Calibration-free Structure-from-Motion with Calibrated Radial Trifocal Tensors Supplementary Material

Viktor Larsson[1], Nicolas Zobernig[2], Kasim Taskin[3], and Marc Pollefeys[1,4]

[1] Department of Computer Science, ETH Zürich,
[2] Dept. of Information Technology and Electrical Engineering, ETH Zürich,
[3] KTH Royal Institute of Technology
[4] Microsoft Mixed Reality & AI Zurich Lab

1 Supplementary Material

In the supplementary material we present more details and reconstruction results. We also provide two videos `supp_init.mp4` and `supp_results.mp4` which show more qualitative results. The files `internal_radial.m` and `internal_mixed.m` are MATLAB scripts which verifies the internal constraints for the radial trifocal tensor and the mixed trifocal tensor.

2 The 1D Radial Camera Model

In this section we present additional explanations for the 1D radial camera model and how it relates to the regular pinhole camera model. We also clarify why the assumption of square pixels and known principal point is necessary.

Consider first a standard pinhole camera $P = K[R\ \mathbf{t}]$, which maps 3D points (in homogeneous coordinates) to image points (in homogeneous coordinates). The matrix K encodes the camera's intrinsic parameters; focal length f, aspect ratio α, skew s and principal point (u_x, u_y),

$$K = \begin{bmatrix} f & s & u_x \\ 0 & \alpha f & u_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

If the principal point (u_x, u_y) is known, we can center the image coordinate system around this point. The new camera matrix corresponding to the centered image points then has principal point in the origin, i.e. $u_x = u_y = 0$.

Let $\mathbf{x} \in \mathbb{R}^2$ be the 2D projection of the 3D point $\mathbf{X} \in \mathbb{R}^3$, i.e.

$$\mathbf{x} = \frac{1}{Z}\left(fX + sY, \alpha f Y\right)^T \quad \text{where} \quad (X, Y, Z)^T = R\mathbf{X} + \mathbf{t} \tag{2}$$

For square pixels, i.e. unit aspect ratio ($\alpha = 1$) and zero skew ($s = 0$), we get

$$\mathbf{x} = \frac{1}{Z}\left(fX,\, fY\right)^T = \frac{f}{Z}\left(X, Y\right)^T \tag{3}$$

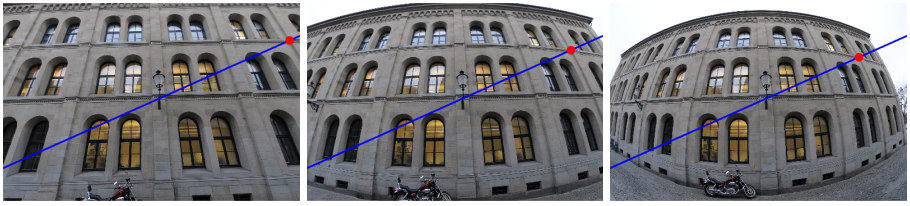

Fig. 1. Radial distortion moves points along the radial lines. The red point moves along the radial line (blue) as the image becomes radially distorted. In the 1D radial camera model the 3D points projects to radial lines which should then pass through the 2D image points. Since this only depends on the direction of the projection in the image plane, and not the radial offset, it is invariant to focal length and radial distortion.

The main idea in the Radial Alignment Constraints (RAC) [8], is that instead of requiring equality in (3), we only require that the vectors are parallel. Geometrically, this restricts the projection of the 3D point lie somewhere on the radial line passing through $\mathbf{x}$. This constraint does not depend on the focal length since,[5]

$$\mathbf{x} \sim \frac{f}{Z}\left(X,Y\right)^T \sim \left(X,Y\right)^T, \tag{4}$$

Similarly any radial distortion simply scales the projection radially. See Figure 1 for an illustration. The vector (X,Y) only depend on the top two rows of $[R\ \mathbf{t}]$,

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{bmatrix} \mathbf{r}_1^T & t_1 \\ \mathbf{r}_2^T & t_2 \end{bmatrix} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix} \tag{5}$$

In the 1D radial camera model (see e.g. [7]) we then interpret this as a camera that projects 3D points onto radial lines (i.e. lines passing through the image center). Similar to a pinhole we can represent this with a matrix, though in this case a 2×4-matrix. In general a 2×4 matrix P_{rad} can be factorized (up to scale) as $P_{rad} = K_{2\times 2}[R_{2\times 3}\ \mathbf{t}_{2\times 1}]$ where

$$K_{2\times 2} = \begin{bmatrix} 1 & s \\ 0 & \alpha \end{bmatrix} \tag{6}$$

Note that the focal length simply scales the entire matrix P_{rad} and can thus be ignored here. Now under the assumption of square pixels we have

$$P_{rad} = \begin{bmatrix} \mathbf{r}_1^T & t_1 \\ \mathbf{r}_2^T & t_2 \end{bmatrix} \tag{7}$$

which is what we call a *calibrated radial camera* in the paper.

[5] Here $\sim$ denotes equality up to scale.

3 Finding the Internal Constraints

To find the internal constraints on the trifocal tensors we leverage the fact that it is very easy to generate random instances of calibrated tensors. This can be done by simply generating cameras of the correct form and computing the corresponding tensor. Let $\mathcal{M}_d(T)$ denote the vector of all degree d monomials in the elements of the tensor T. Then any polynomial of degree d in T can be written as

$$p = \sum_{k=0}^{d} \mathbf{c}_k^T \mathcal{M}_k(T). \tag{8}$$

Note that since the tensors are only defined up to scale, any internal constraint must be a homogeneous polynomial. Thus we only need to consider single terms in the sum above. Now, given a specific tensor T, there are of course infinitely many choices of coefficients $\mathbf{c}$ that satisfy (8). To find constraints that are satisfied for all tensors, we generate multiple random calibrated trifocal tensors $T_1, \ldots, T_N$. Then if there exists a degree d internal constraint we must have

$$\mathbf{c}^T \left[\mathcal{M}_d(T_1) \ \ldots \ \mathcal{M}_d(T_N) \right] = 0. \tag{9}$$

If we choose N sufficiently large, the matrix with the stacked tensor monomials in (9) becomes square and we can simply find the coefficients $\mathbf{c}$ by computing the left nullspace, e.g. using SVD.

For the two tensors considered in the main paper, the radial trifocal tensor and the mixed trifocal tensor, we used this technique to find a degree 4 constraint (radial) and a degree 8 (mixed) constraint. For the radial trifocal tensor (which has 8 elements) the matrix for degree 4 was 330×330. For the mixed trifocal tensor (which has 12 elements) the matrix for degree 8 is 75782×75782. Note that this is a dense matrix (≈ 42 GB).

For the mixed trifocal tensor there is a degree 6 constraint which is satisfied regardless of calibration. For $d > 6$ there are additional vectors in the nullspace corresponding to multiples of the degree 6 constraint. To avoid this problem we add constraints that $\mathbf{c}$ should be orthogonal these multiples. This just adds additional homogeneous linear constraints to (9).

3.1 Internal Constraint on the Radial Trifocal Tensor

The internal constraint for the calibrated radial trifocal tensor (for cameras with intersecting principal axes) is shown below

$$T_{111}^3T_{222} - T_{111}^2T_{211}T_{122} - T_{111}^2T_{121}T_{212} - T_{111}^2T_{221}T_{112} + T_{111}T_{211}^2T_{222} + 2T_{111}T_{211}T_{121}T_{112} - 2T_{111}T_{211}T_{221}T_{212} + T_{111}T_{121}^2T_{222} - 2T_{111}T_{121}T_{221}T_{122} - T_{111}T_{221}^2T_{222} + T_{111}T_{112}^2T_{222} - 2T_{111}T_{112}T_{212}T_{122} - T_{111}T_{212}^2T_{222} - T_{111}T_{122}^2T_{222} - T_{111}T_{222}^3 - T_{211}^3T_{122} + T_{211}^2T_{121}T_{212} + T_{211}^2T_{221}T_{112} + T_{211}T_{121}^2T_{122} + 2T_{211}T_{121}T_{221}T_{222} - T_{211}T_{221}^2T_{122} + T_{211}T_{112}^2T_{122} + 2T_{211}T_{112}T_{212}T_{222} - T_{211}T_{212}^2T_{122} + T_{211}T_{122}^3 + T_{211}T_{122}T_{222}^2 - T_{121}^3T_{212} + T_{121}^2T_{221}T_{112} - T_{121}T_{221}^2T_{212} + T_{121}T_{112}^2T_{212} + 2T_{121}T_{112}T_{122}T_{222} + T_{121}T_{212}^3 - T_{121}T_{212}T_{122}^2 + T_{121}T_{212}T_{222}^2 + T_{221}^3T_{112} - T_{221}T_{112}^3 - T_{221}T_{112}T_{212}^2 - T_{221}T_{112}T_{122}^2 + T_{221}T_{112}T_{222}^2 - 2T_{221}T_{212}T_{122}T_{222} = 0.$$

4 Factorizing the Tensors

The factorization of the radial tensor into three projective cameras is detailed in [5]. Similarly, using the ideas from [2], we can factorize the mixed radial tensor. See Figure 2 for one possible factorization of the mixed $3 \times 2 \times 2$ tensor. Note that these factorizations degenerate for certain tensors. To avoid this problem we perform a random projective change of coordinates in each of the images before performing the factorization.

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a_1 & a_2 & a_1 & a_2 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ b_1 & b_2 & b_3 & b_4 \end{bmatrix}$$

$$P_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ c_1 & c_2 & c_3 & c_4 \end{bmatrix}$$

$$\begin{aligned} a_1 &= -T_{122} \\ a_2 &= -T_{222} \\ b_1 &= (T_{112} - T_{122}T_{312})/T_{122} \\ b_2 &= (T_{212} - T_{222}T_{312})/T_{122} \\ b_3 &= -T_{312} \\ b_4 &= x_2 \\ c_1 &= (T_{121} - T_{122}T_{321})/T_{222} \\ c_2 &= (T_{221} - T_{222}T_{321})/T_{222} \\ c_3 &= x_1 \\ c_4 &= -T_{321} \end{aligned}$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = A^{-1}b, \quad \text{where} \quad A = \begin{bmatrix} T_{122}T_{222}^2T_{312} - T_{112}T_{222}^2 & T_{122}^3T_{321} - T_{121}T_{122}^2 \\ T_{222}^3T_{312} - T_{212}T_{222}^2 & T_{122}^2T_{222}T_{321} - T_{122}^2T_{221} \end{bmatrix}$$

$$b = \begin{pmatrix} T_{122}^2T_{222}(T_{311} - T_{312}T_{321}) - T_{111}T_{122}T_{222} - T_{122}^2T_{222}T_{312}T_{321} + T_{112}T_{122}T_{222}T_{321} + T_{121}T_{122}T_{222}T_{312} \\ T_{122}T_{222}^2(T_{311} - T_{312}T_{321}) - T_{122}T_{211}T_{222} - T_{122}T_{222}^2T_{312}T_{321} + T_{122}T_{212}T_{222}T_{321} + T_{122}T_{221}T_{222}T_{312} \end{pmatrix}$$

Fig. 2. Factorizing the Mixed Trifocal Tensor. Assumes that the tensor is normalized such that $T_{322} = 1$.

4.1 Metric Upgrade

The cameras obtained from the factorization in the previous section are not calibrated. The internal constraints for calibration on the tensor ensure that we can get a calibrated set of cameras by applying a projective transform. Since we assume zero skew and unit aspect ratio the projective transform can be linearly estimated as described in [7].

5 Planar Scene under Radial Projections

In this section we discuss the third case considered for the radial trifocal tensor in [7], when the scene points are planar. As mentioned in the main paper (and in [7]), for the case of intersecting principal axes we can only recover the direction to the 3D point. Since directions correspond to points on the plane at infinity $\boldsymbol{\pi}_\infty$, we can think of the cameras as only imaging this plane. Of course, then by a projective change of coordinates the cameras could be imaging any other plane instead. Now assume the cameras are observing a real plane, which we w.l.o.g. assume is the xy-plane, i.e. $\mathbf{X} = (x, y, 0, 1)$. The radial trifocal tensor describing this camera configuration then only gives us information about the 1,2 and 4 column of the camera matrices (since all imaged points have $z = 0$), i.e. factorizing the tensor we only get

$$P_i = \begin{bmatrix} a_{11} & a_{12} & ? & a_{14} \\ a_{21} & a_{22} & ? & a_{24} \end{bmatrix}, \quad i = 1, 2, 3. \tag{10}$$

Requiring the cameras to be calibrated we can recover the unknown elements a_{13} and a_{23} using the equations

$$a_{11}^2 + a_{12}^2 + a_{13}^2 = a_{21}^2 + a_{22}^2 + a_{23}^2, \tag{11}$$

$$a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23} = 0. \tag{12}$$

Note that we have two unknowns and two equations, so we can always extend the 2×2 block to a calibrated reconstruction. Since we can always do this (assuming the 2×2-block is non-zero) this shows that there does not exist any internal constraint for calibration in the case of planar scene points. In fact it is easy to see that there exist infinitely many calibrated planar reconstructions consistent with a radial trifocal tensor, since we can do this for any projective coordinate change which keeps the xy-plane fixed.

5.1 Reconstruction Ambiguities

For the 1D radial camera model it is not possible to determine if a 3D point is behind or in front the camera since the depth is not observable. This means that there is an unresolvable ambiguity in the reconstruction where we can mirror it in the coordinate axes (i.e. $H = \text{diag}(1, 1, 1, -1)$). Additionally, only fixing the first camera to $[I_2 \ 0]$ leaves a possible reflection in the z-axis (i.e. $H = \text{diag}(1, 1, -1, 1)$).

6 More Experimental Results

Now we present more qualitative results for our reconstruction pipeline. Figures 3, 4, 5, 6, and 7 shows reconstructions from the quantitative evaluation in the main paper.

In Figure 8 we show a comparison with running vanilla COLMAP [6] with default parameters on the fisheye images from [1]. In [1] they also present catadioptric images of the same building. Unfortunately, due to the limited motion in this dataset and the difficulty in matching the catadioptric images, we were unable to get a nice reconstruction using only catadioptric images. In Figure 9 we present a reconstruction using both the fisheye and the catadioptric images. While the reconstruction is noisier, you can also see how more features are being triangulated, e.g. the two smaller structures on the sides as well as the building/trees across the street. Note that these structures are not visible in the fisheye images. To get more accurate matches for the catadioptric images we used the same partial undistortion described in [1] before computing the feature descriptors.

Figure 11 and Figure 12 show two new datasets, *Big Church* and *Building II*, captured using a standard DSLR camera with a fisheye lens. Figure 13 shows a reconstruction from a new dataset captured with a Ricoh Theta Z1 camera. The camera has two 180 degree fisheye lenses pointing in opposite directions to get a omnidirectional field-of-view.

In Figure 14 we show two failure cases for the *Spilled Blood* and the *Big Church* dataset. In both of these cases the incremental SfM method incorrectly registered images to the reconstruction. The reconstructions in Figure 6 and Figure 11 were found by using a different set of images to initialize.

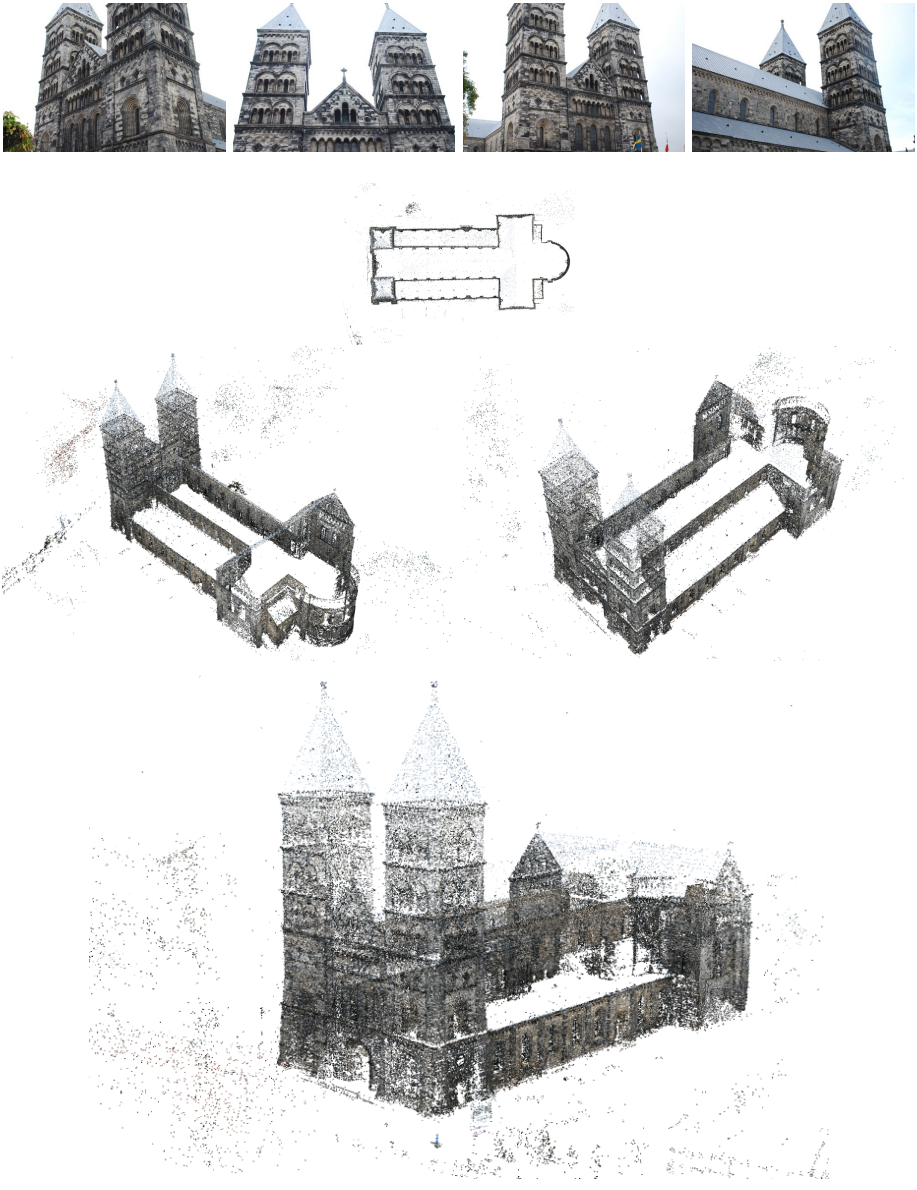

Fig. 3. *Lund Cathedral* [4] 1226 images, 422939 points, 0.29 px average reprojection error.

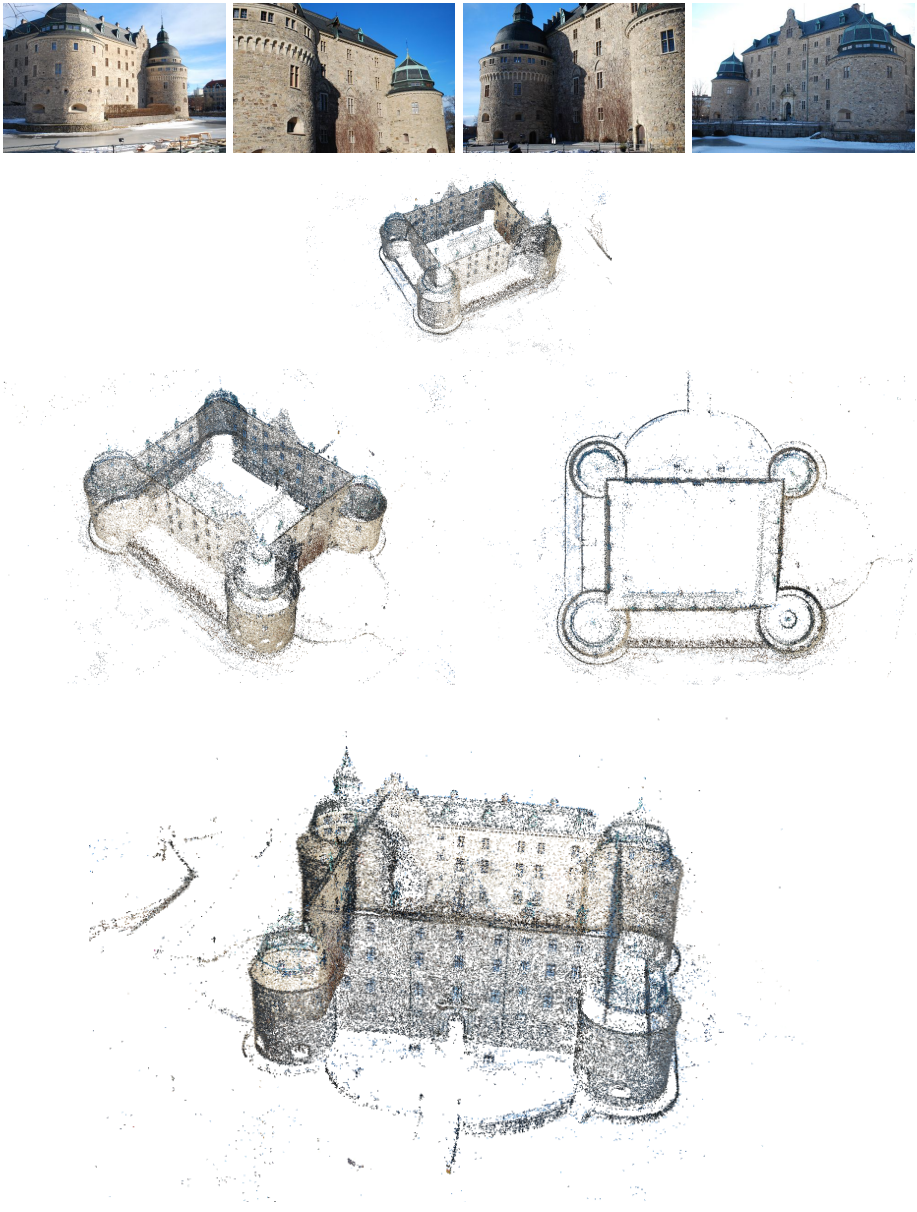

Fig. 4. *Orebro Castle* [4] 763 images, 197300 points, 0.28 px average reprojection error. Note that the actual building is not exactly rectangular.

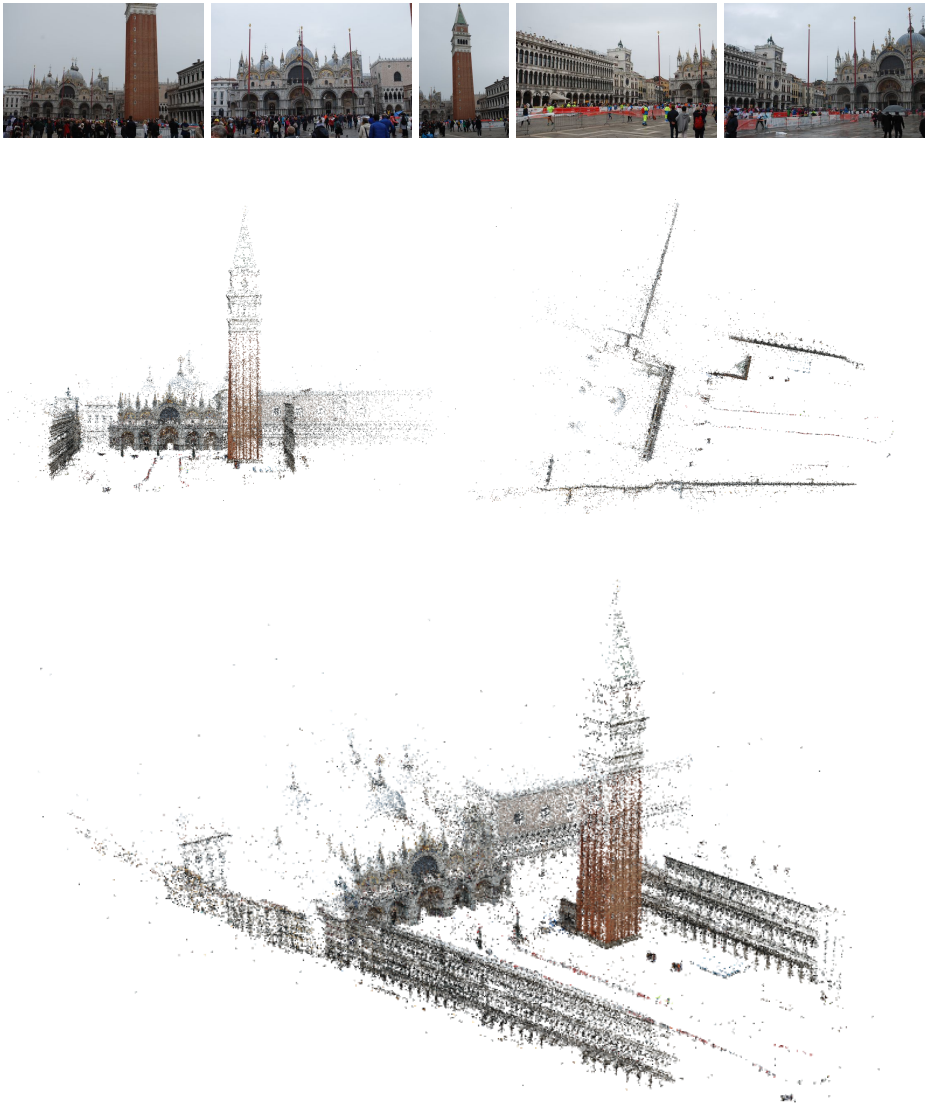

Fig. 5. *San Marco* [4], 1498 images, 293014 points, 0.44 px average reprojection error.

Fig. 6. *Spilled Blood* [4] 781 images, 284979 points, 0.41 px average reprojection error.

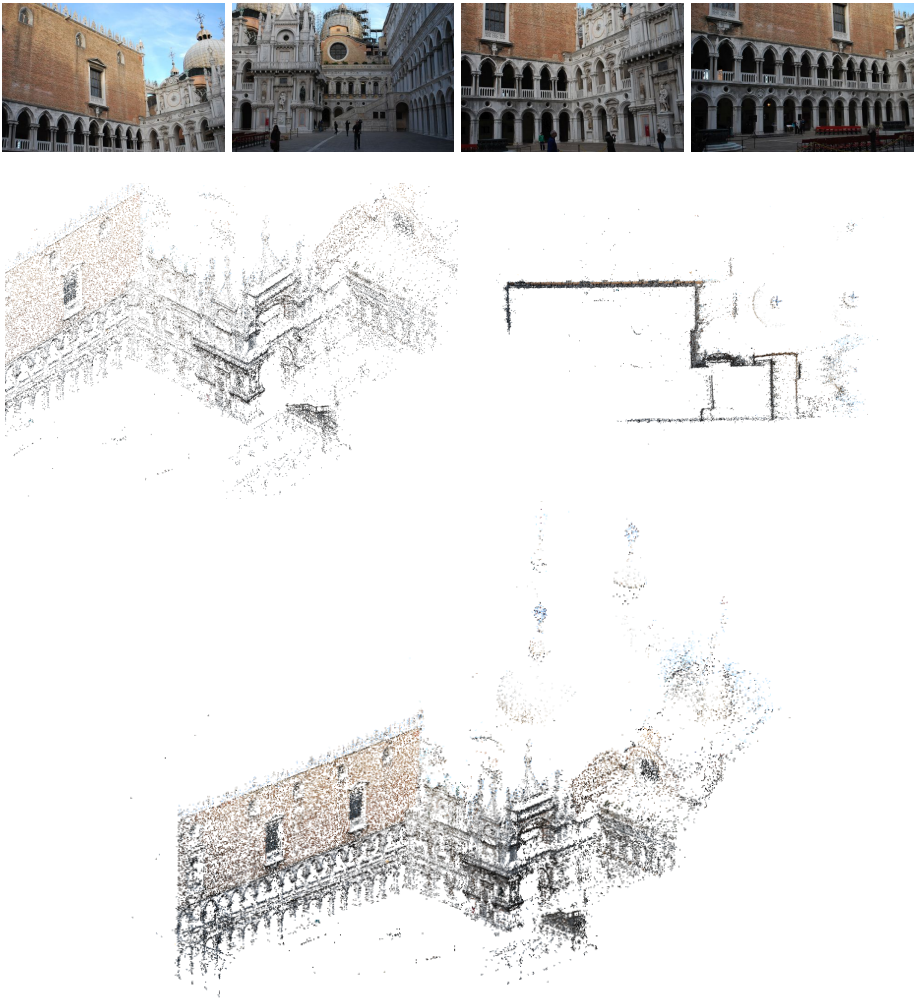

Fig. 7. *Doge Palace* [4] 241 images, 74302 points, 0.29 px average reprojection error.

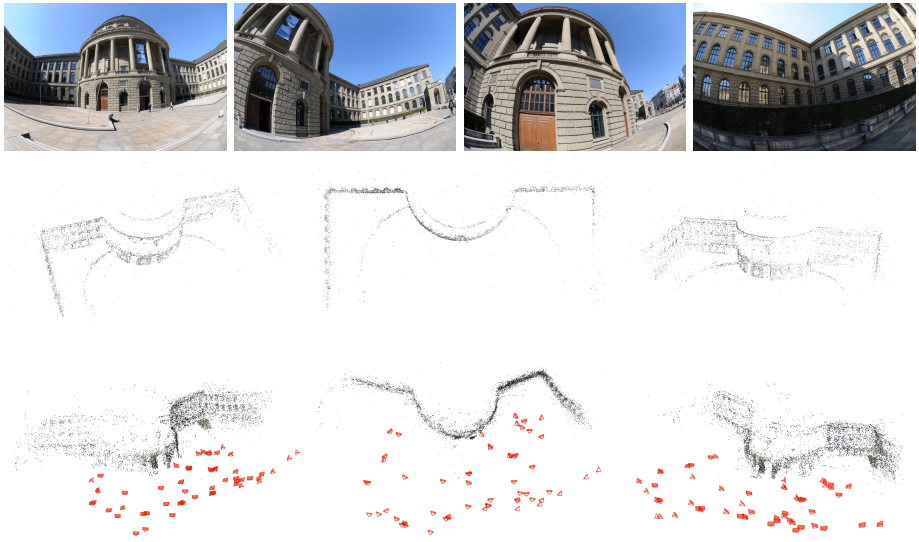

Fig. 8. *Building dataset* using Fisheye images from Camposeco et al. [1]. *Top:* Our reconstruction. *Bottom:* COLMAP [6] without camera intrinsic/distortion parameters.

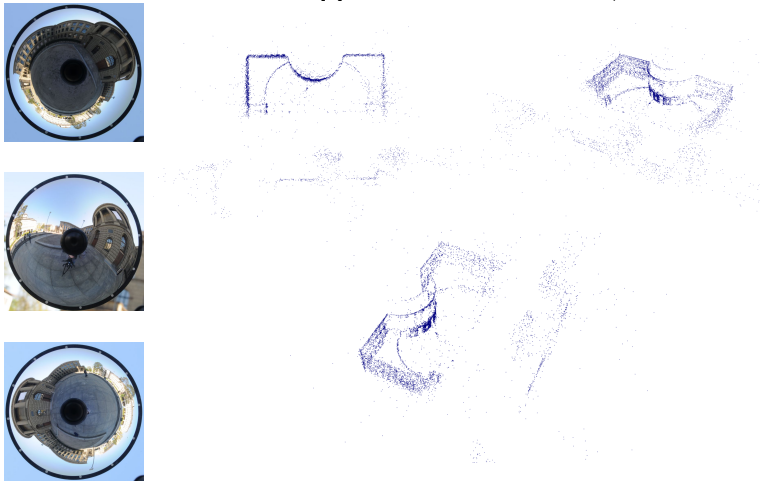

Fig. 9. *Building dataset* using fisheye and catadioptric images from Camposeco et al. [1]. Note that we are able to triangulate additional features not present in the fisheye images, e.g. the trees and facade of the building across the street. For visualization purposes we show the 3D points in blue instead of using the image RGB colors (the other building has a white facade.)

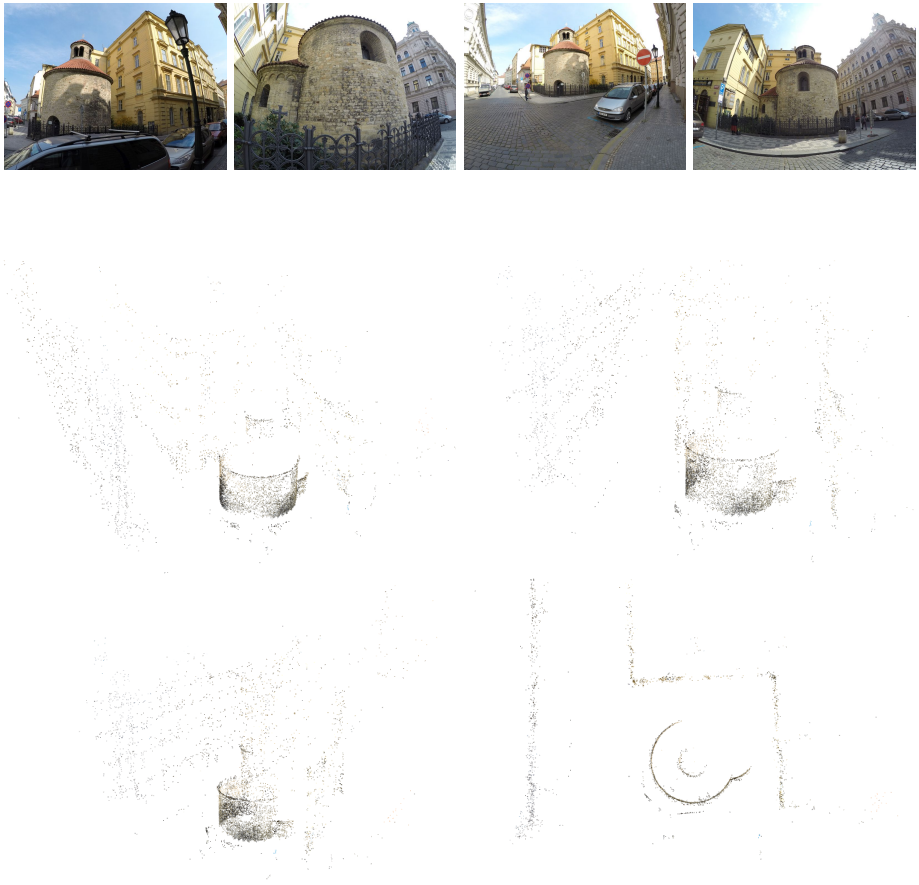

Fig. 10. *Rotunda* [3], 62 images, 16292 points, 0.41 px average reprojection error.

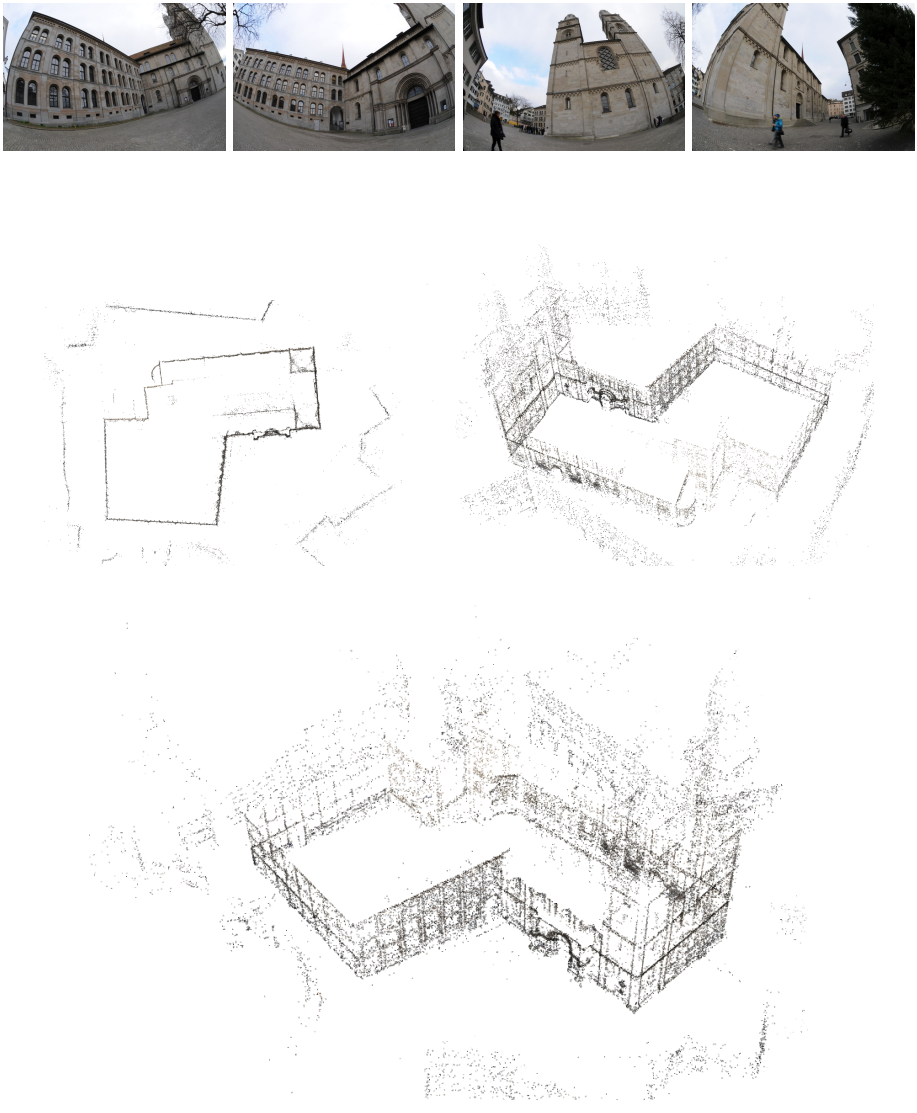

Fig. 11. *Big Church*, 373 images, 89288 points, 0.50 px average reprojection error.

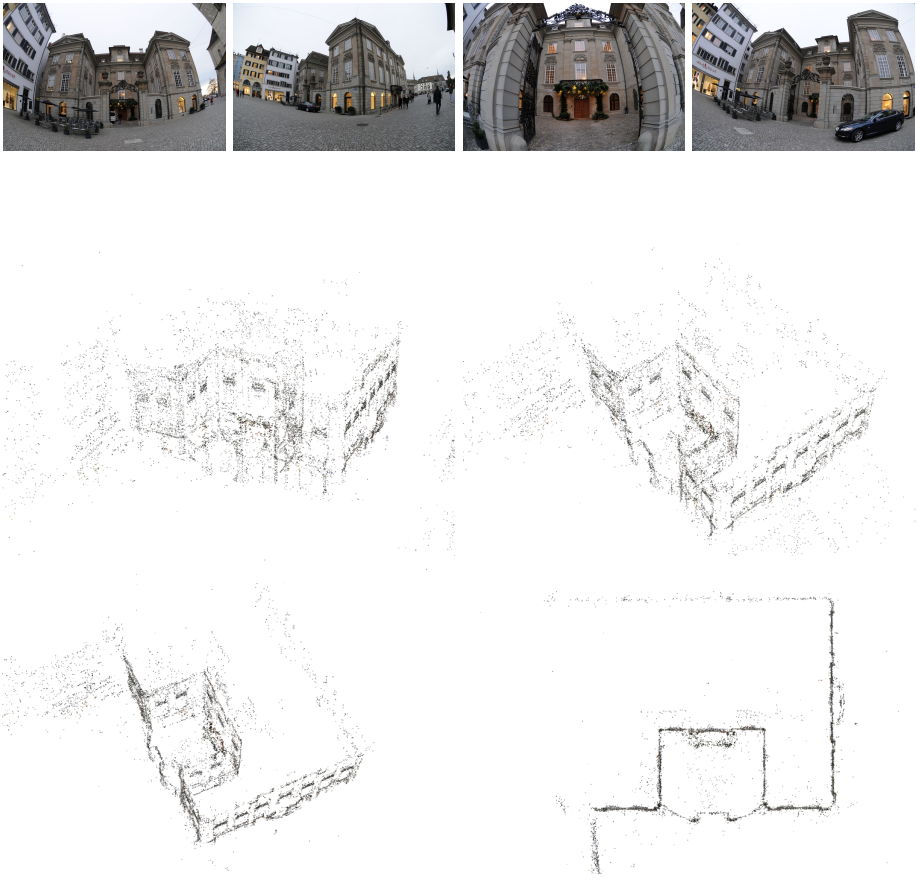

Fig. 12. *Building II*, 126 images, 27740 points, 0.44 px average reprojection error.

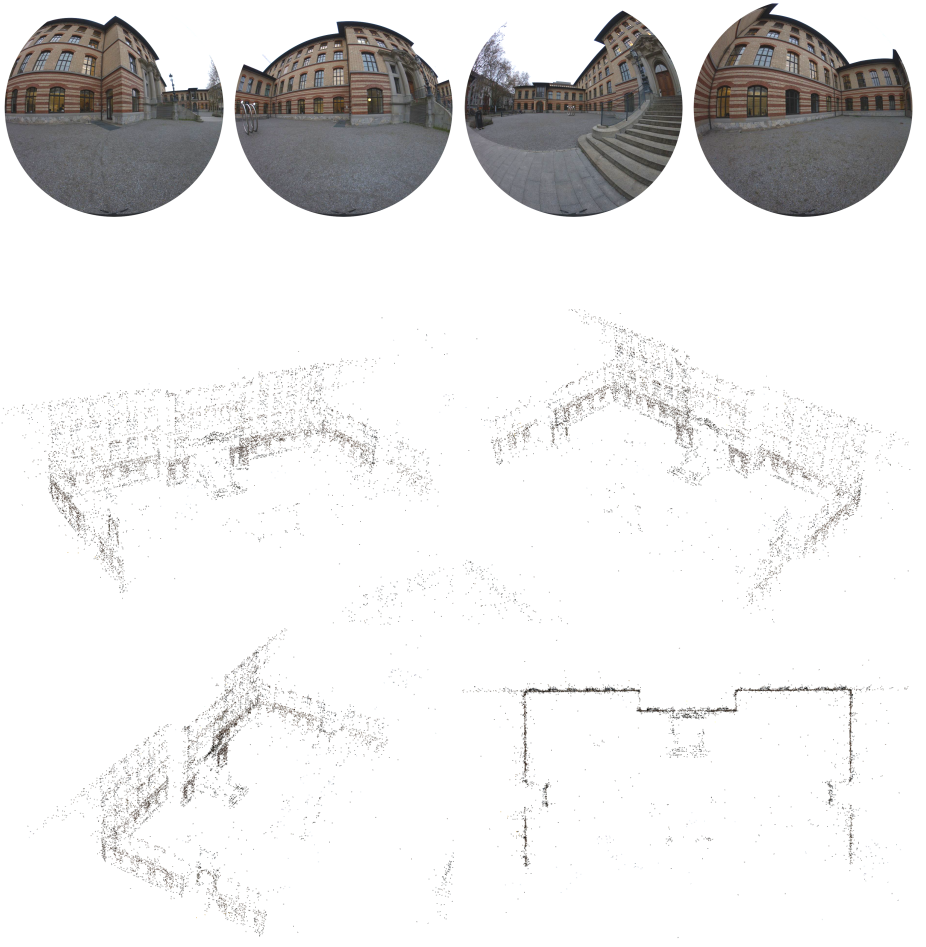

Fig. 13. *Fisheye Dataset*, 148 images, 14893 points, 0.51 px average reprojection error.

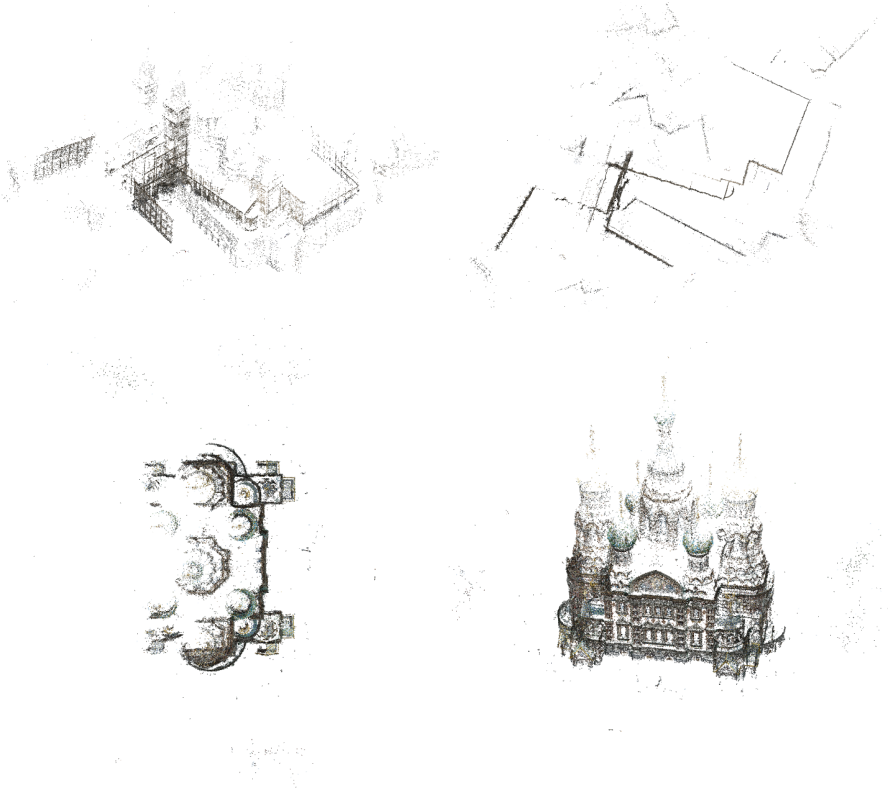

Fig. 14. Failure cases for *Spilled Blood* dataset and *Big Church* dataset.

References

1. Camposeco, F., Sattler, T., Pollefeys, M.: Non-parametric structure-based calibration of radially symmetric cameras. In: International Conference on Computer Vision (ICCV) (2015) 6, 12
2. Hartley, R., Schaffalitzky, F.: Reconstruction from projections using grassmann tensors. International Journal of Computer Vision (IJCV) (2009) 4
3. Kukelova, Z., Heller, J., Bujnak, M., Fitzgibbon, A., Pajdla, T.: Efficient solution to the epipolar geometry for radially distorted cameras. In: International Conference on Computer Vision (ICCV) (2015) 13
4. Olsson, C., Enqvist, O.: Stable structure from motion for unordered image collections. In: Scandinavian Conference on Image Analysis (SCIA) (2011) 7, 8, 9, 10, 11
5. Quan, L., Kanade, T.: Affine structure from line correspondences with uncalibrated affine cameras. Trans. Pattern Analysis and Machine Intelligence (PAMI) (1997) 4
6. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition (CVPR) (2016) 6, 12
7. Thirthala, S., Pollefeys, M.: Radial multi-focal tensors. International Journal of Computer Vision (IJCV) **96**(2), 195–211 (2012) 2, 4, 5
8. Tsai, R.: A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. Journal on Robotics and Automation (1987) 2