Resolving Scale Ambiguity in Multi-view 3D Reconstruction using Dual-Pixel Sensors

Kohei Ashida, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita

Graduate School of Information Science and Technology,
Osaka University, Japan
{kohei.ashida, santo.hiroaki, okura, yasumat}@ist.osaka-u.ac.jp

Abstract. Multi-view 3D reconstruction, namely structure-from-motion and multi-view stereo, is an essential component in 3D computer vision. In general, multi-view 3D reconstruction suffers from unknown scale ambiguity unless a reference object of known size is recorded together with the scene, or the camera poses are pre-calibrated. In this paper, we show that multi-view images recorded by a dual-pixel (DP) sensor allow us to automatically resolve the scale ambiguity without requiring a reference object or pre-calibration. Specifically, the observed defocus blurs in DP images provide sufficient information for determining the scale when paired together with the depth maps (up to scale) recovered from the multi-view 3D reconstruction. Based on this observation, we develop a simple yet effective linear solution method to determine the absolute scale in multi-view 3D reconstruction. Experiments demonstrate the effectiveness of the proposed method with diverse scenes recorded with different cameras/lenses. Code and data are available at `https://github.com/kohei-ashida/dp-sfm`.

Keywords: Multi-view 3D reconstruction· Photogrammetry· Dual-pixel imaging · Scale ambiguity

1 Introduction

Multi-view 3D reconstruction is one of the most important problems in 3D computer vision. From multi-view images of a scene taken from unknown viewpoints, structure-from-motion (SfM) estimates camera poses and sparse 3D tie points of the scene. Subsequently, multi-view stereo (MVS) uses the estimated camera poses to reconstruct the dense shape of the scene. It is known that the scale ambiguity in SfM is inherent even with known camera intrinsics and cannot be resolved without external information. In this setting, the subsequent MVS can only recover dense shapes up to scale. A classical yet most reliable approach to resolving the ambiguity is to place a reference object of known size, such as a calibration target, in the scene to determine the absolute scale. However, it is not always possible to put such a reference object in the scene, and it is desired to relax the demand for diverse application scenarios.

In this paper, we propose a method for automatically determining the absolute scale in multi-view 3D reconstruction using dual-pixel (DP) cameras whose

intrinsics are calibrated. DP sensors are designed to assist the camera's auto-focus system and are becoming popular since they are equipped with Canon EOS cameras and Google Pixel smartphones. Each pixel on the DP sensor has two photodiodes to acquire two sub-aperture views. The two sub-aperture views allow us to estimate *defocus blurs* that are observed at regions off of the focal plane, although they depend on the unknown focus distance. We show that, by relating the DP defocus blur sizes with unknown focus distances with the dense depth maps up to scale obtained from SfM and MVS, the absolute scale of the scene can be fully determined without requiring other external information.

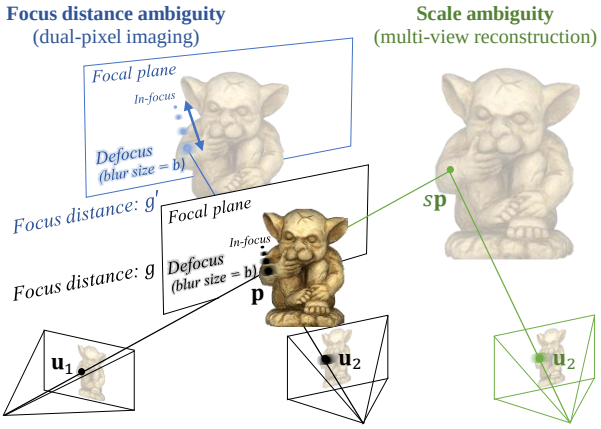


Fig. 1: Our method determines the absolute scene scale of SfM and MVS from multi-view DP images. We put together the scale ambiguity in SfM and MVS and the focus distance ambiguity of DP images. Our method jointly resolves the ambiguities.

The key idea of our method is coupling two ambiguous measurements for resolving the ambiguities: One is the linear *scale ambiguity* of depth values computed from SfM and MVS, and the other is the *focus distance ambiguity* that affinely relates the reciprocal of depths with DP defocus blurs. As illustrated in Fig. 1, the size of defocus blur b is related to unknown focus distance g and unknown scene scale s, with which scene point $\mathbf{p}$ may be scaled to $s\mathbf{p}$ retaining its projections $\mathbf{u}_1$ and $\mathbf{u}_2$. By putting them together, our method solves a linear regression problem for determining both the absolute scale of the scene and the focus distances in the multi-view images. Our method capitalizes on the fact that DP's two sub-aperture views faithfully yield defocus blur estimates, significantly more reliably than just a single view from an ordinary camera.

In summary, the chief contributions of this paper are:

- We propose a method to determine the absolute scale in uncalibrated multi-view 3D reconstruction using DP images.
- We develop a linear solution method for simultaneously determining the global scene scale and per-view focus distances.
- We conduct comprehensive experiments to assess the effectiveness of the proposed method using diverse scenes and camera settings.

2 Related Work

Scale estimation in SfM Since SfM alone cannot recover the absolute scale from the image information, it is necessary to introduce an additional cue that provides absolute scale information. A simple and reliable approach is to place a reference object of known size, such as markers and ground control points (GCPs), which is widely employed in commercial photogrammetry software. Domain-specific cues are often used to avoid manual placement of reference objects, for example, urban objects [9, 34], ground surfaces [33, 42, 43], and human faces [19]. To implicitly exploit the scale information of scene objects, recent learning-based methods use monocular depth estimation providing absolute scale information [27, 28].

Since it is not always possible to assume reference objects in target scenes, external sensors are often used to recover the scale information. The use of stereo cameras with known baselines [7, 35] or using LiDAR [13] are typical examples, while inertial sensors [23, 40] and odometry information [20, 29] are also widely used in robotics applications. Placing a refraction plate in front of the camera provides scale information using multiple capture [30, 31].

Similar to our work, defocus has been used to determine the absolute scale in the literature. *Calibrated* depth-from-defocus is used with SfM-based depths to recover the absolute scale by assuming a known and fixed focus distance [32, 37]. Similar to our method, a recent study [22] relaxes the known focus assumption by jointly estimating the physical focal length and scene scale. Although their use of the thin-lens model is similar to ours, their method requires pre-training or calibration for a learning-based blur size estimation [17] that depends on the hardware setup and focal lengths. Our study instead uses *uncalibrated* defocus blurs, where the focus distance is treated as unknown. The uncalibrated defocus blurs are reliably obtained from DP images, and thereby, our method recovers the absolute scales without prior knowledge, training data, or additional equipment.

DP imaging and applications Unlike conventional sensors, a single pixel in a DP sensor is divided into two photodiodes, one on the left and one on the right, which produces disparity. Beyond its primary use of fast auto-focus [4, 14] implemented in several commercial cameras, applications of DP imaging are actively studied. DP's disparity gives effective cues to estimate the Point Spread Functions (PSFs) of defocus blurs [25], whose direct application is defocus deblurring [1–3, 5, 15, 24, 38, 39] and blur synthesis [36]. The defocus blur information from DPs improves the performance of monocular depth estimation [10, 18, 25]. For this application, a DP image dataset with the ground truth depths has been developed [21]. The DP-based depth has been used for domain-specific applications, such as reflection removal [26] and facial shape recovery [16]. A recent method combines DP imaging and traditional two-view stereo for depth estimation [41]. Taking advantage of DP's effectiveness in acquiring defocus blur information, our work attempts to determine the scene scale automatically in multi-view 3D reconstruction.

3 Preliminary: DP Imaging and Defocus Blur

Unlike conventional sensors, a DP sensor can capture two sub-aperture images by the left and right photodiodes in a single shot. There is a difference between the two images, which is caused by the difference in the PSFs of the left and right DP views [25].

Figure 2 shows the visualization of the difference between two DP views and illustrated corresponding PSFs. There is no difference between the two views if the depth of a scene point is in focus. When the scene point is out of focus, *i.e.*, located further away or closer to the camera from the focal plane, the disparities between the two views appear in opposite directions. Unlike traditional two-view stereo, the disparity is caused by defocus blurs generated by the PSFs for each view. It is known that the signed defocus blur size $b \in \mathbb{R}$ is proportional to the *defocus* disparity d between two views [10].

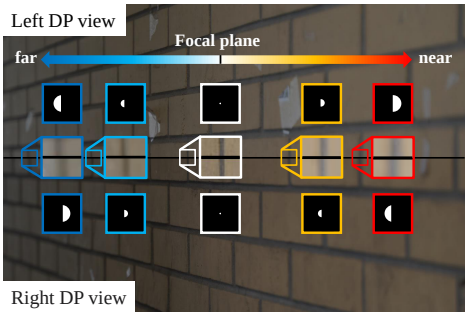


Fig. 2: Visualization of DP disparity and illustrations of the blur, *i.e.*, PSF, kernels. The two sub-aperture views *convert* the defocus blur to disparity, resulting in defocus disparity in opposite directions for near and far scene points. The DP image is from [2].

DP imaging makes the estimation of defocus blur size b tractable [25] without requiring external information or learning-based methods. Since DP PSFs $(\mathbf{H}_l, \mathbf{H}_r)$ can be well approximated by a parametric function of blur size b, when DP image patches $(\mathbf{G}_l, \mathbf{G}_r)$ are given, the blur size can be estimated per pixel by a nonlinear minimization in an unsupervised manner as

$$b = \underset{b}{\operatorname{argmin}} \|\mathbf{G}_l * \mathbf{H}_r(b) - \mathbf{G}_r * \mathbf{H}_l(b)\|_F, \tag{1}$$

where $*$ is a convolution operator. The readers are referred to [25] for details.

4 Proposed Method

Our method resolves the scale ambiguity in multi-view 3D reconstruction using multi-view DP images. The overview is illustrated in Fig. 3. Our method infers the unknown scene scale just by capturing multi-view images with a DP camera. DP images provide a useful cue for blur size estimation, as described above, without requiring any external information. Our method shares the same spirit with [22]; however, there is a key difference in that our method does not require pre-training but capitalizes on the DP sensor's effectiveness in determining blur sizes. We show in Fig. 3 that the estimated blur size and (inverse) depth from SfM and MVS have an affine relationship, and thus, the unknown scene scale and focus distance can be estimated simultaneously by linear regression.

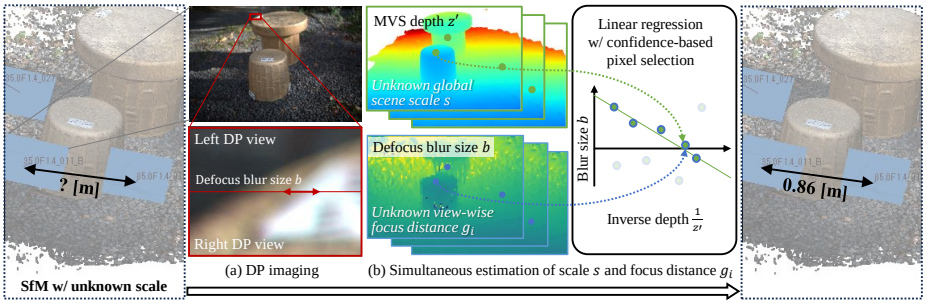


Fig. 3: Overview of our method that provides the absolute scale to multi-view 3D reconstruction (SfM and MVS) just by capturing the scene using DP cameras. (a) The disparity between DP views provides the cue to estimate the defocus blur size. (b) From the affine relation between the (inverse) depth and the defocus blur size, our method simultaneously estimates the scene scale and focus distances.

We begin by describing the ambiguities that arise in our setting, then describe the solution method for resolving them. We also present a method for selecting reliable pixels and views for scale estimation, which is based on the inference by an off-the-shelf blur size estimation method [25].

Assumptions We assume that camera intrinsics, such as focal lengths and aperture sizes (f-numbers), are known as they are usually accessible from the camera system. Although defocus blurs slightly change the center of projection, we assume the change is sufficiently small and ignore their effect on the intrinsics. As stated, we use DP images taken from multiple viewpoints as input. Our method does not assume pre-calibrated viewpoints nor reference objects in a scene.

4.1 Scale and Focus Distance Ambiguities

There are two ambiguities in our setting: One is the scale ambiguity in SfM and MVS, and the other is the focus distance ambiguity in DP images.

Scale ambiguity in SfM and MVS Camera positions and 3D point locations recovered by SfM and the subsequent MVS have a scale ambiguity even with known camera intrinsics. Let us consider a 3D point $\mathbf{p} = [x, y, z]^\top$ projected to the corresponding 2D point $\mathbf{u} = [u, v, 1]^\top$ in the homogeneous coordinates with a projection matrix $[\mathbf{R}|\mathbf{t}]$ formed by a rotation matrix $\mathbf{R} \in SO(3)$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$. With known camera intrinsics, we can set the intrinsics to be identity and write the 3D-to-2D perspective projection in some Euclidean world coordinate system as

$$\mathbf{u} \sim \begin{bmatrix}\mathbf{R}\ \mathbf{t}\end{bmatrix} \begin{bmatrix}\mathbf{p} \\ 1\end{bmatrix} = \frac{1}{s} \begin{bmatrix}\mathbf{R}\ s\mathbf{t}\end{bmatrix} \begin{bmatrix}s\mathbf{p} \\ 1\end{bmatrix}, \tag{2}$$

where $s \in \mathbb{R}$ is an unknown scaling factor. As such, the depth z' obtained by SfM and MVS is linearly related to the true depth z by the unknown scaling factor s as

$$z = sz'. \tag{3}$$

This is the inherent ambiguity in SfM, which we are going to resolve in this paper.

Focus distance ambiguity in DP images As described by Wadhawa *et al.* [36] and Garg *et al.* [10], the signed defocus blur size $b \in \mathbb{R}$ is related to scene depth z, focal length f, lens aperture diameter l, and the focus distance g in $\mathbb{R}$ as

$$b \approx \frac{lf}{1 - f/g}\left(\frac{1}{g} - \frac{1}{z}\right) \tag{4}$$

using the paraxial and thin-lens approximations. The blur size b in DP images varies with the focus distance g even when the same scene depth z is recorded with the same camera parameters (l, f). Therefore, even with the blur size b from DP images, its relationship to depth z is ambiguous due to unknown focus distance g and depth z as

$$b \approx \frac{lf}{1 - f/g}\left(\frac{1}{g} - \frac{1}{z}\right) = \frac{lf}{1 - f/g'}\left(\frac{1}{g'} - \frac{1}{z'}\right), \tag{5}$$

i.e., there exists a set of pairs like (g, z) and (g', z') that yields the same blur size b. The blur size b is signed, *i.e.*, it becomes positive when the scene point is farther away from the focal plane and negative when nearer to the camera.

4.2 Solution Method

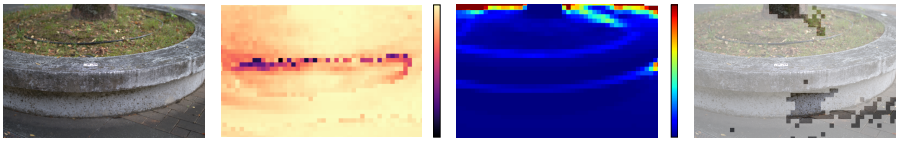
Given the blur sizes $b_{ij} \in \mathbb{R}$ obtained from DP images, depths $z'_{ij} \in \mathbb{R}_+$ from SfM and MVS in the i-th view at j-th pixel, and known per-view focal lengths f_i and lens aperture l_i, our goal is to determine unknown scene scale $s \in \mathbb{R}_+$ and per-view focus distances $g_i \in \mathbb{R}_+$. Since we use per-view intrinsics, our method can be used in a practical setting of multi-view 3D reconstruction, in which camera intrinsics may vary across viewpoints.

We put together the two ambiguous observations, *i.e.*, depths up to scale and blur sizes with unknown focus distances, by substituting Eq. (3) to Eq. (4) as

$$b_{ij} = \frac{l_i f_i}{1 - f_i/g_i}\left(\frac{1}{g_i} - \frac{1}{sz'_{ij}}\right). \tag{6}$$

Denoting $\bar{g}_i = \frac{1}{g_i}$ and $\bar{s} = \frac{1}{s}$ as g_i and s are both non-zero, Eq. (6) can be rewritten as

$$\frac{b_{ij} z'_{ij}}{f_i} = \underbrace{z'_{ij}\left(b_{ij} + l_i\right)}_{\gamma_{ij}} \bar{g}_i - l_i \bar{s} = \gamma_{ij} \bar{g}_i - l_i \bar{s}, \tag{7}$$



(a) Input image **(b)** Confidence score **(c)** Local depth variation **(d)** Selected pixels

Fig. 4: Pixel selection within a view based on confidence scores. We select reliable pixels that contain good blur size estimates from an input image (a). (b) shows the log-scaled confidence scores c that show the reliability of blur size estimates [25]. (c) Local variations (variance) of depths. (d) Selected pixels based on our pixel selection method (highlighted in the figure).

which is linear with respect to the unknown parameters $\bar{g}_i$ and $\bar{s}$.

From observations of multiple views and pixels, we can construct an over-determined system of linear equations from Eq. (7). Given the number of views n and the number of valid pixels p_i in the i-th view, we have $\sum_{i=1}^{n} p_i$ equations, which can be written in a matrix form as:

$$\underbrace{\begin{bmatrix} \gamma_{11} & & & -l_1 \\ \vdots & & & \vdots \\ & \mathbf{0} & & \\ \gamma_{1p_1} & & & -l_1 \\ & \gamma_{21} & & -l_2 \\ & \vdots & & \vdots \\ \mathbf{0} & & \mathbf{0} & \\ & \gamma_{2p_2} & & -l_2 \\ & & \ddots & \vdots \\ & & \gamma_{n1} & -l_n \\ & & \vdots & \vdots \\ & \mathbf{0} & & \\ & & \gamma_{np_n} & -l_n \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \vdots \\ \bar{g}_n \\ \bar{s} \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} b_{11} z'_{11}/f_1 \\ \vdots \\ b_{1p_1} z'_{1p_1}/f_1 \\ b_{21} z'_{21}/f_2 \\ \vdots \\ b_{2p_2} z'_{2p_2}/f_2 \\ \vdots \\ b_{n1} z'_{n1}/f_n \\ \vdots \\ b_{np_n} z'_{np_n}/f_n \end{bmatrix}}_{\mathbf{b}} . \tag{8}$$

Since we have $n+1$ unknowns in vector $\mathbf{x}$, given $\mathbf{b} \neq \mathbf{0}$, the condition for obtaining a unique (approximate) solution is $\mathsf{rank}\,(\mathbf{A}) = n+1$, which is, in practice, easy to achieve because a large number of pixels are available. The minimum setting to determine scale s is selecting two pixels with distinct depths in one of the multi-view images, which makes a full-rank 2×2 matrix $\mathbf{A}$ for two unknowns $\mathbf{x} = [\bar{g}_i, \bar{s}]^\top$.

In practice, we can select many pixels from multiple views to rely on robust estimation. With a sufficiently large number of selected pixels, we construct a tall and skinny matrix $\mathbf{A}$ and use ℓ_1 residual minimization to effectively disregard outliers as

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \left\| \mathbf{A}\mathbf{x} - \mathbf{b} \right\|_1 . \tag{9}$$

We use the Iteratively Reweighted Least Squares (IRLS) method [12] for deriving the approximate solution $\mathbf{x}^*$. Obviously, other choices of robust estimation methods, such as RANdom SAmple Consensus (RANSAC) [8], are viable for alleviating the impact of outliers.

4.3 Confidence-based Pixel Selection

So far, we have assumed that the blur sizes b_{ij} are available to our method. For obtaining the blur sizes, we use the method proposed by Punnappurath *et al.* [25] that optimizes Eq. (1). For our case, we only require blur size estimates for a small number of pixels but not for all the pixels since our goal is determining the scene scale s using Eq. (8). We, therefore, develop a pixel selection heuristic to choose a small number of pixels that have reliable blur size estimates. Specifically, we select the reliable pixels in each view, followed by selecting the views that produce reliable estimates of scene scale.

Pixel-selection within a view We begin by adopting the confidence scores of [25] that indicate the reliability of the blur size estimates based on the residual of the blur size optimization and the textured-ness (magnitude of image gradient) of the pixel of interest. We slightly modified their original confidence score using a depth-based criterion since we observed that the large depth variation in an image patch negatively affects the blur size estimation.

Let us consider one of the views and let c_j be the confidence score defined in [25] at the j-th pixel. In our case, a depth value z'_j up to scale is available at the pixel, and we can assess the variance of the depth values within an image patch $\mathcal{P}_j$ centered at the j-th pixel as $\mathrm{var}\left(\{z'_k\}\right), k \in \mathcal{P}_j$. Based on the variance of the depths, we re-define the confidence score as

$$c_j^* \triangleq \frac{c_j}{\mathrm{var}\left(\{z'_k\}\right)}, \quad k \in \mathcal{P}_j. \tag{10}$$

The greater depth variation in an image patch leads to a larger variance and lower confidence score, and vice versa. Based on the new confidence score c_j^*, we select the top $T_c\%$ of pixels and regard them as reliable pixels that have good defocus blur estimates. Although the variance of depths varies with the depth scale, which is unknown in our case, it does not affect the confidence score c_j^* within a view because the same scale is applied to all the pixels in the view. Figure 4 depicts the pixel selection within a view.

View selection In the multi-view images for SfM, there may be images that are not suitable for scale estimation. For example, a view with small depth variations is not preferred because the small depth variations make the linear system of Eq. (8) unstable. Therefore, we further select reliable views for scale estimation.

Since our method can estimate scale s for each view as long as the depths up to scale are available, we first compute the scale predictions for each view independently using the selected pixels within each view by solving Eq. (8). Ideally, the scales estimated for each view should match; however, in reality, they do not due to outliers. We first disregard the views with clear outliers, *i.e.*, views with excessively small blur size variations and views that yield negative values of scale estimates. The small blur size variations indicate that the scene contains small depth variations. We simply set a threshold to discard the views with small blur size variations. Namely, when the difference between maximum

and minimum blur size estimates becomes less than or equal to T_p pixels, the view is unused for the estimation. The negative scale estimates could arise due to various reasons, such as image degradation by motion blur and inaccurate estimates of either or both blur sizes and depths. We simply discard such views from the estimation.

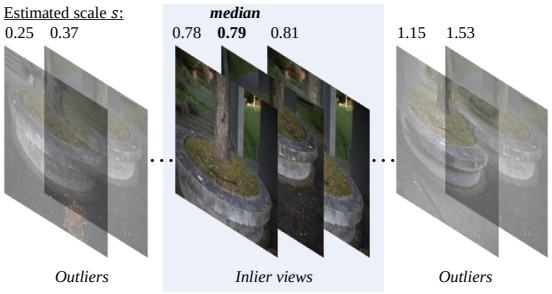


Fig. 5: View selection. We select the *reliable* views for scale estimation based on the median of the scale values estimated for each view.

Finally, from the remaining views, we select N_v views that are close to the median of the estimated scales as illustrated in Fig. 5. Once the appropriate views are selected, by using all the selected pixels within the views, we create the linear system of Eq. (8) and re-estimate the scale s.

4.4 Implementation Details

For SfM and MVS, we use the commercial software Metashape[1]. For DP blur size estimation, we use the method proposed by Punnappurath *et al.* [25] as described, except for the smoothing operation implemented as post-processing. All captured images are undistorted using the parameters estimated by MVS before being fed into the blur size estimation. Our method is not restricted by the choice of SfM+MVS and blur size estimation methods, but we found these to be the best choices today.

For the focal length f, we use the one from the lens specification. The aperture diameter l is calculated using f-number N of the lens as $l = \frac{f}{N}$. For the hyper-parameters, we use $T_c = 10$, $T_p = 2$, and $N_v = 7$ throughout the experiments.

5 Experiment

We assess the accuracy of our scale estimation method both quantitatively and qualitatively. We use multi-view DP images captured in six scenes with different capturing settings, such as, cameras (DSLR and smartphone), lenses, and aperture sizes.

Recording setting We use two different types of cameras equipped with a DP sensor: a DSLR camera and a smartphone. For the DSLR camera setting, we use Canon EOS 5D Mark IV with three different lenses: EF35mm F1.4L II USM, EF50mm F1.8 STM, and EF85mm F1.8 USM. For each scene, we also change the aperture size to assess the effect of the aperture size on our method. The

[1] Agisoft Metashape `https://www.agisoft.com/`, last accessed on July 12, 2024.

maximum aperture ($f/1.4$ or $f/1.8$ depending on the lens), $f/4.0$, and $f/8.0$ are used for the experiment depending on the target scene. The images from Canon's DLSR camera are recorded as Canon Raw images with the CR2 format, and the left and right DP views are extracted using DPRSplit[2]. For a more accessible setup, we also use a smartphone, Google Pixel 4, which has a focal length of 4.38 mm and a fixed aperture of $f/1.73$. To obtain the left and right DP views, we use the capturing software used in [11].

For each scene for each camera setting, we capture about 30 images using a DSLR, and 60 images using a smartphone. For fair comparisons across different lens and aperture sizes, for each viewpoint, we use a tripod to keep the camera fixed and record images with different camera settings. To evaluate the scale estimates, we obtain the ground-truth scene scale by placing markers with known sizes in the scenes.

Evaluation metric As a quantitative evaluation metric, we use the scale ratio r_s of the estimated scene scale s_{est} to the ground-truth scale s_{gt}, which is defined as follows:

$$\text{Scale ratio} \quad r_s = \frac{s_{\text{est}}}{s_{\text{gt}}}.$$

In the above metric, closer to 1 indicates good scale estimation. The scale ratio smaller than 1 indicates that the scale is under-estimated, *i.e.*, the estimated scale is smaller than the true scale, and the greater scale ratio corresponds to the over-estimation. While scale ratio r_s is defined for each scene and camera setting, we also define the average error of the scale ratios e_s to characterize the overall effectiveness as

$$\text{Average error of scale ratios} \quad e_s = \frac{1}{n_s} \sum \left(\max(r_s, r_s^{-1}) - 1 \right),$$

where n_s is the number of settings, and $\max(r_s, r_s^{-1})$ is for taking into account the under- and over-estimations.

Quantitative results The results of the scale estimation for six scenes using two DP cameras are summarized in Table 1. The DSLR camera uses three lenses and different aperture sizes. Overall, our method can reasonably well estimate the absolute scene scale just by using a DP camera for multi-view capture, considering the fact that scale ratio r_s can range in $(0, \infty)$. The average error e_s across scenes and lenses is approximately 0.219 as shown in the table. From comparing the average errors for each lens, we see that the difference in lenses does not significantly affect the overall performance.

Visual results Figures 6 and 7 show visual examples of our results by the DSLR and smartphone, as well as the estimated blur sizes and the ground-truth depth maps. We can confirm the same trend as in the quantitative results, where the

[2] DPRSplit `https://www.fastrawviewer.com/DPRSplit`, last accessed on July 12, 2024.

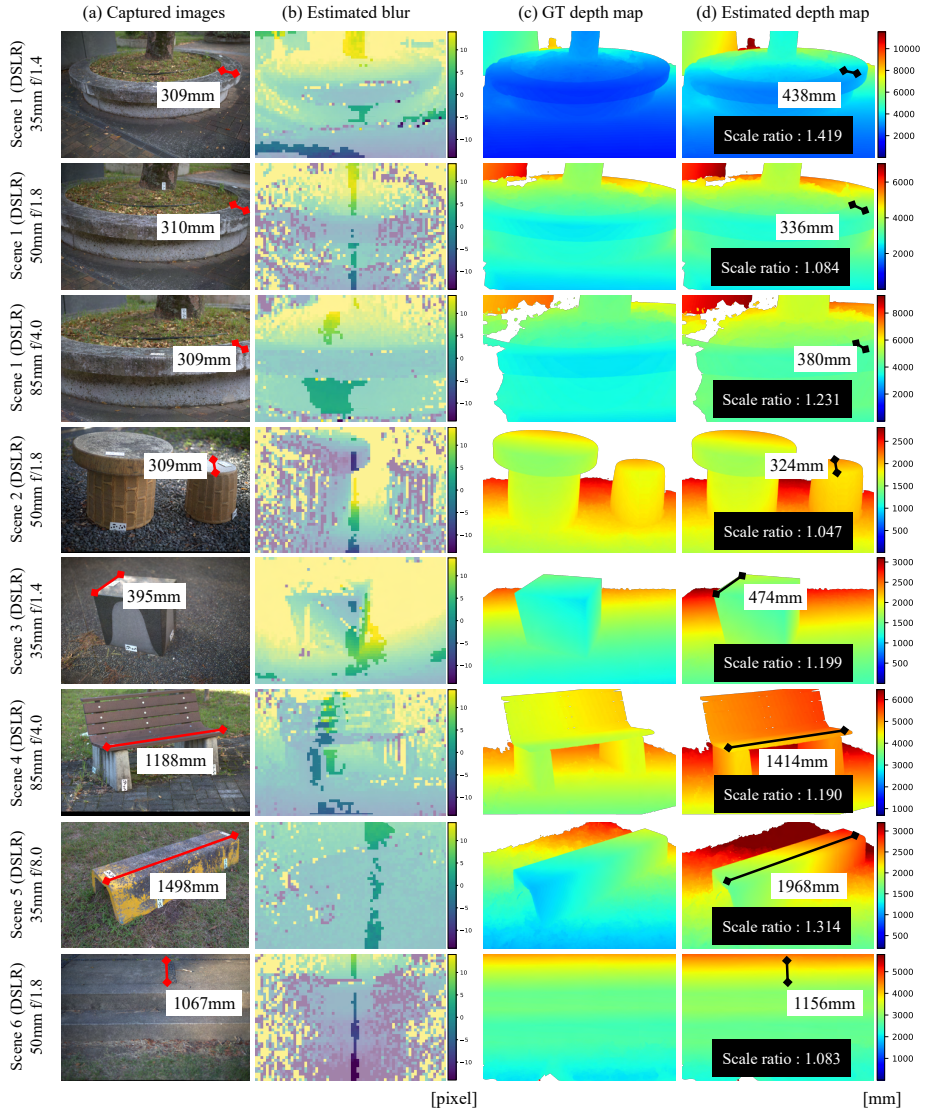


Fig. 6: Visual results of real scenes captured by the DSLR (Canon EOS 5D Mark IV). Our method yields reasonable estimates of scene scales just by using a DP camera for multi-view capture, which is theoretically unrecoverable by standard SfM alone. The top three rows show the results of the same scene (and from almost the same viewpoint) but with different aperture sizes or focal lengths. The bottom rows show the results of different scenes. In (a), the distance indicates the true distance obtained from the GT 3D model. In (d), the distance in the estimated scale is shown. (b) shows estimated blur size and selected pixels based on our pixel selection method (highlighted in the figure).

Table 1: Scale ratio r_s between the ground truth and the estimated scale for each scene, lens, and aperture size. Maximum aperture varies depending on the lens (*i.e.*, $f/1.4$ or $f/1.8$). e_s represents the average errors of the corresponding row/column/total. "-" indicates that we did not record the scene under the setting.

	DSLR (35mm)			DSLR (50mm)			DSLR (85mm)			Smartphone	e_s
	$f/1.4$	$f/4.0$	$f/8.0$	$f/1.8$	$f/4.0$	$f/8.0$	$f/1.8$	$f/4.0$	$f/8.0$	$f/1.73$	
Scene 1	1.419	1.088	-	1.084	1.084	-	1.100	1.231	-	1.085	0.156
Scene 2	1.225	1.238	-	1.047	1.232	-	1.247	1.302	-	1.252	0.220
Scene 3	1.199	1.229	-	1.055	1.352	-	1.238	1.226	-	1.519	0.260
Scene 4	1.181	1.107	1.477	1.027	1.266	1.469	1.225	1.190	1.296	1.034	0.227
Scene 5	1.171	1.072	1.314	1.067	1.099	1.285	1.038	1.205	1.425	1.068	0.174
Scene 6	1.148	1.073	-	1.083	1.029	-	1.123	1.126	-	1.388	0.139
e_s	0.224	0.135	0.396	0.061	0.177	0.377	0.162	0.213	0.361	0.224	0.219

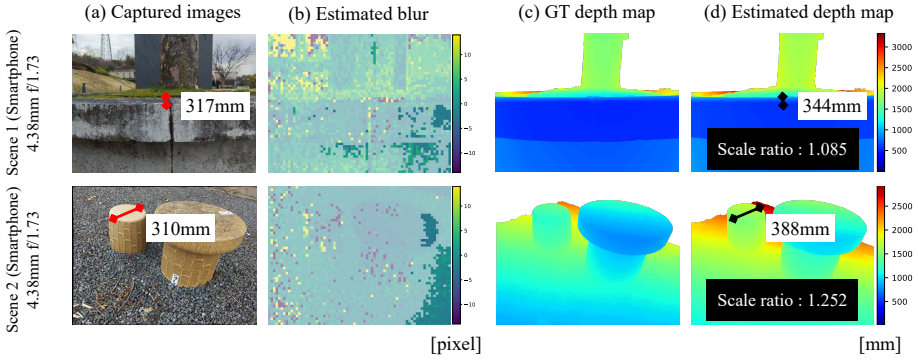


Fig. 7: Visual results of real scenes captured by a smartphone (Google Pixel 4). Similar to the results by DSLR, our method estimates reasonable scene scales despite the DP blur being relatively smaller than DSLR's.

depth maps with estimated scale well represent the actual depths. Our method estimates the scene scale with less than 0.2 error ratio for several settings, where the blur size estimation is relatively stable. For these cases, 3D reconstruction with our absolute scale estimation allows high-quality measurements with an error of several centimeters in wide outdoor scenes.

Stability of our method To assess the stability of our method against variation of input images, we capture additional $200 \sim 300$ images for Scenes 1 and 2 using the DSLR equipped with a 35mm lens and $f/1.4$ aperture size and repeatedly estimate the scale using different sets of images. Figure 8 shows an overlay of estimated models with repeated experiments and the minimum and maximum estimated scale ratios, $r_{s\min}$ and $r_{s\max}$, with 10 trials.

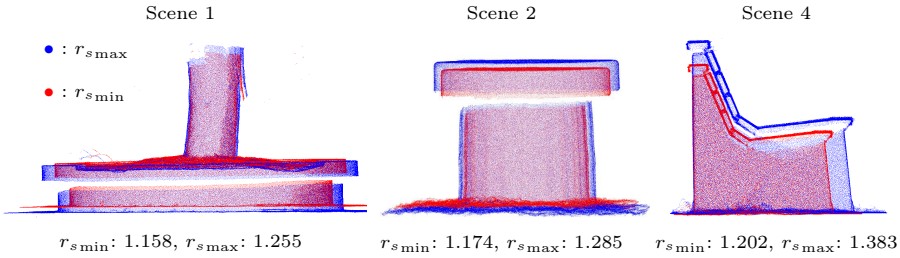


Fig. 8: Overlay of estimated models with repeated experiments and the minimum and maximum estimated scale ratios with 10 trials. Blue and red point clouds represent the ones with scale ratios $r_{s\max}$ and $r_{s\min}$, respectively.

Comparison with scale estimation by monocular depth estimation We compare our method with a single-image metric depth estimation, ZoeDepth [6], for the same scenes and lens setup used. Since it estimates metric depth for each pixel, we compute the per-image scale as the average ratio between ZoeDepth and MVS depth maps. We then average the per-image scale over seven inlier viewpoints selected using the same procedure as our view selection.

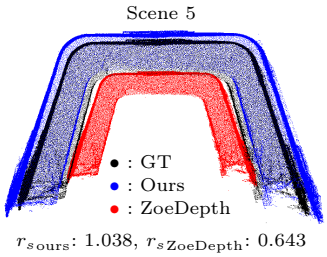
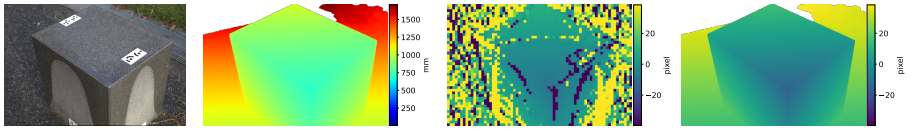


Fig. 9: Comparison with monocular scale estimation. Blue and red points represent the ones of the scale ratio estimated by our method ($r_{s\text{ours}}$) and ZoeDepth ($r_{s\text{ZoeDepth}}$), respectively. Black points represent the GT model.

The average error e_s across scenes and lenses using ZoeDepth is 0.679, which is more than 3 times larger than ours (0.219). Figure 9 confirms this trend through the visualization of the estimated scale. More detailed comparisons are shown in the supplementary material.

Limitations and failure cases Although our method successfully estimates the scene scale in many cases, we find several failure modes in our experiments. Since our method is based on defocus blur, scale estimation from pan-focus images becomes challenging. In particular, since the aperture of smartphone cameras is smaller than that of DSLR's, it was necessary for us to include photographs that exhibit large depth variations to produce relatively large blur like the top row of Fig. 7. Although the pan-focus images are preferable in SfM and MVS for better correspondence matching, in practice, small apertures cannot always be used. This is because the small aperture limits the amount of incoming light, requiring longer exposure times, and resulting in camera motion blurs.

The success of our method relies on the accurate estimation of both defocus blur sizes and depth maps up to scale. We found several examples where the scale estimation is affected by the blur size estimation errors, while the depth maps up to scale are almost always stably estimated. Figure 10 shows a failure



(a) Input image **(b)** GT depth **(c)** Estimated blur size **(d)** Ideal blur size

Fig. 10: An example of failure case (Scene 3 captured with a 50mm lens and $f/4.0$ aperture size) producing scale ratio $r_s = 1.352$ due to the large error in blur size estimation. Ideal blur sizes (d) are calculated from the ground-truth (GT) scene depth (b) and Eq. (6). Since the target scene is mostly textureless or a repetition of fine textures, most of the pixels in the estimated blur sizes (c) deviate from ideal values, making the selection of the inlier pixels challenging.

mode of blur size estimation as well as the estimated scene depth, whose scale ratio r_s reaches as large as 1.352 (Scene 3 captured with 50mm lens and $f/4.0$ aperture size). To assess the quality of blur size estimates shown in Fig. 10 (c), we synthesize the *ideal* blur sizes in Fig. 10 (d) using the ground-truth depth of Fig. 10 (b) that is computed from the markers and the image formation model of Eq. (6). Most of the blur size estimates are far from the ideal blur sizes due to the target object being almost textureless or the repetition of fine-grained texture, which makes blur size optimization challenging. We further analyze the accuracy and characteristics of the existing blur size estimation methods in the supplementary material.

6 Discussions

This paper has shown that the scale information in multi-view 3d reconstruction, which is theoretically unsolvable by SfM and MVS itself, can be inferred from the defocus blur information provided by DP imaging. Our method estimates the absolute scale without relying on reference objects or data priors but by simply capturing multi-view images with a DP camera. The key to the scale recovery is the unique feature of DP that relates defocus disparity, blur size, and scene depth, from which we can affinely relate the reciprocal of the depth and the defocus blur sizes given by DP images. Together with the pixel selection method for improving robustness using the confidence information from the off-the-shelf DP blur estimation method as well as the statistics of the depth information, we simultaneously recover the scene scale and the unknown focus distance via a simple linear regression.

Experiments show that our method successfully estimates the absolute scale in many cases. Since the estimation accuracy of the blur size strongly affects the performance of our method, the analysis of DP blur characteristics will be beneficial in the future. With a method for a more accurate and stabler estimation of blur sizes in DP images, our method will produce more accurate absolute scales for multi-view 3D reconstruction.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23H05491.

References

1. Abuolaim, A., Afifi, M., Brown, M.S.: Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 82–90 (2022)
2. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 111–126 (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 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2289–2298 (2021)
4. Abuolaim, A., Punnappurath, A., Brown, M.S.: Revisiting autofocus for smartphone cameras. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 545–559 (2018)
5. Abuolaim, A., Timofte, R., Brown, M.S.: NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 578–587 (2021)
6. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
7. Engel, J., Stückler, J., Cremers, D.: Large-scale direct SLAM with stereo cameras. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1935–1942 (2015)
8. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
9. Frost, D., Prisacariu, V., Murray, D.: Recovering stable scale in monocular SLAM using object-supplemented bundle adjustment. IEEE Transactions on Robotics **34**(3), 736–747 (2018)
10. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7628–7637 (2019)
11. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7628–7637 (2019)
12. Gentle, J.: Matrix Albegra. Springer Texts in Statistics, Springer, New York (2007)
13. Giubilato, R., Chiodini, S., Pertile, M., Debei, S.: Scale correct monocular visual odometry using a lidar altimeter. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3694–3700 (2018)
14. Herrmann, C., Bowen, R.S., Wadhwa, N., Garg, R., He, Q., Barron, J.T., Zabih, R.: Learning to autofocus. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2227–2236 (2020)
15. Jung, S.H., Heo, Y.S.: Disparity probability volume guided defocus deblurring using dual pixel data. In: Proceedings of International Conference on Information and Communication Technology Convergence (ICTC). pp. 305–308 (2021)
16. Kang, M., Choe, J., Ha, H., Jeon, H.G., Im, S., Kweon, I.S.: Facial depth and normal estimation using single dual-pixel camera. In: Proceedings of European Conference on Computer Vision (ECCV) (2022)
17. Kashiwagi, M., Mishima, N., Kozakaya, T., Hiura, S.: Deep depth from aberration map. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV) (2019)

18. Kim, D., Jang, H., Kim, I., Kim, M.H.: Spatio-focal bidirectional disparity estimation from a dual-pixel image. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5023–5032 (2023)
19. Knorr, S.B., Kurz, D.: Leveraging the user's face for absolute scale estimation in handheld monocular SLAM. In: Proceedings of IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 11–17 (2016)
20. Lee, S.H., de Croon, G.: Stability-based scale estimation for monocular SLAM. IEEE Robotics and Automation Letters **3**(2), 780–787 (2018)
21. Li, F., Guo, H., Santo, H., Okura, F., Matsushita, Y.: Learning to synthesize photorealistic dual-pixel images from RGBD frames. In: International Conference on Computational Photography (ICCP). pp. 1–11 (2023)
22. Mishima, N., Seki, A., Hiura, S.: Absolute scale from varifocal monocular camera through SfM and defocus combined. In: Proceedings of British Machine Vision Conference (BMVC) (2021)
23. Nützi, G., Weiss, S., Scaramuzza, D., Siegwart, R.: Fusion of IMU and vision for absolute scale estimation in monocular SLAM. Journal of Intelligent & Robotic Systems **61**(1–4), 287–299 (2011)
24. Pan, L., Chowdhury, S., Hartley, R., Liu, M., Zhang, H., Li, H.: Dual pixel exploration: Simultaneous depth estimation and image restoration. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4340–4349 (2021)
25. Punnappurath, A., Abuolaim, A., Afifi, M., Brown, M.S.: Modeling defocus-disparity in dual-pixel sensors. In: International Conference on Computational Photography (ICCP). pp. 1–12 (2020)
26. Punnappurath, A., Brown, M.S.: Reflection removal using a dual-pixel sensor. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1556–1565 (2019)
27. Roussel, T., Van Eycken, L., Tuytelaars, T.: Monocular depth estimation in new environments with absolute scale. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1735–1741 (2019)
28. Rukhovich, D., Mouritzen, D., Kaestner, R., Rufli, M., Velizhev, A.: Estimation of absolute scale in monocular SLAM using synthetic data. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 803–812 (2019)
29. Scaramuzza, D., Fraundorfer, F., Pollefeys, M., Siegwart, R.: Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1413–1419 (2009)
30. Shibata, A., Fujii, H., Yamashita, A., Asama, H.: Absolute scale structure from motion using a refractive plate. In: Proceedings of IEEE/SICE International Symposium on System Integration (SII). pp. 540–545 (2015)
31. Shibata, A., Fujii, H., Yamashita, A., Asama, H.: Scale-reconstructable structure from motion using refraction with a single camera. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 5239–5244 (2015)
32. Shiozaki, T., Dissanayake, G.: Eliminating scale drift in monocular SLAM using depth from defocus. IEEE Robotics and Automation Letters **3**(1), 581–587 (2017)
33. Song, S., Chandraker, M.: Robust scale estimation in real-time monocular sfm for autonomous driving. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1566–1573 (2014)

34. Sucar, E., Hayet, J.B.: Bayesian scale estimation for monocular SLAM based on generic object detection for correcting scale drift. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 5152–5158 (2018)
35. Sumikura, S., Sakurada, K., Kawaguchi, N., Nakamura, R.: Scale estimation of monocular SfM for a multi-modal stereo camera. In: Proceedings of Asian Conference on Computer Vision (ACCV). pp. 281–297 (2019)
36. 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) pp. 1–13 (2018)
37. Wöhler, C., d'Angelo, P., Krüger, L., Kuhl, A., Groß, H.M.: Monocular 3D scene reconstruction at absolute scale. ISPRS Journal of Photogrammetry and Remote Sensing **64**(6), 529–540 (2009)
38. 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. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2228–2238 (2021)
39. Yang, Y., Pan, L., Liu, L., Liu, M.: K3DN: Disparity-aware kernel estimation for dual-pixel defocus deblurring. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13263–13272 (2023)
40. Zhang, S., Zhang, J., Tao, D.: Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating IMU motion dynamics. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 143–160 (2022)
41. Zhang, Y., Wadhwa, N., Orts-Escolano, S., Häne, C., Fanello, S.R., Garg, R.: Du2Net: Learning depth estimation from dual-cameras and dual-pixels. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 582–598 (2020)
42. Zhou, D., Dai, Y., Li, H.: Reliable scale estimation and correction for monocular visual odometry. In: Proceedings of IEEE Intelligent Vehicles Symposium (IV). pp. 490–495 (2016)
43. Zhou, D., Dai, Y., Li, H.: Ground-plane-based absolute scale estimation for monocular visual odometry. IEEE Transactions on Intelligent Transportation Systems **21**(2), 791–802 (2019)