Absolute Scale from Varifocal Monocular Camera through SfM and Defocus Combined

Nao Mishima[1]
nao.mishima@toshiba.co.jp
Akihito Seki[1]
akihito.seki@toshiba.co.jp
Shinsaku Hiura[2]
hiura@eng.u-hyogo.ac.jp

[1] Corporate Research and Development Center, Toshiba Corp.
[2] University of Hyogo

Abstract

The absolute scale estimation of monocular structure from motion (SfM) is still under-explored even though it is essential for robotic tasks or real-world interaction. Typically, the use of physical scale cues requires a calibration process while context scale cues introduce geometric assumptions. In this paper, we propose a novel method to obtain absolute scales of the scene and camera motion by combining monocular SfM and uncalibrated depth from defocus (DfD) which is free for zooming and focusing on each shot independently. Specifically, we exploit that the scene structure and field of view (FoV) of each camera estimated by SfM are tightly coupled to the focal length and focused distance of DfD, and the radius of the effective aperture of the lens constrains the absolute scale of the entire estimation. The effectiveness of the proposed method is verified by using a commercially available camera with a varifocal lens through various experiments.

1 Introduction

Accurate depth estimation is a key component in many robotic tasks, including perception, navigation, and planning. In the last few decades, there has been significant progress in investigating methods to acquire depth using only a monocular camera, using motion parallax [2, 45, 47], context [12, 15, 16], and physical cues [13, 20, 41]. Structure from motion (SfM) is a typical method based on motion parallax [2, 45, 47]. Various methods introduce absolute scale by utilizing prior knowledge such as the height of the camera installation and the flatness of the ground [4, 5, 17, 36, 37, 53]. This introduces assumptions and reduces versatility. Another approach is to utilize additional sensors, such as stereo cameras and inertial sensors to overcome the scale ambiguity [25, 26, 31, 50]. In small robots, space-limitation makes additional sensors unfeasible. The context-based method that uses deep learning to directly regress the target depth based on the context is another typical approach for monocular depth estimation [12, 15, 16, 21, 34]. The combination of context and SfM has also been used to estimate the absolute scale of monocular SfM [44, 49]. However, if the target application has a different scale from the pre-training, the pre-trained model will have

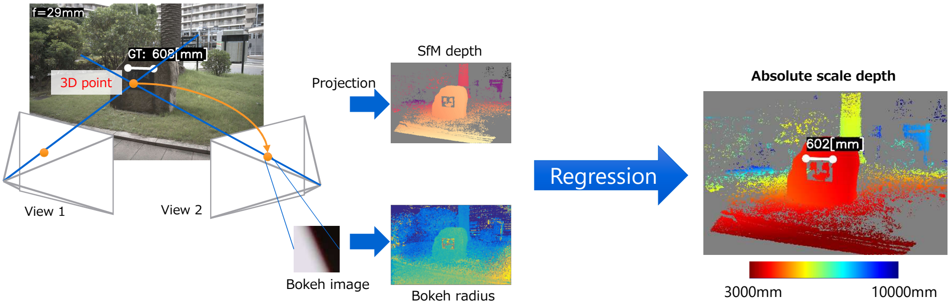


Figure 1: We have made it possible to simultaneously obtain an unknown scale β and the unknown physical focal length f by associating the scene structure of the geometric side (SfM depth) and the bokeh radius of the physical side. Object size: GT=608[mm], ours=602[mm].

a large error due to the tight coupling of context and scale. Finally, we note all methods still have problems in obtaining absolute scale due to the scale ambiguity in the 3D geometry.

Good property	Context [44]	Prior knowledge [53]	Sensor [31]	Physical cue [20]	Ours
Context-free		✓	✓	✓	✓
Prior knowledge-free	✓		✓	✓	✓
Only monocular camera	✓	✓		✓	✓
Pre-calibration-free	✓		✓		✓

Table 1: While existing methods require some information or introduce assumptions, our method has the fewest assumption.

A scene geometry independent method is the depth from defocus (DfD). In DfD, the absolute scale depth can be obtained from the change in bokeh radius of two captured images with known focal length and focus distance [9, 13, 32, 38, 41, 46]. It has been shown in [20] that the bokeh radius with lens aberration uniquely determines the depth even from a single monocular image. However, to convert from the bokeh radius to the absolute scale depth, the physical focal length and focused distance must be calibrated. Therefore, it cannot be applied to uncalibrated cameras such as varifocal lenses that cannot be calibrated in advance.

In this paper, we target absolute scale estimation from varifocal monocular cameras. To the best of our knowledge, there is no method for estimating the absolute scale without making any use of context, prior knowledge, additional sensors, and pre-calibration of its focal length (Table 1). Our goal is to find the absolute scale using only a monocular camera without loss of flexibility. In particular, we propose a novel method to obtain absolute scales by combining monocular SfM and uncalibrated DfD which is free for zooming and focusing on each shot independently (Figure 1). Specifically, the scene structure and field of view (FoV) of each camera estimated by SfM are tightly coupled to the focal length and focused distance of DfD, and the radius of the effective aperture of the lens constrains the absolute scale of the entire estimation. To demonstrate the effectiveness of the proposed method, we conducted various experiments using a varifocal lens. As mentioned above, if the lens is fixed, the absolute scale depth can be obtained by pre-calibration of the focal length [20]. If the context does not change, the absolute scale can be estimated by learning the scale contained in the context [44, 49]. However, the problem we are tackling is difficult to learn in an end-to-end manner because neither the lens nor the context is fixed. Therefore, we believe that the combination of physical cue-based deep learning and geometric information will be the breakthrough for absolute scale estimation with high flexibility.

The contributions of this paper are as follows. (1) We propose the first framework to estimate the absolute scale independent of the scene using only an uncalibrated monocular camera, as we combine monocular SfM and uncalibrated DfD. (2) We formulate the problem as a nonlinear regression problem. Moreover, we show that the nonlinear regression problem is a convex function with a unique local minimum where a global solution is easily found. (3) We demonstrate the effectiveness of the proposed method on our challenging dataset with varifocal cameras and provide an ablation study in comparison to the state of the art.

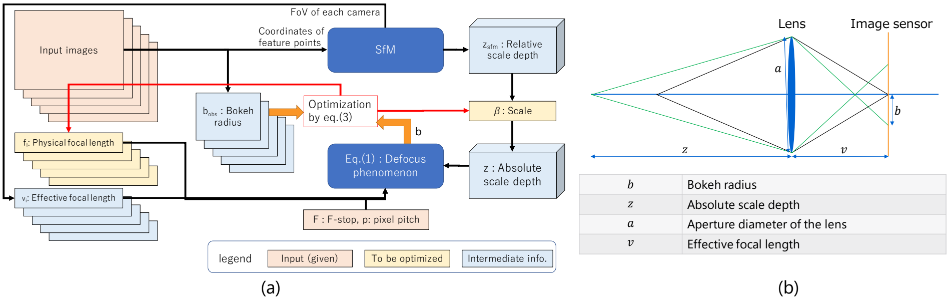
2 Related work

Depth estimation for monocular camera There are four major methods of depth estimation: based on motion parallax, shading, defocus, and context. The shading-based method utilizes the shading produced by additional light sources [24, 51]. The defocus-based method, which estimates depth from two images taken with a different focus, requires a special device to be able to accurately obtain the distance of image focus [9, 13, 32, 38, 41, 42, 46]. A typical method based on motion parallax is SfM [2, 45, 47]. When performing monocular SfM, both the camera positions and the 3D points are optimized based on geometric constraints between multiple viewpoints. The context-based method uses deep learning to directly regress the target depth based on the context in the image [12, 15, 16, 21, 34]. Recently, several methods based on physical cues have been proposed [8, 10, 18, 20, 27, 28].

Absolute scale estimation Application-specific studies are carried out by utilizing prior knowledge: known 3D model[35], the height of the camera installation and the flatness of the ground[17, 36, 37, 53] and the size of known object [4, 5, 19, 22, 40]. The other approach is to utilize additional sensors, such as stereo cameras [7, 11, 14, 33] and inertial sensors [25, 26, 30, 31, 43, 48]. The combination of context and SfM has also been used to estimate the absolute scale of monocular SfM [44, 49]. For monocular SfM, no method can estimate the absolute scale based solely on the monocular camera without any context information, prior knowledge, and additional sensors. Physical cue-based methods need pre-calibration for the absolute scale estimation [20]. No method can be used in scenarios where pre-calibration would be impossible, such as with varifocal lenses.

3 Convex regression for scale

Conventionally, to acquire the absolute scale in DfD, the physical focal length f and the effective focal length v need to be calibrated in advance. In contrast, we have made it possible to simultaneously optimize an unknown scale β and the unknown f by associating the scene structure of the geometric side and the bokeh radius of the physical side. The schematic diagram of our method is shown in Figure 2 (a). The relative scene structure z_{sfm} from SfM, called SfM depth, can be converted to a bokeh radius based on the defocus model via the unknown β, the unknown f, and v estimated by SfM as the FoV. Our method can optimize the above unknown parameters so that the converted bokeh radius matches the bokeh radius observed from the image at the same point. The above matching problem can be formulated as a nonlinear regression problem. Moreover, in this section, we show that the above nonlinear regression problem is a convex function, i.e., its unique local minimum is equal to the



b	Bokeh radius
z	Absolute scale depth
a	Aperture diameter of the lens
v	Effective focal length

Figure 2: (a) Schematic diagram of our method and (b) defocus model.

global minimum and the problem can be solved easily.

Optimization function: We use the following defocus model, introduced in [32], and shown in Figure 2 (b): $z = \frac{fv}{v-f-2pbF}$, where b is the bokeh radius, z is the absolute scale depth, $a = f/F$ is the aperture diameter of the lens, and F and p are F stop and pixel pitch, respectively. We define the bokeh radius on the near side to be negative and on the far side to be positive. Then, we rearrange the above equation to find the bokeh radius

$$b = \frac{fv}{2pF}\left(\frac{1}{f} - \frac{1}{z} - \frac{1}{v}\right). \tag{1}$$

Using SfM depth z_{sfm} and unknown scale factor β, the absolute scale depth can be represented as $z = \beta z_{sfm}$. Thus, (1) can be expressed as the following function with f and β as the unknown parameters:

$$b(z_{sfm}; f, \beta) = \frac{fv}{2pF}\left(\frac{1}{f} - \frac{1}{\beta z_{sfm}} - \frac{1}{v}\right). \tag{2}$$

Let b_{obs} be the observed bokeh radius. To estimate the bokeh radius, we use CNN framework [20]. As b_{obs} and $b(z_{sfm}; f, \beta)$ should be equal at the same point of the object, the following regression problem can be derived:

$$\hat{f}_i, \hat{\beta}, (\forall i \in \mathbf{I}) = \arg \min_{f_i, \beta, (\forall i \in \mathbf{I})} \sum_{i \in \mathbf{I}} \sum_{b_{obs}, z_{sfm}} L\left(b_{obs} - \frac{f_i v_i}{2pF}\left(\frac{1}{f_i} - \frac{1}{\beta z_{sfm}} - \frac{1}{v_i}\right)\right), \tag{3}$$

where L is an arbitrary distance function, i is the index number of the captured images, and the index set $\mathbf{I}$ is $\{0, \cdots, N-1\}$. As the images are acquired by uncalibrated cameras, such as varifocal lenses, we assume that f_i is different for each image. This is a nonlinear regression problem for f_i and β. In the following, we show that the subproblems divided into each image are convex and the original problem in (3), which combines them, is also convex.

Showing convexity: To simplify the problem, let us first consider finding a solution for each image, and use the L2 norm as follows:

$$\hat{f}_i, \hat{\beta}_i = \arg \min_{f_i, \beta_i} \sum_{b_{obs}, z_{sfm}} \left(b_{obs} - \frac{f_i v_i}{2pF}\left(\frac{1}{f_i} - \frac{1}{\beta_i z_{sfm}} - \frac{1}{v_i}\right)\right)^2, \tag{4}$$

where β_i is the per-image scale parameter. After applying the variable transformation $d_{sfm} =$

$\frac{1}{z_{sfm}}$, (4) becomes the following linear regression problem:

$$\hat{a_0}_i, \hat{a_1}_i = \arg \min_{a_{0i}, a_{1i}} \sum_{b_{obs}, d_{sfm}} \left(b_{obs} - (a_{0i} + a_{1i} d_{sfm})\right)^2 . \tag{5}$$

With respect to the solutions $\hat{a_0}_i$ and $\hat{a_1}_i$, the following relations are obtained from (4) and (5):

$$\begin{cases} \hat{a_0}_i = \frac{v_i}{2pF} - \frac{f_i}{2pF} \\ \hat{a_1}_i = -\frac{f_i v_i}{2pF\beta_i} \end{cases} . \tag{6}$$

The analytical solution $\hat{f_i}$ and $\hat{\beta_i}$ can be obtained by transforming the above equation. This solution is guaranteed to be a global minimum because (5) is a linear regression problem.

Next, we derive the original nonlinear regression problem in (3) from the per-image linear regression problem in (5). Concatenating the per-image regression problem in (5) and introducing a regularization term such that the per-image scale β_i gets closer to the same value as each other, the following equation is obtained:

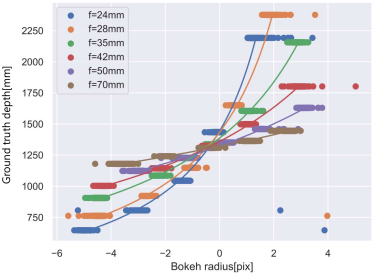
$$\hat{f_i}, \hat{\beta_i}, (\forall i \in \mathbf{I}) = \arg \min_{f_i, \beta_i, (\forall i \in \mathbf{I})} \sum_{i \in \mathbf{I}} \left\{ \sum_{b_{obs}, d_{sfm}} \left(b_{obs} - (a_{0i} + a_{1i} d_{sfm})\right)^2 + \lambda \sum_{j \in \mathbf{I}} (\beta_i - \beta_j)^2 \right\} , \tag{7}$$

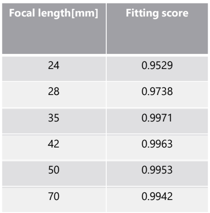
where $\lambda > 0$ is a balancing parameter of two terms. Thus, in the above equation, the first term on the right-hand side is convex, because it involves linear regression problems in (5). The second term is also convex because of its quadratic form, and thus, (7) is also convex. When λ is set to infinite, β_i for each image converges to the same value, i.e. the single scale value β. Then, in the case of L2 norm, (7) is equivalent to the original regression problem in (3). From the above discussion, the nonlinear regression problem in (3) can be solved easily by any nonlinear solver [6, 29] because of its convexity. In this paper, we use Trust Region Reflective [6] implemented in scikit-learn [1] as a nonlinear solver and Cauchy (aka Lorentzian) [3] as the distance function for reducing the effect of outliers. We will confirm the difference between L2 norm and Cauchy in the experiment. Once $\hat{\beta}$ is obtained, the absolute scale depth z can be obtained as $z = \hat{\beta} z_{SfM}$.

4 Experiment

To demonstrate the effectiveness of the proposed method, we conducted various experiments using a digital SLR camera (Nikon D810) and a varifocal lens (AF-S NIKKOR 24-70mm f/2.8G ED). Throughout all experiments, the F-number of the lens was fixed at 2.8. The captured images were saved as RAW data and resized to 1845x1232 pixels for the experiment. The pixel pitch p was calculated as 0.0195 based on the horizontal sensor size (35.9[mm]) and the horizontal number of pixels. We used COLMAP [39] to obtain a dense 3D point cloud from the multi-view image. Although COLMAP refers focal length based on EXIF information if accessible, we input images without EXIF which are converted from the RAW data. In this setting, COLMAP estimates reasonable focal length during SfM pipeline. We trained the CNN which estimates bokeh radius for each pixel [20] by fixing the focal length of the zoom lens at 48mm.

Input and output: The input of our experiment is multi-view images taken by the above camera and lens. The outputs are the absolute scale depth z and the estimated size of objects.





Focal length[mm]	Fitting score
24	0.9529
28	0.9738
35	0.9971
42	0.9963
50	0.9953
70	0.9942

Figure 3: Fitting result from the bokeh radius to the ground truth depth according to the defocus model in (1) for various focal lengths. The bokeh radius is predicted by the CNN trained on a single focal length at 48mm.

Ground truth: The ground truth is the size of the objects measured by a ruler.
Oracle (fixed setup): As mentioned in the related work, there is no method for estimating absolute scale using only a monocular camera in a setting where zoom and focus are variable for each captured image. In contrast, in a fixed lens setup, the absolute scale can be estimated by pre-calibrating the defocus cue. As an oracle, we utilize a single image DfD [20] with the fixed lens and full calibrated settings (**FixedOracle**). The single image DfD was calibrated with f and v in the defocus model (1) using the ground truth depth obtained by a chessboard [52]. From the absolute scale depth z_{dfd} obtained by the calibrated DfD, we find the scale parameter as $\beta = median(z_{dfd}/z_{sfm})$.
Metric: We evaluate the mean absolute error (MAE) between the estimated size of the objects and the ground truth. We also use the following metric to evaluate the error between the estimated value and the ground truth in the ablation study.

$$ErrRate(x, x_{GT}) = \begin{cases} (1 - \frac{x}{x_{GT}}) * 100, \text{ if } x < x_{GT} \\ (1 - \frac{x_{GT}}{x}) * 100, \text{ otherwise} \end{cases} . \quad (8)$$

Robustness of the bokeh extraction CNN trained with a single focal length: We tested how well the CNN trained at a single focal length would adapt to other focal lengths of the zoom lens. The CNN was trained with a focal length of only 48mm. Then, we took images at several different zoom-in increments. Thus, for each zoom, we took five images at different distances and fitted the blur radius predicted by the CNN to the ground truth depth using the defocus model in (1). The focal length and the effective focal length are the fitting parameters. The fitting results is shown in Figure 3. The fitting scores (coefficient of determination) are above 0.95 for all focal lengths. It can be confirmed that the CNN trained at a single focal length fits the defocus model well for various focal lengths.
Evaluation in the wild: To evaluate our method in the wild, we captured 11 outdoor scenes; each scene contained 9 images. The examples of the scenes are shown in Figure 4. This is a challenging setting that includes a variety of focal lengths within a single scene. To demonstrate the effectiveness of our method for the situations where pre-calibration was not possible, we randomly changed the zoom factor and autofocused on an object for each image. This is the same condition as the last row of Table 2 (**Ours**). The distance to the objects is about 4∼7[m]. The examples of the measurement point, the SfM depth, the bokeh radius, and the absolute scale depth are shown in Figure 5. The quantitative result is shown in Figure 6. The average MAE of all scenes is 23.035±4.014[mm], which is about 0.42%

error against the object distance from the camera. In the fixed lens setup, the average MAE of **FixedOracle** is 13.855±4.204[mm]. Although our method is not quite as good as **FixedOracle**'s, it achieves good accuracy under the difficult conditions of variable lens settings.

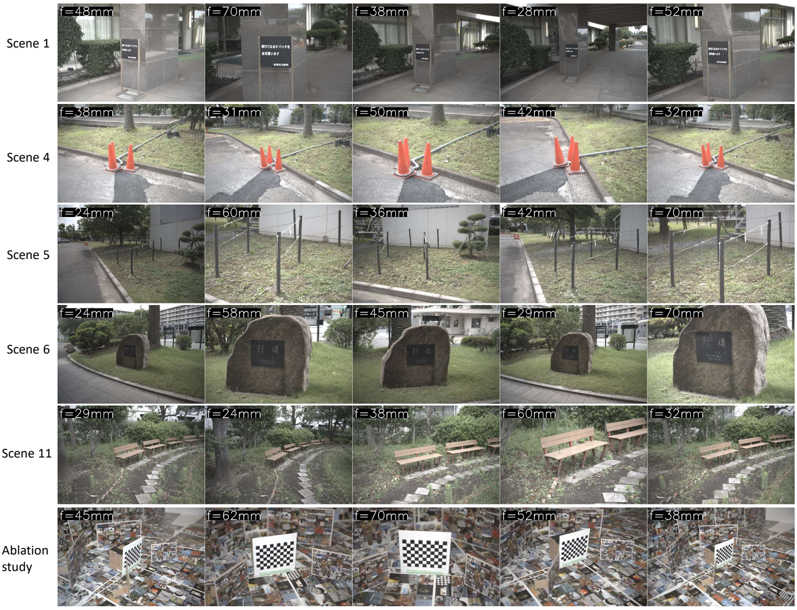


Figure 4: Examples of the multi-view images under both variable zoom and focus settings. The number in each image indicates its focal length. This is a challenging setting that includes a variety of focal lengths within a single scene. The last row is an example of the multi-view image for the ablation study.

Ablation study: For the ablation study, we captured several indoor multi-view images at a distance of about 1.0 [m] from a chessboard (each scene contains 9 multi-view images, and the baseline length between two views is about 15[cm]) as shown in Figure 4. We randomly sampled and resized photos from the dataset [23], and they were printed on some planes so that the absolute scale depth could not be determined from the context. To avoid using prior knowledge of the camera height, we captured sparse multi-view images by a handheld camera. The experimental setup is shown in Table 2. The conditions (zoom: Z, focus: F, distance function: L, and effective focal length: v) for our method are shown in the last row of Table 2 (**Ours**). The ablated conditions are appended to the method name. We define the inlier as the rate of pixels in which the absolute difference between b_{obs} and $b(z_{sfm}; f, \beta)$ is below the certain threshold corresponding to 100 [mm].

(1) FixedOracle vs. Ours_Z_F_v: We compare our method (**Ours_Z_F_v**) with **FixedOracle** under same conditions of **FixedOracle**. We evaluated the accuracy of β and the grid size of the chessboard estimated by **FixedOracle** and **Ours_Z_F_v**. Comparing β and the grid size of the chessboard of **FixedOracle** and **Ours_Z_F_v**, it can be seen that **FixedOracle** is very slightly more accurate as shown in Table 2. However, **FixedOracle** and **Ours_Z_F_v** have almost the same accuracy, and their errors are around 1%, which shows that our method is working very well, even without the need for calibration.

(2) L2 norm (Ours_Z_F_L_v) vs. Cauchy (Ours_Z_F_v): We verify the difference

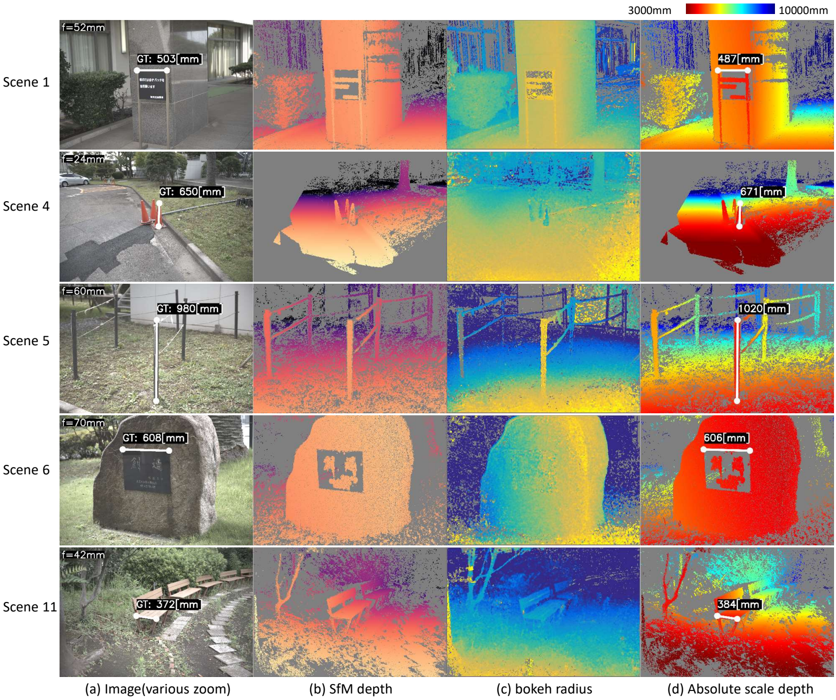


Figure 5: Examples of SfM depth, bokeh radius, and estimated absolute scale depth in the wild. The ground truth of the object size is overlaid on the captured image, and the corresponding estimate is overlaid on the absolute scale depth.

between L2 norm and Cauchy as the distance function in (3). In Figure 7, we show the error maps of (3). It can be seen that the error maps of both Cauchy and L2 norm indicate convexity. However, the solution of L2 norm deviated from the ground truth due to outliers. Comparing **Ours_Z_F_L_v** and **Ours_Z_F_v** in Table 2, we can see that L2 norm is much less accurate for both f, β, and the grid size than Cauchy.

(3) Calibrated v (Ours_Z_F_v) vs. estimated v (Ours_Z_F): We investigated whether the estimated v by SfM (**Ours_Z_F**) causes any accuracy degradation compared to the calibrated v (**Ours_Z_F_v**). Although v should not change in the fixed camera setting, the estimated v (**Ours_Z_F**) changes as shown in Figure 8(a). Comparing **Ours_Z_F_v** and **Ours_Z_F** in Table 2, we can see the accuracy of the estimated f of **Ours_Z_F** is slightly worse than that of **Ours_Z_F_v**. However, the estimated β of **Ours_Z_F** was better than that of **Ours_Z_F_v**, and the grid size was estimated almost the same accuracy. From the above results, we confirmed that the influence of the estimated v by SfM is small.

(4) Scalability to both variable zoom and focus: We checked the scalability for both variable zoom and focus. Under the variable zoom setting, we changed the zoom magnification randomly for each image. Under variable focus setting, the camera was autofocused on the chessboard in the center for each captured image. We evaluated the grid size of the chessboard because there is no way to know the ground truth of β and f under variable zoom and focus setting. The results are shown in Figure 8 and Table 2. We note that the variable zoom setting (**Ours_F** and **Ours**) is less accurate than the fixed zoom setting (**Ours_Z_F**

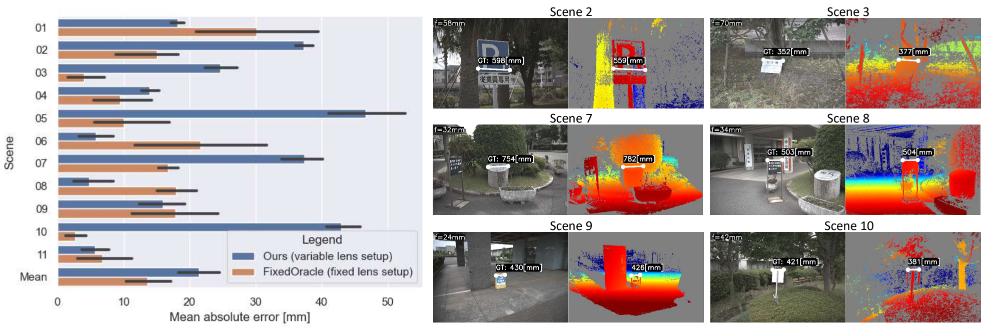


Figure 6: Mean absolute error between the estimated size and the ground truth for each scene in the wild, and captured images of the scenes out of Figure 5. The black solid line indicates the standard deviation of the MAE. The average MAE of all scenes is 23.035±4.014 [mm] at the distance of about 4∼7 [m] to each object, which is about 0.42% error against the object distance from the camera. In the fixed lens setup, the average MAE of FixedOracle is 13.855±4.204[mm].

Method	Zoom	Focus	L	v		↓Error rate[%]		↑Inlier
					f	β	Grid size	
FixedOracle [20]	Fix	Fix	-	Calib	Calibrated	0.779±0.080	1.087±0.79	0.809
Ours_Z_F_L_v	Fix	Fix	L2	Calib	0.122±0.047	3.514±0.931	3.595±1.364	0.824
Ours_Z_F_v	Fix	Fix	Cauchy	Calib	0.078±0.010	0.998±0.110	1.139±0.784	0.794
Ours_Z_F	Fix	Fix	Cauchy	SfM	0.242±0.208	0.606±0.350	1.134±0.806	0.794
Ours_Z	Fix	Variable	Cauchy	SfM	-	-	1.001±0.735	0.857
Ours_F	Variable	Fix	Cauchy	SfM	-	-	2.530±3.113	0.692
Ours	Variable	Variable	Cauchy	SfM	-	-	1.485±1.246	0.784

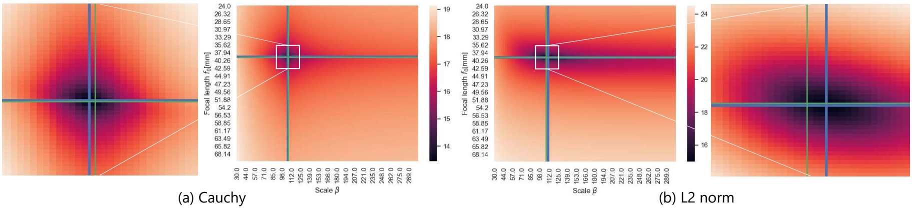
Table 2: The error rate of the focal length f, the scale β, and the grid size of the chessboard.

and **Ours_Z**). This is because the focal length (f=48mm) at which the CNN is trained is close to the focal length of the fixed zoom setting (f=42mm). Because of the nature of regression problems, a high inlier rate leads to high accuracy. Comparing the inlier ratio of the fixed zoom setting with that of the variable zoom setting, we can see that the former is higher. Furthermore, the variable focus (**Ours_Z** and **Ours**) has better accuracy than the fixed focus (**Ours_Z_F** and **Ours_F**). This is because the bokeh extraction CNN is highly accurate at near the focus [20]. In the variable focus setting, we found that all of the images were taken well. However, it was confirmed that in the autofocused setting, the error of the grid size is relatively high even with the variable zoom setting.

Effect of the number of images: To evaluate the effect of the number of images on the accuracy, we experimented with varying the number of images in **Ours_Z_F** setting. As shown in Figure 9, The accuracy of the estimated grid size deteriorates as the number of images decreases. The error is especially large for 2 or 3 images, and for practical use, 5 or more images are desirable.

Robustness to textureless surfaces: Even if the scene has a lot of texture less area, our method can estimate absolute scale by bokeh radius and SfM. As shown in Figure 10, we confirmed that the absolute scale can be obtained even in scenes with few textures.

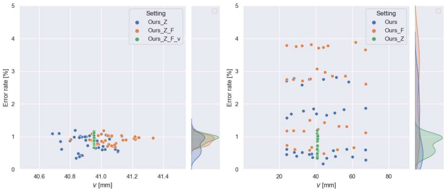
Limitations: Our method degenerates accuracy for pan-focus images, where bokeh cues are difficult to obtain. It is also inaccurate in scenes where the depth is concentrated in one



(a) Cauchy

(b) L2 norm

Figure 7: The error map of (3). The horizontal axis is β, the vertical axis is f_0, and the heat map represents the error value. Solid green and blue lines indicate the ground truth and estimated values of β and f_0, respectively.



(a) Fixed zoom (f=42mm)

(b) Variable zoom and focus

Figure 8: Plots of the effective focal length v vs. the error rate of the estimated grid size of the chessboard. Each dot represents v and the error rate for each image.

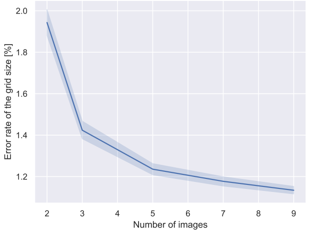
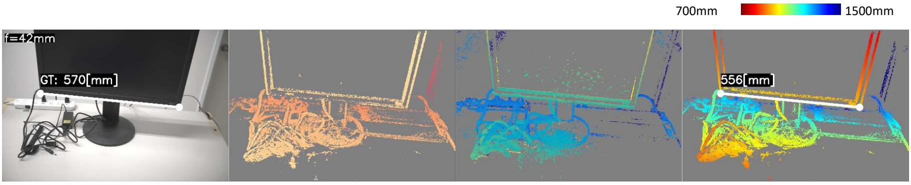


Figure 9: Plot of the number of images vs. the error rate of the estimated grid size of the chessboard.

place due to poor regression accuracy.

5 Conclusion

In this paper, we have proposed a novel method to obtain absolute scales by combining monocular SfM and bokeh cues without making any use of context, prior knowledge, additional sensors, and pre-calibration. Since our method can optimize unknown focal length as well as scale, it can be applied to uncalibrated cameras. We have formulated the above optimization problem as a nonlinear regression problem. Moreover, we have shown that the above non-linear regression problem is convex. To demonstrate the effectiveness of the proposed method, we conducted various experiments using a varifocal lens.



(a) Image

(b) SfM depth

(c) bokeh radius

(d) Absolute scale depth

Figure 10: Results for the textureless scene. Our method can estimate absolute scale. Estimated and GT of the monitor width are 556 mm and 570 mm, respectively.

References

[1] scikit-learn. https://scikit-learn.org/.

[2] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building Rome in a day. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 72–79. IEEE, 2009.

[3] Michael J Black and Paul Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996.

[4] Tom Botterill, Steven Mills, and Richard Green. Bag-of-words-driven, single-camera simultaneous localization and mapping. *Journal of Field Robotics*, 28(2):204–226, 2011.

[5] Tom Botterill, Steven Mills, and Richard Green. Correcting scale drift by object recognition in single-camera slam. *IEEE transactions on cybernetics*, 43(6):1767–1780, 2012.

[6] Mary Ann Branch, Thomas F Coleman, and Yuying Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, 21(1):1–23, 1999.

[7] Martin Buczko and Volker Willert. How to distinguish inliers from outliers in visual odometry for high-speed automotive applications. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 478–483. IEEE, 2016.

[8] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10193–10202, 2019.

[9] Subhasis Chaudhuri and Ambasamudram N Rajagopalan. *Depth from defocus: a real aperture imaging approach.* Springer Science & Business Media, 2012.

[10] Eric Cristofalo and Zijian Wang. Out-of-focus: Learning depth from image bokeh for robotic perception. *arXiv preprint arXiv:1705.01152*, 2017.

[11] Igor Cvišić and Ivan Petrović. Stereo odometry based on careful feature selection and tracking. In *2015 European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2015.

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[13] Paolo Favaro. Recovering thin structures via nonlocal-means regularization with application to depth from defocus. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1133–1140. IEEE, 2010.

[14] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3D reconstruction in real-time. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 963–968. IEEE, 2011.

[15] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017.

[16] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.

[17] Johannes Gräter, Tobias Schwarze, and Martin Lauer. Robust scale estimation for monocular visual odometry using structure from motion and vanishing points. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 475–480. IEEE, 2015.

[18] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2019.

[19] Sebastian Hilsenbeck, Andreas Möller, Robert Huitl, Georg Schroth, Matthias Kranz, and Eckehard Steinbach. Scale-preserving long-term visual odometry for indoor navigation. In *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–10. IEEE, 2012.

[20] Masako Kashiwagi, Nao Mishima, Tatsuo Kozakaya, and Shinsaku Hiura. Deep depth from aberration map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4070–4079, 2019.

[21] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[22] Hyon Lim and Sudipta N Sinha. Monocular localization of a moving person onboard a quadrotor MAV. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 2182–2189. IEEE, 2015.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[24] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1052–1061, 2019.

[25] Todd Lupton and Salah Sukkarieh. Removing scale biases and ambiguity from 6DoF monocular SLAM using inertial. In *2008 IEEE International Conference on Robotics and Automation*, pages 3698–3703. IEEE, 2008.

[26] Agostino Martinelli. Vision and imu data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination. *IEEE Transactions on Robotics*, 28(1):44–60, 2011.

[27] Maxim Maximov, Kevin Galim, and Laura Leal-Taixé. Focus on defocus: bridging the synthetic to real domain gap for depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1071–1080, 2020.

[28] Nao Mishima, Tatsuo Kozakaya, Akihisa Moriya, Ryuzo Okada, and Shinsaku Hiura. Physical cue based depth-sensing by color coding with deaberration network. In *BMVC2019*, 2019.

[29] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[30] Ryo Nakashima and Akihito Seki. Uncertainty-based adaptive sensor fusion for visual-inertial odometry under various motion characteristics. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3119–3125. IEEE, 2020.

[31] Gabriel Nützi, Stephan Weiss, Davide Scaramuzza, and Roland Siegwart. Fusion of imu and vision for absolute scale estimation in monocular slam. *Journal of intelligent & robotic systems*, 61(1):287–299, 2011.

[32] Alex Paul Pentland. A new sense for depth of field. *IEEE transactions on pattern analysis and machine intelligence*, (4):523–531, 1987.

[33] Mikael Persson, Tommaso Piccini, Michael Felsberg, and Rudolf Mester. Robust stereo visual odometry from monocular techniques. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 686–691. IEEE, 2015.

[34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019.

[35] DAN Rosenholm and KENNERT TORLEGARD. Three-dimensional absolute orientation of stereo models using digital elevation models. *Photogrammetric engineering and remote sensing*, 54(10):1385–1389, 1988.

[36] Davide Scaramuzza and Roland Siegwart. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE transactions on robotics*, 24(5):1015–1026, 2008.

[37] Davide Scaramuzza, Friedrich Fraundorfer, Marc Pollefeys, and Roland Siegwart. Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In *2009 IEEE 12th international conference on computer vision*, pages 1413–1419. IEEE, 2009.

[38] Yoav Y Schechner and Nahum Kiryati. Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, 39(2):141–162, 2000.

[39] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[40] Shiyu Song, Manmohan Chandraker, and Clark C Guest. High accuracy monocular sfm and scale correction for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):730–743, 2015.

[41] Murali Subbarao and Gopal Surya. Depth from defocus: A spatial domain approach. *International Journal of Computer Vision*, 13(3):271–294, 1994.

[42] Huixuan Tang, Scott Cohen, Brian Price, Stephen Schiller, and Kiriakos N Kutulakos. Depth from defocus in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2740–2748, 2017.

[43] Jean-Philippe Tardif, Michael George, Michel Laverne, Alonzo Kelly, and Anthony Stentz. A new approach to vision-aided inertial navigation. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4161–4168. IEEE, 2010.

[44] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6243–6252, 2017.

[45] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9(2):137–154, 1992.

[46] Masahiro Watanabe and Shree K Nayar. Rational filters for passive depth from defocus. *International Journal of Computer Vision*, 27(3):203–225, 1998.

[47] Matthew J Westoby, James Brasington, Niel F Glasser, Michael J Hambrey, and Jennifer M Reynolds. ' structure-from-motion ' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314, 2012.

[48] Minjun Xiong, Huimin Lu, Dan Xiong, Junhao Xiao, and Ming Lv. Scale-aware monocular visual-inertial pose estimation for aerial robots. In *2017 Chinese Automation Congress (CAC)*, pages 7030–7034. IEEE, 2017.

[49] Xiaochuan Yin, Xiangwei Wang, Xiaoguo Du, and Qijun Chen. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields. In *Proceedings of the IEEE international conference on computer vision*, pages 5870–5878, 2017.

[50] Ji Zhang and Sanjiv Singh. Visual-lidar odometry and mapping: Low-drift, robust, and fast. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2174–2181. IEEE, 2015.

[51] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999.

[52] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[53] Dingfu Zhou, Yuchao Dai, and Hongdong Li. Ground-plane-based absolute scale estimation for monocular visual odometry. *IEEE Transactions on Intelligent Transportation Systems*, 21(2): 791–802, 2019.