geometrically consistent and high-quality depth and surface normal, which has been demonstrated in previous works [38, 23].

4 Conclusion

We present a high-quality facial DP dataset incorporating 135,744 face images for 101 subjects with corresponding depth maps in metric scale and surface normal maps. Moreover, we introduce DP-oriented StereoDPNet for both depth and surface normal estimation. StereoDPNet successfully shows impressive results in the wild in Figure 8 by effectively handling the narrow baseline problem in DP.

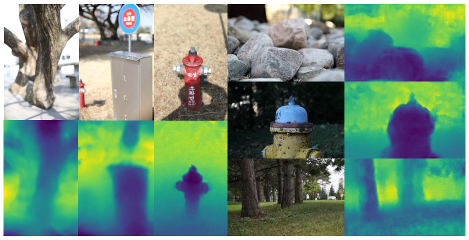


Fig. 8. Depth from single DP images in the wild. We show scene depth estimation results of StereoDPNet on outdoor photos, which are directly captured by us (**left col**) and in a public real-world DP dataset [2] for deblurring (**right col**).

Remarks. This paper is a re-publishing (summary presentation) of the paper which has been published in ”ECCV 2022” by request of the IW-FCV2023 program committee to share the research results.

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) pp. 1–16 (2016)
2. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: Proceedings of the European conference on computer vision (ECCV). pp. 111–126. Springer (2020)
3. Abuolaim, A., Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Learning to reduce defocus blur by realistically modeling dual-pixel data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2289–2298 (2021)
4. Apple: Apple iphone 11 pro. https://www.apple.com/iphone-11-pro/ (2019), accessed: 2019-09-20
5. ARCore: Augmented faces. https://developers.google.com/ar/develop/java/augmented-faces (2019), accessed: 2019-12-18
6. Bai, Z., Cui, Z., Rahim, J.A., Liu, X., Tan, P.: Deep facial non-rigid multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5850–5860 (2020)
7. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
8. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: CVPR (2022)
9. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5799–5809 (June 2021)
10. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
11. Chen, C.H., Zhou, H., Ahonen, T.: Blur-aware disparity estimation from defocus stereo images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 855–863 (2015)
12. Chen, W., Mirdehghan, P., Fidler, S., Kutulakos, K.N.: Auto-tuning structured light by optical stochastic gradient descent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
13. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European conference on computer vision (ECCV) (2018)
14. Galaxy: Samsung galaxy s10. https://www.samsung.com/us/mobile/galaxy-s10/ (2019), accessed: 2019-03-08
15. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

16. Google: Google photos: One year, 200 million users, and a whole lot of selfies. https://blog.google/products/photos/google-photos-one-year-200-million/ (2016), accessed: 2016-05-27
17. Google: More controls and transparency for your selfies. https://blog.google/outreach-initiatives/digital-wellbeing/more-controls-selfie-filters/ (2020), accessed: 2020-10-01
18. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Proceedings of the European conference on computer vision (ECCV). pp. 152–168. Springer (2020)
19. Ha, H., Oh, T.H., Kweon, I.S.: A multi-view structured-light system for highly accurate 3d modeling. In: International Conference on 3D Vision (3DV) (2015)
20. Ha, H., Park, J., Kweon, I.S.: Dense depth and albedo from a single-shot structured light. In: International Conference on 3D Vision (3DV). pp. 127–134 (2015)
21. Han, Y., Lee, J.Y., So Kweon, I.: High quality shape from a single rgb-d image under uncalibrated natural illumination. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2013)
22. Hu, P., Ramanan, D.: Finding tiny faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 951–959 (2017)
23. Im, S., Ha, H., Choe, G., Jeon, H.G., Joo, K., Kweon, I.S.: High quality structure from small motion for rolling shutter cameras. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
24. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
25. Jeon, H.G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.W., So Kweon, I.: Accurate depth map estimation from a lenslet light field camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
26. Keselman, L., Iselin Woodfill, J., Grunnet-Jepsen, A., Bhowmik, A.: Intel realsense stereoscopic depth cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2017)
27. Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 573–590 (2018)
28. Kinect2: Kinect for windows sdk 2.0. https://developer.microsoft.com/en-us/windows/kinect/ (2014), accessed: 2014-10-21
29. Kusupati, U., Cheng, S., Chen, R., Su, H.: Normal assisted stereo depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
30. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
31. Long, X., Lin, C., Liu, L., Li, W., Theobalt, C., Yang, R., Wang, W.: Adaptive surface normal constraint for depth estimation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
32. Long, X., Liu, L., Theobalt, C., Wang, W.: Occlusion-aware depth estimation with adaptive normal constraints. In: Proceedings of the European conference on computer vision (ECCV). pp. 640–657. Springer (2020)

33. Luo, H., Nagano, K., Kung, H.W., Xu, Q., Wang, Z., Wei, L., Hu, L., Li, H.: Normalized avatar synthesis using stylegan and perceptual refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11662–11672 (2021)
34. Nehab, D., Rusinkiewicz, S., Davis, J., Ramamoorthi, R.: Efficiently combining positions and normals for precise 3d geometry. ACM Transactions on Graphics (ToG) **24**(3), 536–543 (2005)
35. Pan, L., Chowdhury, S., Hartley, R., Liu, M., Zhang, H., Li, H.: Dual pixel exploration: Simultaneous depth estimation and image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4340–4349 (June 2021)
36. Punnappurath, A., Abuolaim, A., Afifi, M., Brown, M.S.: Modeling defocus-disparity in dual-pixel sensors. In: 2020 IEEE International Conference on Computational Photography (ICCP) (2020)
37. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 283–291 (2018)
38. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
39. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1259–1268 (2017)
40. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1 (2003)
41. Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 53–70. Springer (2020)
42. Shi, B., Wu, Z., Mo, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
43. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) - Workshop on 3D Representation and Recognition (2011)
44. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7346–7355 (2018)
45. Wadhwa, N., Garg, R., Jacobs, D.E., Feldman, B.E., Kanazawa, N., Carroll, R., Movshovitz-Attias, Y., Barron, J.T., Pritch, Y., Levoy, M.: Synthetic depth-of-field with a single-camera mobile phone. ACM Transactions on Graphics (ToG) **37**(4), 1–13 (2018)
46. Wu, S., Rupprecht, C., Vedaldi, A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

47. Wu, S., Rupprecht, C., Vedaldi, A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–10 (2020)
48. Wu, X., Zhou, J., Liu, J., Ni, F., Fan, H.: Single-shot face anti-spoofing for dual pixel camera. IEEE Transactions on Information Forensics and Security **16**, 1440–1451 (2020)
49. Xin, S., Wadhwa, N., Xue, T., Barron, J.T., Srinivasan, P.P., Chen, J., Gkioulekas, I., Garg, R.: Defocus map estimation and deblurring from a single dual-pixel image. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
50. Yang, F., Wang, J., Shechtman, E., Bourdev, L., Metaxas, D.: Expression flow for 3d-aware face component transfer. In: ACM SIGGRAPH 2011 papers, pp. 1–10 (2011)
51. Ying, X., Wang, L., Wang, Y., Sheng, W., An, W., Guo, Y.: Deformable 3d convolution for video super-resolution. IEEE Signal Processing Letters **27**, 1500–1504 (2020)
52. Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., Zhao, G.: Deep learning for face anti-spoofing: A survey. arXiv preprint arXiv:2106.14948 (2021)
53. Zhang, Y., Wadhwa, N., Orts-Escolano, S., Häne, C., Fanello, S., Garg, R.: Du 2 net: Learning depth estimation from dual-cameras and dual-pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 582–598. Springer (2020)
54. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)