Spatio-Focal Bidirectional Disparity Estimation from a Dual-Pixel Image

Donggun Kim Hyeonjoong Jang Inchul Kim Min H. Kim

KAIST

Abstract

Dual-pixel photography is monocular RGB-D photography with an ultra-high resolution, enabling many applications in computational photography. However, there are still several challenges to fully utilizing dual-pixel photography. Unlike the conventional stereo pair, the dual pixel exhibits a bidirectional disparity that includes positive and negative values, depending on the focus plane depth in an image. Furthermore, capturing a wide range of dual-pixel disparity requires a shallow depth of field, resulting in a severely blurred image, degrading depth estimation performance. Recently, several data-driven approaches have been proposed to mitigate these two challenges. However, due to the lack of the ground-truth dataset of the dual-pixel disparity, existing data-driven methods estimate either inverse depth or blurriness map. In this work, we propose a self-supervised learning method that learns bidirectional disparity by utilizing the nature of anisotropic blur kernels in dual-pixel photography. We observe that the dual-pixel left/right images have reflective-symmetric anisotropic kernels, so their sum is equivalent to that of a conventional image. We take a self-supervised training approach with the novel kernel-split symmetry loss accounting for the phenomenon. Our method does not rely on a training dataset of dual-pixel disparity that does not exist yet. Our method can estimate a complete disparity map with respect to the focus-plane depth from a dual-pixel image, outperforming the baseline dual-pixel methods.

1. Introduction

Dual-pixel is an image sensor technology that a single pixel has two photodiodes, while a pixel on the traditional image sensor has only a single photodiode. The dual-pixel is originally invented to leverage the phase difference of two photodiodes for efficient autofocusing [1, 14, 27]. Nowadays, dual-pixel image sensors can be easily found in multiple camera platforms such as Canon EOS 5D Mark IV DSLR and Google Pixel phone cameras. The dual-pixel cameras make possible not only traditional autofocusing but also a wide range of interesting applications of RGB-D computational photography, such as depth estimation [9], refocusing [29], deblurring [2, 34] or reflection removal [24]. The captured dual-pixel image can be separated into two images by gathering pixels from the left and ride sides of photodiodes, respectively. Although the baseline between dual-pixel photodiodes is short, these two images capture slightly different light fields, producing disparity by parallax.

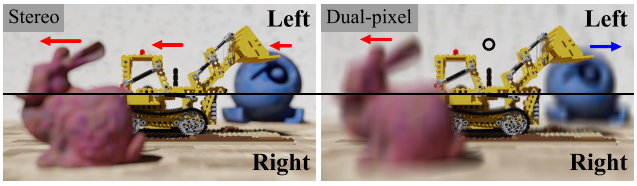


Figure 1. Captured images from the traditional binocular stereo camera and dual-pixel camera, with a focus-plane depth at the middle object. Upper part of the image is a left image, and lower part is a right image from each camera. Arrows note parallax, which is unidirectional in stereo imaging and bidirectional in dual-pixel imaging.

One characteristic of the dual-pixel setup compared with a traditional binocular stereo is that it yields a *bidirectional* disparity shown in Figure 1, depending on the *focus-plane depth* that is the distance between a camera lens and the perfect point of focus in an image (Figure 2). The dual-pixel disparity estimation should account for the direction of horizontal shift changes by the focus-plane depth. In addition, the magnitude of a dual-pixel disparity is proportional to the size of a circle of confusion. To guarantee enough disparity in dual-pixel imaging, optics with a large aperture are required to obtain a shallow depth of field for a large circle of confusion, which is the area around the plane of focus that appears to be in focus. As a result, the disparity is estimated from a pair of blurry dual-pixel images caused by the shallow depth of field, which raises another challenge. The conventional camera has an *isotropic* blur kernel when an image is out of focus (Figure 2). However, a dual-pixel camera produces *anisotropic* blur kernels, i.e., the kernel shapes of the left/right photodiodes are anisotropic and symmetrically

flipped [3, 23]. It is impossible to learn these kernels explicitly in a supervised manner since no training dataset with bidirectional disparity is available.

To address those challenges, several data-driven supervised learning-based methods have been proposed [9, 21]. However, the main challenge in this supervised approach is that it is difficult to create a ground-truth disparity dataset specifically designed for a dual-pixel system unless it is produced by simulating complete light transport through camera optics with the variation of focus-plane depths in each image. Up to date, there is no bidirectional disparity dataset available for supervised learning. Due to this, these methods alternatively estimate either *inverse depth* [9, 21], or *blurriness maps* [34] only, which cannot directly estimate the disparity of a dual-pixel image without focus-plane depth information. And thus, only traditional optimization-based approaches [23, 29] have handled the bidirectional disparity estimation problem in dual-pixel photography. No learning-based approach tackles the bidirectional disparity problem in dual-pixel photography.

In this work, we propose a learning-based bidirectional disparity estimation method specially tailored for dual-pixel imaging. Our contribution consists of two main parts: (1) Assuming an isotropic blur kernel, we first pretrain a conventional stereo network using a stereo dataset with left-right inverted. (2) Based on our observation that the dual-pixel left/right images have *reflective-symmetric* anisotropic kernels that their sum has to be equivalent to that of a conventional image, we employ our novel self-supervised training with the novel kernel-split symmetry loss accounting for the phenomenon. In addition, we demonstrate that our model represents more accurate 3D geometric relationships among objects with comprehensive evaluations.

2. Related Work

Dual-pixel depth estimation. Since dual-pixel cameras have been released, many works have been proposed to estimate depth from a dual-pixel image. Garg *et al*. [9] present a learning-based dual-pixel depth estimation using the affine invariant objective to estimate inverse depth. Zhang *et al*. [38] introduce a supervised learning method similar to [9] to estimate disparity by using two dual-pixel cameras. Recently, Pan *et al*. [21] propose a dual-pixel simulator and also presented the learning-based inverse depth estimation method, which is trained with their simulated data. Xin *et al*. [34] present an unsupervised optimization method based on the estimated defocus map, which can also be shown as an inverse depth map in dual-pixel data. These methods estimate unidirectional information of inverse depth maps [9, 21, 34, 38] only, often assuming that the focus plane is fixed to the nearest or the farthest location of the scene [34].

Punnappurath *et al*. [23] show that the sign of a bidirectional dual-pixel disparity changes depending on the focus-plane depth in an image. Also, Wadhwa *et al*. [29] apply traditional stereo matching on the separated stereo images from a dual-pixel image to estimate bidirectional disparity. These two traditional methods [23, 29] discover the bidirectional nature of dual-pixel disparity and anisotropic blur kernels. However, their performance often degrades due to severe defocus blur of a shallow depth of field and the affine ambiguity of bidirectional disparity w.r.t. focus-plane depth. A learning-based approach would be impactful in mitigating these challenges as the problem is severely ill-posed. However, as mentioned earlier, there is no dual-pixel dataset, including pairs of dual-pixel photographs and bidirectional disparities for every focus-plane depth.

In contrast, our self-supervised method estimates bidirectional disparities while implicitly imposing the reflective-symmetry constraint in anisotropic kernels.

Learning-based stereo. Learning-based stereo methods can be divided into two groups. First, *supervised* learning-based approaches [7, 16, 17, 36, 37] have been proposed that train neural networks with traditional stereo image datasets [18, 20, 25, 26]. These network architectures resemble many aspects of the traditional stereo algorithms, for instance, searching correspondences between two rectified images in a coarse-to-fine manner using an image pyramid. However, it is challenging to create a supervision dataset with ground-truth depth labels in the real world. Available datasets are still insufficient to cover the variety of daily stereo-imaging scenarios.

The other group of approaches is *self-supervised*-based to address the limitation of acquiring the true dense depth labels in stereo depth estimation [11, 13, 15, 30, 35, 39]. The key advantages of the self-supervised approach are that it does not require any ground-truth depth labels for learning and that once a self-supervised network is pretrained with a large number of observation samples in advance, the network can infer depth at a faster speed than the traditional binocular stereo methods [8, 12].

As mentioned earlier, there is no publicly available dual-pixel dataset with the ground-truth bidirectional disparity labels with a wide range of variation of focus-plane depths. We therefore adopt this self-supervised learning scheme for the dual-pixel disparity problem. Note that the traditional learning-based stereo scheme is not directly applicable to dual-pixel photography since the sign of a bidirectional disparity should change according to an arbitrary focus-plane depth, and the blur kernels should be anisotropic for left/right photodiodes, respectively. To the best of our knowledge, this is the first learning-based solution that can estimate a bidirectional disparity w.r.t. an arbitrary focus-plane depth in dual-pixel photography.

3. Disparity of a Dual-Pixel Image

In the conventional stereo setup with two rectified pinhole cameras, the left and right image relationship at a scene point (x, y) is given as $I^l(x, y) = I^r(x - d, y)$. The disparity d can be computed as $d = fB/z$, where f is a focal length and B is a baseline, and z is the depth at (x, y) in the left image. Since the disparity is proportional to inverse depth, it has a zero value at the infinite depth, and the disparity is unidirectional, as shown in the left image in Figure 1.

However, unlike the stereo setup, the disparity in a dual-pixel image pair can be represented as: $d = \alpha b$, where d is the dual-pixel disparity, α is the ambiguous positive scale coefficient and b is the circle of confusion size. Based on thin-lens optics, the circle of confusion size can be formulated as $b = \frac{Af}{1-f/z_f}\left(\frac{1}{z_f} - \frac{1}{z}\right)$, where A is the size of the aperture and z_f is the depth of the focus plane. Note that there is no absolute operator, and the b value can have a negative sign, unlike conventional imaging, which takes an absolute value to get the circle of confusion size. Lastly, the bidirectional disparity can be formulated as follows:

$$d = \alpha b = \alpha \frac{Af}{1 - f/z_f}\left(\frac{1}{z_f} - \frac{1}{z}\right) \equiv \theta_0 + \theta_1 \frac{1}{z}. \quad (1)$$

With two coefficients θ_0 and θ_1, the dual-pixel disparity d and the inverse depth $1/z$ are in an *affine* relation, whereas in the conventional stereo setup, they are just *proportional*. We summarize the key difference between disparity in conventional stereo and disparity in dual-pixel stereo from this relationship in the following sections.

3.1. Disparity-Blurriness Trade-off

The dual-pixel disparity proportionally increases with the size of the circle of confusion (Equation (1)). If we set the camera to have an ultra-wide depth of field and capture an all-in-focus image ($b \to 0$), i.e., this is the case of the Dirac-delta point spread function, the separated left and right images are exactly identical, and there is no disparity ($d \to 0$). Therefore, if we want to estimate disparity from a single dual-pixel image, the image must be captured with *defocus blur*. This trade-off increases the ill-posedness of the disparity estimation problem as the defocus blur and the range of disparities are closely related, so one can only get a small range of disparity [29] or degraded disparity results.

3.2. Direction of Disparity

Based on Equation (1), the disparity in dual-pixel can have either positive or negative signs depending on focus plane depth. This occurs when separating the PSFs from left and right of dual-pixel images [2, 3, 23]. Figure 2 describes this phenomenon. The point source P_2 is imaged as a point on the sensor, and P_1 and P_3 are blurred. If a scene point is at the focus plane (P_2), there is no disparity since both left and right blur kernels are identical. On the other hand, if the point is behind the focus plane depth (P_3), then the rays are inverted, and the disparity occurs in the left-to-right direction, while if the point is closer to the lens than the focus plane (P_1), it is the other way around. As a result, the focus plane depth comes into play as another variable in estimating dual-pixel disparities. Therefore, in this setup, both directions of disparity should be considered if the focus-plane position is unknown, whereas the traditional stereo searches only one direction along the scan line.

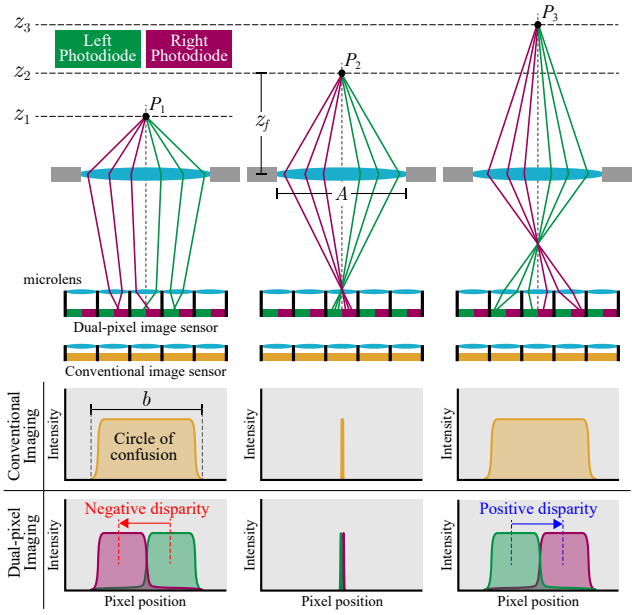


Figure 2. A pictorial description of the dual-pixel setup with a thin lens. The point sources P_1 and P_3 are out of the focus plane, so they are imaged on the dual pixel sensor forming a circle of confusion (CoC) with a direction. Note that all the point sources are placed on the optical axis of the lens. P_2 is in-focus, so it does not create a CoC. The conventional sensor collects rays from all paths, so the CoC in any case is *isotropic* (yellow plot). On the other hand, the dual-pixel sensor gathers rays coming to left and right diodes separately, so the CoCs are *anisotropic* (magenta and green plots). We show two interesting phenomena from this observation with the dual-pixel setup: The shape of CoCs changes according to the focus plane depth, and the summation of the two anisotropic CoC equals to the isotropic counterpart [3, 23].

3.3. Blur-Kernel Ambiguity in Disparity

In Equation (1), the parameter α cannot be obtained directly, while A, f, and z_f can be directly set or can be obtained easily by a traditional camera calibration. The α value is closely related to the shape of PSF, which has wide variation, depending on the camera and lens camera architectures [5]. Hence, obtaining the ground-truth dual-pixel disparity is technically challenging, and also it depends on the imaging hardware. Recent learning-based dual-pixel

stereo methods can estimate *inverse depth* only rather than the *bidirectional* disparity due to this difficulty. Our method accounts for this blur kernel ambiguity by employing a network.

4. Dual-pixel Disparity Estimation

Since we aim to estimate dual-pixel disparity without any PSF calibration nor a specific focus plane setup, we adopt a learning-based binocular stereo method [17] as our backbone network to make it predict disparity from any given dual-pixel image input. Given the binocular stereo network $\mathcal{N}$ and dual-pixel left and right image I^l and I^r, the *left-referenced* disparity map $\hat{D}^{l\to r}$ representing the horizontal shift from the left to the right image is estimated as

$$\hat{D}^{l\to r} = \mathcal{N}(I^l, I^r). \tag{2}$$

By reversing the input image order $\mathcal{N}(I^r, I^l)$, we can obtain another disparity map of the right image $\hat{D}^{r\to l}$, the right-referenced disparity.

However, we found that recent binocular stereo networks [7, 17] fail to estimate bidirectional disparity and are not directly applicable to dual-pixel stereo.

To this end, we build our disparity network upon the employed stereo model based on the optical characteristic of the dual-pixel setup. Our method consists of two stages: We first pretrain our network with the stereo dataset in a supervised way, and then our novel self-supervised training that accounts for dual-pixel physics follows.

4.1. Pretraining of Directional Disparity

Since there is no dual-pixel dataset with bidirectional disparity, we first use a set of popular binocular stereo datasets [6, 18, 28, 31] to pretrain our network. The training data is unidirectional, so to account for the bidirectional nature of dual-pixel disparity, we randomly flip the ground truth data with probability 0.5 in the data augmentation. This simple step generates disparities in the other direction, and although this can only create *bidirectional* disparity maps but with only one direction per each map, we found that this is enough to estimate bidirectional disparity for one pair of images. Along with this, we use the baseline training using the sequential L1 loss and data augmentation such as spatial scaling, crop, color jitter, and eraser transform [17]. Lastly, we change the batch normalization in the baseline network to instance normalization, which is more applicable for small batch size [33] since the next self-supervised step uses the smaller batch size.

Note that we do not use synthetic bidirectional disparity generation method using plane induced homography with RGB-D image [4] nor render an image by modifying the center of projection, since it creates artificial disparities and lacks the blur effect. We discuss this in the supplemental material.

4.2. Self-supervised Learning of Disparity

In the pretraining stage, we have trained our network to learn bidirectional disparity with the existing stereo datasets; however, since the stereo dataset does not obey the optics of dual-pixel photography with anisotropic PSFs, we recover a physically-plausible dual-pixel disparity that preserves the *reflective symmetry of kernels split to left/right*, implicitly accounting for the dual-pixel optics.

Utilizing the reflective symmetry. As shown in Figure 2, the blur kernel of the conventional imaging is isotropic. In dual-pixel photography, it is split into two anisotropic kernels, which are *reflective symmetry*, i.e., the summation of the left/right blur kernel is equal to an isotropic kernel. Refer to the supplemental for details of an experiment we conduct that shows this phenomenon. As a result, we obtain the anisotropic dual-pixel kernels shown in Figure 3(a). In our self-supervised training, we employ a center image I^c, the combined charges of the left and right diodes. In practice, we estimate I^c by computing the mean of left and right images $I^c = \left(I^l + I^r\right)/2$, as shown in Figure 3(b). I^c plays a role as a proxy geometry between the left and right images. With this, we implicitly impose the anisotropic nature of dual-pixel kernels by computing a detour disparity $D^{l\to c\to r} = -D^{c\to l} + D^{c\to r}$ instead of the direct one $D^{l\to r}$, where $D^{c\to l}$ is the disparity from the (I^c, I^l) pair and $D^{c\to r}$ is the disparity from the (I^c, I^r) pair. Note that in the following description, the losses are of a pair (left and right), and we only show the left ones for the sake of brevity. Our convention that comes in handy to know is that $\hat{D}$ is an estimated disparity by our trained network.

Kernel-split symmetry loss. As stated above, we estimate two disparity maps with our network: $\hat{D}^{c\to l}, \hat{D}^{c\to r}$. Since the kernel is evenly split from an isotropic one [23, 29], the disparity maps $\hat{D}^{c\to l}$ and $\hat{D}^{c\to r}$ should have the same absolute value while their signs are the opposite. To pursue this property, we define our kernel-split symmetry term as

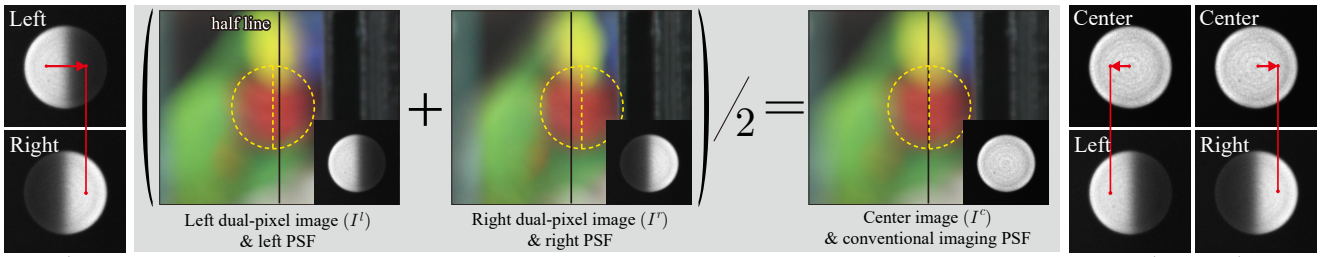
$$\mathcal{L}_{\text{kernel}} = \frac{1}{N}\sum L_\beta(\hat{D}^{c\to l} + \hat{D}^{c\to r}), \tag{3}$$

where N is the number of pixels in an image and L_β is a robust Huber loss [10]. Note that this term is by definition bidirectional, so we do not compute left and right losses separately like the other terms.

Photometric loss. The photometric loss measures the projection error by comparing pixel values of a left image I^l and a warped image using a right image and its inferred disparity $\mathcal{W}(I^r, \hat{D}^{l\to c\to r})$:

$$\mathcal{L}^l_{\text{photo}} = \frac{1}{N}\sum \psi\left(I^l, \mathcal{W}(I^r, \hat{D}^{l\to c\to r})\right), \tag{4}$$

where $\mathcal{W}(I, D)$ is an warping operator that takes an image I and disparity map D to generate a forward warped



(a) $\hat{D}^{l\to r}$ (b) Creating the *center* image (c) $\hat{D}^{c\to l}$ and $\hat{D}^{c\to r}$

Figure 3. An illustrative example showing the kernel-split symmetry in Section 4.2. We experiment to obtain the real dual-pixel PSF kernels of the Canon 5D Mark IV camera (refer to the supplemental for more detail). (a) Measured left and right PSFs of the Canon dual-pixel camera. (b) The dual-pixel center image I^c with a full blur kernel is computed by averaging the left and right dual-pixel image pair. (c) Using the center image I^c as a proxy, we estimate $\hat{D}^{l\to c\to r} = -\hat{D}^{c\to l} + \hat{D}^{c\to r}$. The constraint that the summation of the left and right PSFs is equivalent to the center PSF is implicitly imposed.

image as $\mathcal{W}(I, D) = I(\mathbf{x} + D(\mathbf{x}))$, and $\mathbf{x}$ is a pixel position (x, y). The $\psi(\cdot,\cdot)$ measures the difference between two image inputs, using the soft hamming distance of the census-transformed images combined with the robust generalized Charbonnier penalty function [19] so we can mitigate an outlier influence.

Smoothness loss. To smooth out disparity maps while preserving sharp edge-discontinuity, we use the bilateral gradient smoothness loss [11, 32] defined as

$$\mathcal{L}^l_{\text{smooth}} = \frac{1}{N}\sum \Big(\exp\left(-\lambda\left|\nabla_x I^l\right|\right) \odot \left|\nabla_x \hat{D}^{l\to c\to r}\right| + \exp\left(-\lambda\left|\nabla_y I^l\right|\right) \odot \left|\nabla_y \hat{D}^{l\to c\to r}\right|\Big), \quad (5)$$

where $\odot$ is an element-wise multiplication, and λ is a parameter controlling the power of edge-awareness in the gradient domain. The gradient of an image is computed by averaging the gradient values of each color channel. We use $\lambda = 5$ for all experiments.

Finally, we train the network with the weighted sum of all the losses:

$$\mathcal{L} = w_{\text{kernel}}\mathcal{L}_{\text{kernel}} + w_{\text{photo}}\mathcal{L}_{\text{photo}} + w_{\text{smooth}}\mathcal{L}_{\text{smooth}}. \quad (6)$$

We use $w_{\text{kernel}} = 1.0$, $w_{\text{photo}} = 1.0$, and $w_{\text{smooth}} = 0.01$ for all experiments.

5. Experiments and Results

5.1. Implementation Details

We implement our method using PyTorch [22]. We train our 2-stage method with different schemes. In the pretraining stage (Section 4.1), we perform training for 40k iterations with a batch size of 40. We used a mix of multiple synthetic stereo datasets, SceneFlow [18], Sintel stereo [6], Falling Things [28] and Tartan Air [31]. In the second stage, we train our model for 1.5k iterations with a batch size of 8. Since the second stage is self-supervised, we only need dual-pixel left and right image pairs for training. We use DSLR dual-pixel images from Abuolaim *et al*. [2]. Refer to the supplemental document for more details.

Table 1. Bidirectional disparity evaluation by RMSE, comparison to the dual-pixel method that estimates bidirectional disparity. A lower number is better. The best scores are highlighted in bold.

	Dataset by [23]	Dataset by [2]
Wadhwa *et al*. [29]	4.1113	6.7384
Ours	**3.7943**	**5.5075**

5.2. Bidirectional Evaluation

We evaluate our *bidirectional* disparity estimation results. The quantitative results are shown in Table 1, and the qualitative comparison is presented in Figure 4.

Quantitative comparison. Since there is no dataset with ground truth bidirectional disparity, we measure projection error by computing the root mean square error (RMSE) of the photometric difference between a reference image and a warped image as $\sqrt{\frac{1}{N}\sum\left(I^l - \mathcal{W}(I^r, \hat{D})\right)^2}$. We use the dataset provided by Punnappurath *et al*. [23] and Abuolaim *et al*. [2], which provides DSLR dual-pixel image pairs. Table 1 shows the result.

We quantitatively compare the existing bidirectional disparity estimation method by Wadhwa *et al*. [29]. Since Punnappurath *et al*. [23] do not estimate disparities but rather signed kernel sizes, we do not include their method for this comparison. Our method shows a strong performance over other methods in this metric.

Qualitative comparison. We present the qualitative comparison in Figure 4. The optimization-based methods [23, 29] produce plenty of artifacts across the disparity maps since the problem is severely ill-posed. In contrast, our network shows consistent bidirectional disparities with fewer artifacts.

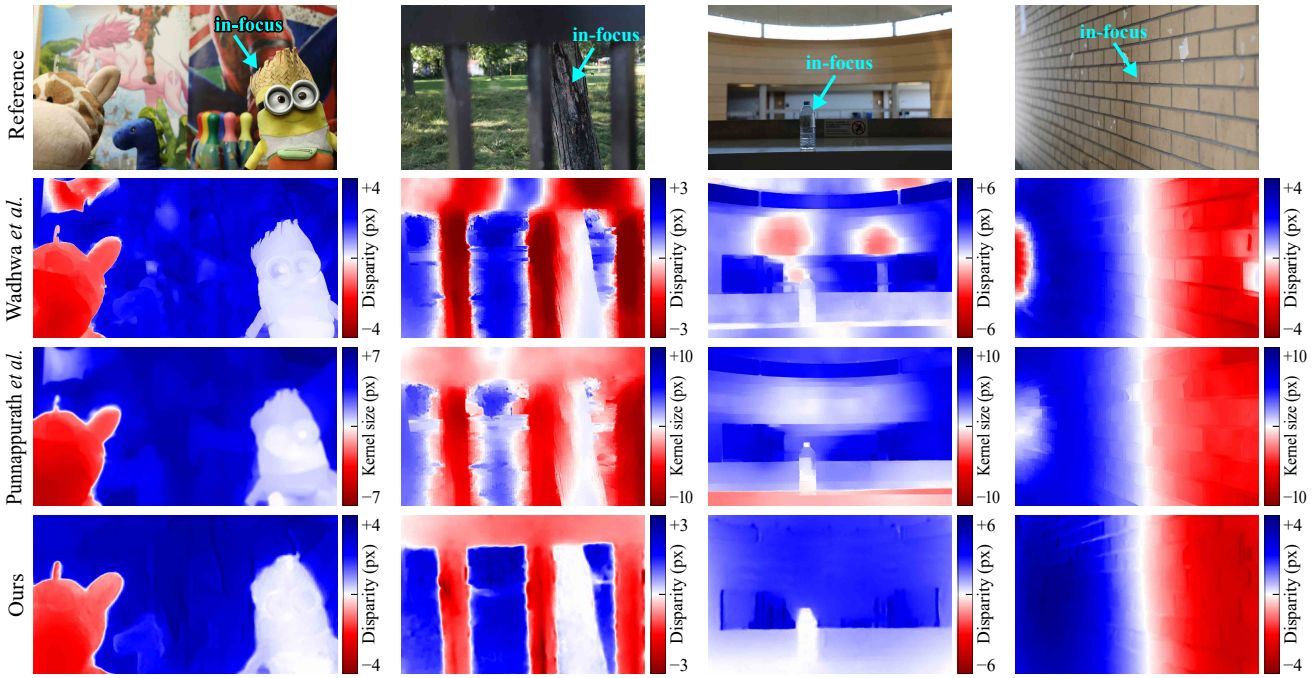


Figure 4. Bidirectional disparity estimation results. Note that our method and Wadhwa *et al*. [29] estimate the disparity in pixel, and Punnappurath *et al*. [23] estimates the signed kernel size. The first column image is from [23], and the other column images are from [2]. Refer to the supplemental material for more disparity results.

Figure 5. Bidirectional disparity estimation by varying the focus plane depth. All three images are captured with the same setup, *i.e.,* the same depth, but with different focus plane positions. Note that the white values in the second row show zero disparity; the objects in focus have zero shift, and the near and far objects that are out-of-focus show negative/positive disparities, respectively.

Focally-varying bidirectional disparity. We provide the results of bidirectional disparity maps by varying the focus plane depth in the same scene in Figure 5. As we move the focus position from near to far, our method successfully estimates that the in-focus objects have zero disparity (white in the second row), and the near/far objects have negative/positive disparities, respectively.

5.3. Unidirectional Evaluation

Although our estimation results are bidirectional, we can also evaluate these results by following the commonly used dual-pixel *inverse depth* evaluation methods [9, 21, 23, 34].

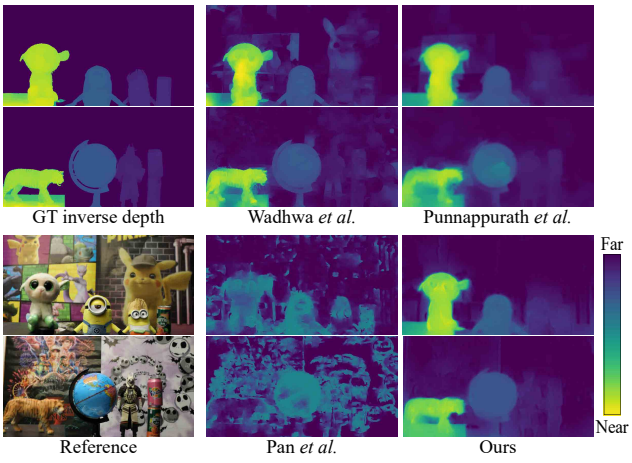
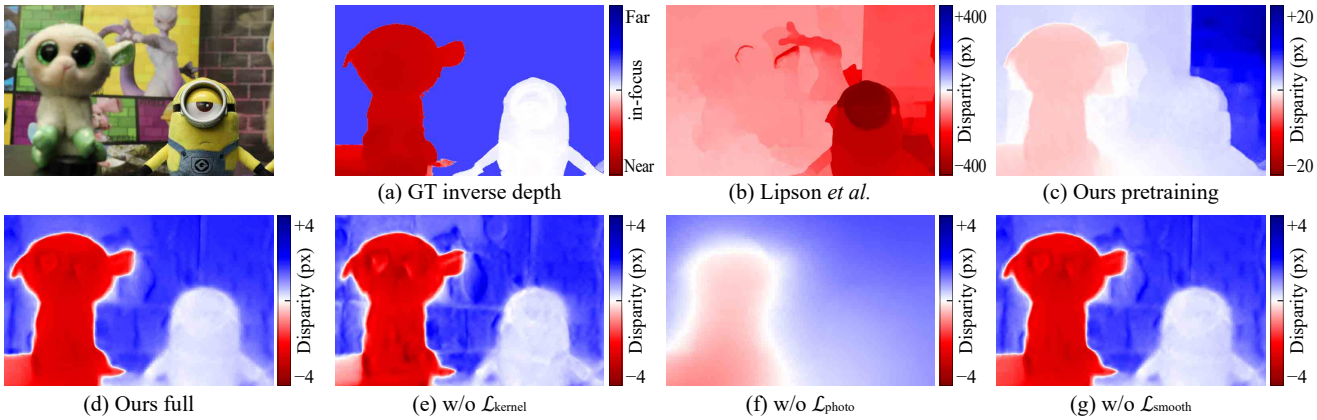


Figure 6. Inverse depth estimation result using dataset from [23]. The estimated inverse depths have been affine transformed using the coefficient from AI(2), then colormapped. Our method is consistent at homogeneous depth regions and has relatively better edge quality. Refer to the supplementary material for more results.

The inverse depth is evaluated by the affine invariant version of mean absolute error (MAE) and root mean squared error (RMSE). The affine invariant metric $\mathrm{AI}(p)$ is defined as:

$$\mathrm{AI}(p) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \left(\frac{1}{N} \sum \left| D - \left(\beta_0 + \beta_1 \hat{D} \right) \right|^p \right)^{1/p}, \quad (7)$$



(a) GT inverse depth (b) Lipson *et al.* (c) Ours pretraining

(d) Ours full (e) w/o $\mathcal{L}_{\text{kernel}}$ (f) w/o $\mathcal{L}_{\text{photo}}$ (g) w/o $\mathcal{L}_{\text{smooth}}$

Figure 7. Qualitative ablation studies. (a) Ground truth inverse depth. Color mapped so that the in-focus object has white color. (b) Baseline model [17] using author-provided network weights. The baseline model fails to estimate the positive disparity region. (c) Our pretrained model. It can estimate bidirectional disparity. (d) Trained after our self-supervised step using all the loss functions. (e)–(g) Results by excluding one loss term each in Equation (6). We present the quantitative ablation results in Table 3.

where D is the ground truth inverse depth map, $\hat{D}$ is the estimated inverse depth map, and β_0 and β_1 are the optimized affine coefficients. And changing p provides the error selection, either MAE ($p = 1$) or RMSE ($p = 2$). The optimization for computing AI(1) is computed by the iteratively re-weighted least squares (IRLS) with 5 iterations, and AI(2) is computed with a single iteration since it has deterministic optimization. We also report Spearman's rank correlation coefficient ρ_s.

We compare our method with three dual-pixel depth estimation methods. For Punnappurath *et al*. [23] and Pan *et al*. [21], we directly run their publicly released source code. Since Wadhwa *et al*. [29] has no released source code, we implement their method following their descriptions in their work. We again use the dataset by Punnappurath *et al*. [23], which also provides the ground-truth inverse depth for testing.

As a result, the evaluation results are shown in Table 2, and the inverse depth estimation result is presented in Figure 6. Our method shows the best score on all evaluation metrics among the compared methods. And again, our result shows higher consistency and uniformity than other methods, which reveals our learning-based method outperforms traditional patch-matching-based methods.

5.4. Ablation Study

We first describe the effectiveness of our pretraining. As shown in Figure 7(c), the naïve baseline [17] fails to estimate the positive disparity. On the other hand, our pretraining (Figure 7d) enables us to estimate bidirectional disparity.

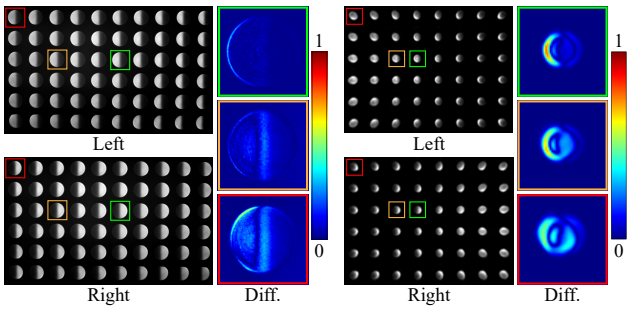
We also show the effect of our multiple loss in self-supervised learning. For this ablation study, we begin with the pretrained stereo network that already learns to produce bidirectional disparity. We conduct self-supervised learning with combinations of different loss functions with the same hyperparameters and a fixed number of iterations. The setup and the quantitative results are shown in Table 3 and estimated disparity map is shown in Figure 7(d)-(g). We use the same dataset and metrics for all ablation experiments. As a result, we discover that our self-supervised learning method

Table 2. Unidirectional evaluation result with comparison to the other dual-pixel depth estimation methods. The last column is the geometric mean of three evaluation metrics. A lower number is better for all AI(1), AI(2), and $1 - |\rho_s|$. The best scores are highlighted in bold.

	AI(1)↓	AI(2)↓	$1 - \lvert\rho_s\rvert$↓
Wadhwa *et al*. [29]	0.0463	0.0740	0.2898
Punnappurath *et al*. [23]	0.0449	0.0724	0.2703
Pan *et al*. [21]	0.0894	0.1491	0.5152
Ours	**0.0391**	**0.0682**	**0.2619**

Table 3. Quantitative ablation study results. We use the same metrics used for quantitative evaluation (Section 5.3). We conduct experiments by turning off each loss component in Equation (6). The best scores are highlighted in bold. Note that all the experiments have been done with pretraining. The best scores are highlighted in bold.

$\mathcal{L}_{\text{kernel}}$	$\mathcal{L}_{\text{photo}}$	$\mathcal{L}_{\text{smooth}}$	AI(1)↓	AI(2)↓	$1-\lvert\rho_s\rvert$↓
–	–	–	0.0826	0.1301	0.4286
✓	–	–	0.0660	0.1011	0.3057
–	✓	–	0.0656	0.1051	0.3156
–	–	✓	0.1122	0.1798	0.7349
✓	✓	–	0.0398	0.0685	0.2640
✓	–	✓	0.0689	0.1047	0.3113
–	✓	✓	0.0502	0.0787	0.2818
✓	✓	✓	**0.0391**	**0.0682**	**0.2619**



(a) Canon EOS 5D Mark IV DSLR (b) Google Pixel 4 smartphone

Figure 8. The first and third columns show spatially-varying blur kernels of two different dual-pixel cameras: Canon EOS 5D (Mark IV) DSLR camera with Sigma Art 50 mm $f/1.4$ lens (our measurement) and Google Pixel 4 camera. The second and fourth columns present the absolute difference maps between the sampled left kernels and the horizontally-inverted right kernels. If they are perfectly symmetric, the different maps should include only zeros (blue). Refer to the supplemental for more details of this PSF experiment setup. Our DSLR measurement exhibits consistent symmetric PSFs across the image, whereas the Pixel shows apparent inconsistency, particularly at the corners.

with all four loss functions improves the performance from pretraining.

As shown in Table 3 and Figure 7, the photometric loss is essential to perform self-supervised learning. Our kernel-split symmetry loss reduces the artifact significantly over homogeneous depth regions. Smoothness loss also quantitatively improves results along with the photometric loss, but the impact of the smooth term is less powerful than the other two terms.

6. Discussion

Occlusion. Since dual-pixel cameras produce a dual-pixel image pair with a very small horizontal shift, so the age-long occlusion problem in stereo disparity estimation does not appear significantly in the dual-pixel setup. We show some example occlusion maps in the supplemental material.

DSLR vs. Google Pixel dataset. Our bidirectional disparity estimation method relies on the kernel symmetry (Section 4.2), so the quality of blur kernels on a dual-pixel image is crucial. Our method is trained and tested using data captured by a high-end DSLR camera, Canon 5D Mark IV, so that we can obtain high-quality images with minimal optical aberrations. We compare the PSF of the DSLR camera with that of Google Pixel 4 provided by Xin *et al*. [34] (Figure 8). The result shows that the DSLR creates reflection-symmetric kernels; on the other hand, the Pixel camera exhibits asymmetric ones. We argue that the asymmetry comes from optical aberrations of the smartphone camera's optics. In addition, the Pixel dataset suffers from severe pincushion distortion (see the corners in Figure 8(b)). For this reason, we do not include this dataset in training our network. Additionally, we conduct an experiment with the Pixel dataset and our method outperforms uncalibrated methods. We refer to these results in the supplemental.

Pre-training. There are two main purposes for using the pre-training step. First, pre-training helps us keep our self-supervised learning process tractable by narrowing down the optimization space close to the optimal solution. It enhances the speed of self-supervised training. Second, the size of the dual-pixel training dataset is relatively small. The pre-training step prevents an overfitting problem in self-supervised training. We confirm that when we modify hyperparameters, such as learning rates and iterations, only self-supervised learning can estimate disparity at an excessively slow speed.

Benefits of Estimating Bidirectional Disparity. The main benefit of estimating bidirectional disparity is that a zero value in the disparity map indicates the focus plane's depth in an image, which is available only from metadata in conventional photography. Also, the focus plane location indicates where the blurred bokeh is flipped about the focal plane. The explicit information of the focus plane location in an image will be highly beneficial for a computational photography application such as high-fidelity synthetic blurring with occlusion. In addition, our method estimates bidirectional disparity from a blurry dual-pixel image. By using the disparity-blur relationship, the estimated disparity can be used to deblur dual-pixel images. We leave it as future work.

7. Conclusion

We present a novel self-supervised dual-pixel bidirectional disparity estimation method. Without relying on any supervision of dual-pixel dataset, our method can estimate bidirectional disparity from a given dual-pixel pair image with high accuracy. Our method shows strong performance compared with existing dual-pixel disparity estimation methods, and we also present that our method is physically-plausible by showing the change of disparities' sign according to the focus plane position.

Acknowledgements

Min H. Kim acknowledges the MSIT/IITP of Korea (RS-2022-00155620, 2022-0-00058, and 2017-0-00072) and the Samsung Research Funding Center (SRFC-IT2001-04) for developing partial 3D imaging algorithms, in addition to the support of the NIRCH of Korea (2021A02P02-001), Samsung Electronics, and Microsoft Research Asia.

References

[1] Abdullah Abuolaim and Michael Brown. Online lens motion smoothing for video autofocus. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1

[2] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 5, 6

[3] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[4] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3D: Stereo depth estimation via binary classifications. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[5] Seung-Hwan Baek, Hayato Ikoma, Daniel S. Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H. Kim. Single-shot hyperspectral-depth imaging with learned diffractive optics. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, 2012. 4, 5

[7] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4

[8] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*. Springer, 2003. 2

[9] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T. Barron. Learning single camera depth estimation using dual-pixels. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 6

[10] Ross Girshick. Fast R-CNN. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 5

[12] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 2008. 2

[13] Baoru Huang, Jian-Qing Zheng, Stamatia Giannarou, and Daniel S Elson. H-net: Unsupervised attention-based stereo depth estimation leveraging epipolar geometry. *arXiv preprint arXiv:2104.11288*, 2021. 2

[14] Jinbeum Jang, Yoonjong Yoo, Jongheon Kim, and Joonki Paik. Sensor-based auto-focusing system using multi-scale feature extraction and phase correlation matching. *Sensors*, 15(3), 2015. 1

[15] Sunghun Joung, Seungryong Kim, Bumsub Ham, and Kwanghoon Sohn. Unsupervised stereo matching using correspondence consistency. In *International Conference on Image Processing (ICIP)*, 2017. 2

[16] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[17] L. Lipson, Z. Teed, and J. Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, 2021. 2, 4, 7

[18] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4, 5

[19] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *Conference on Artificial Intelligence (AAAI)*, 2018. 5

[20] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[21] Liyuan Pan, Shah Chowdhury, Richard Hartley, Miaomiao Liu, Hongguang Zhang, and Hongdong Li. Dual pixel exploration: Simultaneous depth estimation and image restoration. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6, 7

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5

[23] Abhijith Punnappurath, Abdullah Abuolaim, Mahmoud Afifi, and Michael S. Brown. Modeling defocus-disparity in dual-pixel sensors. In *International Conference on Computational Photography (ICCP)*, 2020. 2, 3, 4, 5, 6, 7

[24] Abhijith Punnappurath and Michael S Brown. Reflection removal using a dual-pixel sensor. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[25] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*. Springer, 2014. 2

[26] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[27] Przemysław Śliwiński and Paweł Wachel. A simple model for on-sensor phase-detection autofocusing algorithm. *Journal of Computer and Communications*, 1(06), 2013. 1

[28] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3D object detection and pose estimation. In *CVPR Workshops*, 2018. 4, 5

[29] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics*, 37(4), 2018. 1, 2, 3, 4, 5, 6, 7

[30] Hengli Wang, Rui Fan, and Ming Liu. Co-teaching: An ark to unsupervised stereo matching. In *International Conference on Image Processing (ICIP)*, 2021. 2

[31] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *International Conference on Intelligent Robots and Systems (IROS)*, 2020. 4, 5

[32] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[33] Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision (ECCV)*, 2018. 4

[34] Shumian Xin, Neal Wadhwa, Tianfan Xue, Jonathan T. Barron, Pratul P. Srinivasan, Jiawen Chen, Ioannis Gkioulekas, and Rahul Garg. Defocus map estimation and deblurring from a single dual-pixel image. *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6, 8

[35] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[36] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[37] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[38] Yinda Zhang, Neal Wadhwa, Sergio Orts-Escolano, Christian Häne, Sean Fanello, and Rahul Garg. Du2net: Learning depth estimation from dual-cameras and dual-pixels. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[39] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *International Conference on Computer Vision (ICCV)*, 2017. 2