DeepPTZ: Deep Self-Calibration for PTZ Cameras

Chaoning Zhang
KAIST, South Korea
chaoningzhang1990@gmail.com

Francois Rameau
KAIST, South Korea
rameau.fr@gmail.com

Junsik Kim
KAIST, South Korea
mibastro@gmail.com

Dawit Mureja Argaw
KAIST, South Korea
dawitmureja@kaist.ac.kr

Jean-Charles Bazin
KAIST, South Korea
bazinjc@kaist.ac.kr

In So Kweon
KAIST, South Korea
iskweon@kaist.ac.kr

Abstract

Rotating and zooming cameras, also called PTZ (Pan-Tilt-Zoom) cameras, are widely used in modern surveillance systems. While their zooming ability allows acquiring detailed images of the scene, it also makes their calibration more challenging since any zooming action results in a modification of their intrinsic parameters. Therefore, such camera calibration has to be computed online; this process is called self-calibration*. In this paper, given an image pair captured by a PTZ camera, we propose a deep learning based approach to automatically estimate the focal length and distortion parameters of both images as well as the rotation angles between them. The proposed approach relies on a dual-Siamese structure, imposing bidirectional constraints. The proposed network is trained on a large-scale dataset automatically generated from a set of panoramas. Empirically, we demonstrate that our proposed approach achieves competitive performance with respect to both deep learning based and traditional state-of-the art methods. Our code and model will be publicly available at https://github.com/ChaoningZhang/DeepPTZ.*

1. Introduction

PTZ cameras are free to rotate and zoom to obtain high-quality images of a particular region of interest [1, 22]. This type of camera is used in surveillance systems, as well as for other purposes such as panorama creation or robotics applications [18, 22, 30]. To facilitate their use in these applications, accurately estimating their orientation and intrinsic parameters is highly desirable [18, 22]. If the camera has fixed intrinsic parameters (Pan-Tilt camera), off-line calibration methods with calibration objects [32] are adopted to provide an accurate estimation of the sensor's geometry. However, off-line calibration methods are hardly applicable for a PTZ camera since its intrinsic parameters (i.e. focal length and distortion) can constantly change through zooming [1]. Therefore, automatically estimating these parameters online is critical [12]. This process is known as camera self-calibration [1, 22].

Inspired by the success of deep learning in related geometric tasks, such as optical flow [16] and homography estimation [10], we propose to adopt convolutional neural networks (CNNs) for predicting both the intrinsic and extrinsic parameters of a PTZ camera from a pair of images. An extensive body of traditional methods exist for PTZ camera self-calibration [14, 7, 1]. Traditional techniques rely on robust feature matching between successive images [14, 7]. While these traditional approaches are effective, they rely on strong assumptions such as distortion free images [7, 1], similar intrinsic parameters for the two images [5], or the intrinsic parameters of one image being known from the previous calibration [26]. Our proposed deep learning based approach can achieve competitive performance without these assumptions. In contrast to existing techniques, we propose to estimate the focal length and distortion parameters of both images in the input pair as well as the rotation angle between them. To our best knowledge, no existing work solves this particular problem (i.e., varying focal lengths *plus* varying distortion parameters between both views in the context of a PTZ camera).

Recently CNN based camera calibration has been explored in previous works [3, 15], which estimate camera parameters from a single image. While CNN-based single image calibration is practically interesting, its accuracy is lower compared with traditional calibration methods [3]. In contrast, our approach reaches a better level of accuracy by using geometrical constraints (explicit feature matching) existing between two successive views acquired by a PTZ camera. Our method relies on an architecture called Dual-Siamese Network (DSNet), imposing bidirectional geometric constraint. To train our network, similar to previous deep learning based (single image) camera cali-

bration works [3, 15], we leverage panoramic images available online to generate a large number of synthetic pairs of views with various inter-camera rotation angles and intrinsic parameters.

The contributions of our work include: (1) To our best knowledge, this work is the first attempt to apply deep learning to the problem of PTZ camera calibration. (2) From an image pair, we jointly estimate the focal length and the distortion parameters of both images, which has not yet been introduced for the case of PTZ camera self-calibration. (3) We propose DSNet, imposing bidirectional geometric constraint, which achieves competitive performance with state-of-the-art methods. Moreover, our code and generated dataset will be publicly available to facilitate future research for applying deep learning to PTZ camera calibration.

2. Related works

Our proposed approach is inspired by both traditional and CNN based techniques. In this section, we summarize related works with respect to these two methods.

2.1. Traditional approaches for PTZ camera calibration

The self-calibration of zooming and rotating cameras has attracted lots of attention in the past decades. The pioneering work [14] focused on the automatic calibration of Pan-Tilt (no zooming) cameras. The homography between two successive images is computed in order to linearly estimate the Dual Image of the Absolute Conic (DIAC) from which the intrinsic parameters can be extracted via a Cholesky decomposition [14]. This method presents the advantage that multiple DIACs can be utilized together for the estimation of the camera's parameters, assuming they remain constant over the sequence. However, this approach is inapplicable for PTZ cameras where the intrinsic parameters can change for every new image. To cope with this limitation, Agapito et al. [1] propose a reformulation of the problem using the Image of the Absolute Conic (IAC). This novel formulation allows to introduce new linear constraints on the elements of the IAC (i.e. unit pixel aspect ratio, zero skew) and, in turn, permits to estimate the varying intrinsic parameters of the sensor. As an extension, later works [22, 19] propose a more robust optimization scheme based on linear matrix inequalities. Nevertheless, these techniques still rely on the strong assumption that the images are not affected by any geometrical distortion, which is rarely true in practice for PTZ cameras. To deal with the particular case of distortion, Byrod et al. [5] take advantage of the division distortion model [11] to map the radial distortion of the lens. Using this model, they propose a Gröbner basis based minimal solver requiring 3 correspondence points between two images to estimate the focal length and the distortion parameter. While taking the radial distortion into consideration undeniably improves the range of application of the approach, it assumes the intrinsic parameters to be the same for the image pair. To deal with varying distortion and focal length, Galego et al. [12] propose to track these parameters across zoom levels assuming one of the image in the pair has been already calibrated. While this approach is practical, it requires prior information about one image in the pair, which implies that more than two images are needed in the first place. In contrast, we propose a novel approach to estimate the distortion parameters and focal lengths from a single pair of images without any prior computation. Our technique is able to perform this estimation even when the intrinsic parameters (i.e. distortion parameters and the focal lengths) are different for the two views.

2.2. CNN approaches for camera calibration

The recent development of deep learning for computer vision tasks represents a good alternative to solve the problem of self-calibration. Existing approaches mostly focus on the particular case of single image self-calibration. DeepFocal [29] is the first work attempting to tackle this problem by training a CNN on a Structure-From-Motion dataset. While this seminal work demonstrates the feasibility of the technique, its generalization and accuracy remain problematic due to the limited number of training samples. As an extension, Hold-Geoffroy et al. [15] propose to estimate together the focal length and the horizon line from a single view. This joint estimation results in a significant improvement of the prediction accuracy, which is also partially due to the large quantity of synthetic training images, generated from panoramic images [31]. More recently, DeepCalib [3] includes the distortion parameter in the single image self-calibration problem by generating distorted images using the unified projection model [2]. The above methods only consider a single input image, which makes them versatile but also less robust than multi-view based techniques. In this paper, we explore the first deep learning based multi-view PTZ camera calibration approach.

While this particular configuration has never been addressed using CNN, deep learning has already been applied to related geometric regression tasks, such as optical flow [9, 27], stereo matching [20], fundamental matrix estimation [23], and homography estimation [8]. For the tasks involving dense correspondence regression, the encoder-decoder style network is often adopted [9, 27, 6]. While for parametric model estimation, such as homography, the VGG style network or other similar backbone structures are often employed [8]. Pioneering works, attempting to solve geometric regression problems through deep learning architectures, stack the two images as the input [9, 8]; whereas more recent works demonstrate that the performance can be improved by processing the two images separately first via a Siamese network and then combining the two correspond-

ing embeddings through correlation [27, 6]. Thus, in our work we adopt the later treatment as our baseline approach.

3. PTZ camera model with radial distortion

Our work relies on a widely adopted assumption that the PTZ camera performs a purely rotational motion between two successive images I_1 and I_2, this rotation is represented as a 3×3 orthogonal rotation matrix $\mathbf{R}_{12}$. While this constraint is practically impossible to enforce in a real vision system, this is a common and realistic assumption for PTZ cameras [22]. Under this assumption, the projection of a 3D point $\mathbf{P} = (X\ Y\ Z)^T$ into both image planes at the pixel location $\mathbf{p}_j = (x_j, y_j, 1)^T$ can be written $\mathbf{p}_1 \sim \mathbf{K}_1\mathbf{P}$ for the first image (referential) and $\mathbf{p}_2 \sim \mathbf{K}_2\mathbf{R}_{12}\mathbf{P}$ for the second one, where $\mathbf{K}_j$ is the 3×3 intrinsic matrix (of the j^{th} image) mapping the perspective projection of a 3D point onto the image plane. The matrix $\mathbf{K}_j$ encapsulates the focal length f_j, the principal point location $\mathbf{c}_j = (u_j, v_j)^T$, the pixel aspect ratio λ_j, and the skew parameter s_j.

The above camera model does not take distortion into account. To include this parameter, we follow the parameterization utilized in [3], namely the unified spherical model [21]. This parameterization maps the radial distortion of the image via a double projection on a Gaussian sphere with a single distortion parameter noted ξ_j. For further details concerning this parameterization, refer to [3, 21].

4. Proposed approach

In this section, we first formulate the problem we target to tackle and then propose deep learning based network for solving it. To enable the training of such a network, a large number of image pairs and their corresponding ground truth parameters are needed. The process of generating such datasets from panoramas will also be illustrated.

4.1. Problem formulation

In this paper, we propose a deep learning based approach to self-calibrate a PTZ camera including both the intrinsic parameters and the rotation between the two images. The techniques introduced in Sec. 2 rely on different constraints on the parameters of a camera. Similarly, to simplify the calibration problem, we enforce a certain number of commonly admitted assumptions [1]: the principal point is located at the center of the image, the skew is equal to zero and the aspect ratio $\lambda = 1$, leaving only four intrinsic parameters to be estimated (per image pair): the focal lengths f_1 and f_2 and the distortions ξ_1 and ξ_2. Additionally, we predict the Tait–Bryan angles y_{12}, p_{12} and r_{12} (yaw, pitch, roll) as the rotation parameters between the two images I_1 and I_2. To perform self-calibration of a PTZ camera, we propose two different architectures: (single) Siamese network and dual-Siamese network (DSNet).

4.2. Single Siamese network

In this work we adopt the widely used Siamese CNN as the baseline for predicting the camera parameters. The Siamese network was first proposed in [4] for signature verification. It has shown compelling performance in a wide range of vision applications, such as optical flow [9, 27] and stereo matching [6]. The philosophy of the Siamese CNN is to use the same weight to process two images to obtain the corresponding features. For our application, the single Siamese network is rather straightforward (see Figure 1(a)), two weight-sharing CNNs are utilized to extract the features (equivalent to the role of feature descriptor) from the two images separately. These features are then combined (via correlation or concatenation, see analysis Sec. 5.1), in order to regress the camera's parameters $\hat{\theta}_f = \{r_{12}, p_{12}, y_{12}, f_1, f_2, \xi_1, \xi_2\}$ through the regression CNN. It is worth noting that the features extraction and the regression networks are fundamentally different even if both of them adopt CNN blocks. To construct the two CNN blocks, we choose Inception-v3 [28], which delivers compelling performance with fewer parameters than the VGG style networks. More specifically, we divide the Inception-v3 model into two parts. The convolutional layers before (and including) the feature resolution of 35×35 are used for feature extraction, while the remaining part is used for the regression network. The last fully connected layer (with 1000 heads as the output for ImageNet classification) is replaced with seven separate heads for our task.

Interestingly, this architecture follows the traditional self-calibration pipeline in three steps: (1) feature extraction (Siamese network encoder), (2) feature matching (correlation), and (3) parameter estimation (regression network).

4.3. Dual Siamese network

From the perspective of geometric understanding, the network predicts a set of parameters that map the correspondence between I_1 and I_2. It is interesting to note that this geometric correspondence is bidirectional: forward matching (from I_1 to I_2) and backward matching (from I_2 to I_1). We conjecture that imposing this bidirectional constraint is beneficial for improving the network performance.

To leverage the benefit of the bidirectional geometric constraint, we flip the order of the two images and to perform a second forward pass within the same network. This process is depicted in Figure 1(b). Note that there is redundant computation performed for feature extraction, which can be mitigated by switching the extracted features instead of flipping the images. To distinguish from the single Siamese network, we term this architecture dual-Siamese network (DSNet).

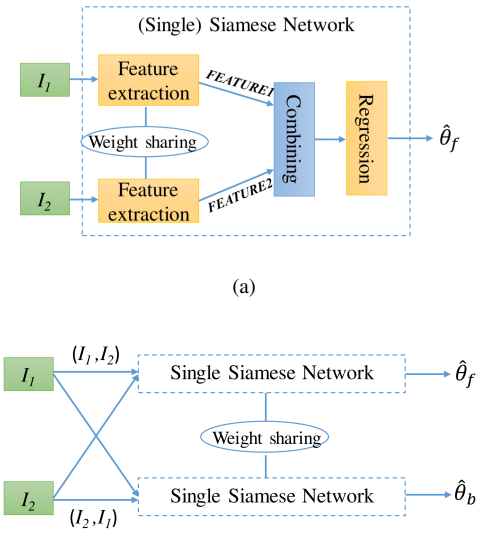

Figure 1. (a) Proposed (Single) Siamese Network, (b) Proposed DSNet.

This structure outputs two sets of parameters, respectively $\hat{\theta}_f$ for the forward estimation and $\hat{\theta}_b = \{r_{21}, p_{21}, y_{21}, f_2', f_1', \xi_2', \xi_1'\}$ for its backward counterpart. With both forward and backward estimation outputs, the bidirectional constrained can be enforced by minimizing the error with their corresponding ground-truth values. In both single Siamese network and DSNet, we employ correlation to combine the extracted features. The correlation technique was first introduced in [9] and has then been widely used in deep learning based geometric works [16, 6, 27]. In those works, the correlation search region is limited to the neighbourhood of the corresponding feature. This strategy is effective for predicting optical flow since the pixel displacement is usually assumed to be relatively small [9, 16] and multi-scale correlation can also be used [6, 27]. For PTZ camera calibration, the displacement can be quite large, thus it is reasonable to search either with a large neighborhood or globally. Since the correlation is single-scale in our work and operated with the resolution of 35×35, it is not computationally heavy to conduct global correlation, we choose to adopt global search region. For more details on the correlation techniques, refer to [16].

Compared with single Siamese network, the proposed DSNet has the advantage to explicitly enforce the bidirectional constraint. Note that DSNet has exactly the same number of parameters as single Siamese regression structure. During the test stage, we can leverage another advantage of the proposed DSNet, which is to do an average of the forward and backward outputs (e.g., taking the average of f_1 and f_1' as the prediction of the I_1 focal length). This averaging operation is similar to the philosophy of model ensembling. However, traditional model ensembling requires extra independent model(s), implying extra parameters, whereas our proposed DSNet can achieve the ensembling effect with only one independent model.

4.4. Dataset generation

There is no publicly available large-scale dataset for training a DNN to perform PTZ camera calibration and evaluating its performance. To overcome this obstacle, panoramas have been used by [3, 15] for automatically generating large-scale datasets to train and evaluate their proposed approaches. Following [3, 15], we choose publicly available SUN360 database [31] to generate the dataset in this work. Contrary to [15], which exclusively considers a pure pinhole camera model, our work takes distortion into account and follows the parameterization described in [3].

The generation of an image from an input panorama involves two steps. First, the panorama is projected onto a unit sphere; second, a distorted image is generated given a focal length and a distortion parameter [2]. Contrary to [3, 15] which estimate the parameters from a single image, our method requires image pairs with sufficient overlapping. To guarantee sufficient overlapping between successive images, we enforce the inter-camera rotation to be comprised between $\pm 15^\circ$ for each rotation angle (i.e., roll, pitch, and yaw). Similarly, the focal length is randomly generated in a range between 50 to 500 pixels with an extra constraint that the difference between the two focal lengths is within ± 50 pixels. Accordingly, the distortion parameter ξ is limited from 0 to 1 with an extra constraint that the difference between the two distortion parameters is within ± 0.1.

It is important to notice that while a PTZ camera mechanically admits two degrees of freedom, the relative rotation between two successive images cannot always be modelled by only two rotation parameters (if no other prior information is given). As a practical illustration, if a PTZ camera installed to the ceiling of a room is oriented downward, any pan rotation of the camera will lead to a pure roll rotation (rotation around the optical axis of the camera) between successive images. For this reason, a very large number of works assume a rotation matrix with 3 degrees of freedom for PTZ camera calibration [1, 22, 14].

One of the main obstacles in applying deep learning to PTZ camera applications is the lack of publicly available dataset. To our best knowledge, our generated dataset is the first large-scale dataset that can be used for training a DNN to perform PTZ camera calibration. To improve research reproducibility and facilitate future research, we will make our code used for generating dataset publicly available.

Model	y_{12} (degree)	p_{12} (degree)	r_{12} (degree)	$f_1(pixel)$	ξ_1
DSNet-corre	0.374	0.377	0.172	12.889 (11.432)	0.085(0.076)
Siamese-corre	0.510	0.520	0.241	17.110	0.112
DSNet-concat	0.568	0.558	0.354	20.449 (19.308)	0.114(0.107)
Siamese-concat	0.752	0.748	0.564	27.284	0.138

Table 1. Performance comparison for different models. The number inside '()' is the average value of forward and backward predictions.

Model	y_{12} (degree)	p_{12} (degree)	r_{12} (degree)	$f_1(pixel)$	ξ_1
DSNet	0.374	0.377	0.172	12.889 (11.432)	0.085(0.076)
DeepCalib[3]	N/A	N/A	N/A	41.734	0.179
DeepHomo[8]	2.516	3.762	2.224	53.797	0.196
DeepHomo-DS	1.386	1.457	1.146	39.216	0.174

Table 2. Comparison to different CNN based approaches. The number inside '()' is the average value of forward and backward predictions.

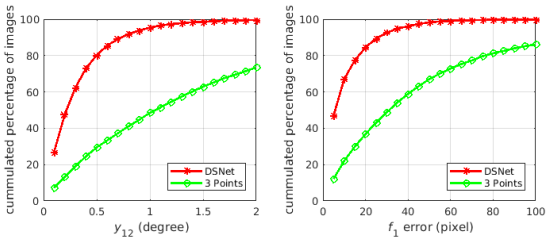

Figure 2. Comparison to 3-point [5] method when images have distortion.

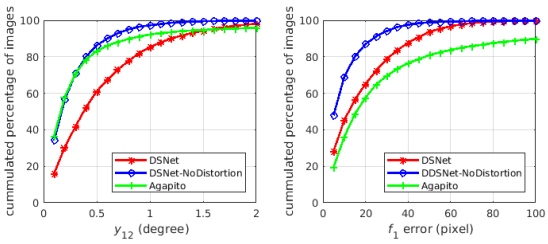

Figure 3. Comparison to Agapito [1] method when images have no distortion.

5. Results and discussion

The experimental results are presented in this section. Our network is trained such that it minimizes the L1-smooth loss [13, 24] between the ground-truth value and the regressed parameters. Since the outputs have different scales, we empirically found that shifting the values to zero average and then multiplying the cost for the focal length and distortion by a factor of 0.1 and 10 respectively achieves better performance. The model is trained on a single GPU with the batch size of 32. We adopt the Adam optimizer [17] with an initial learning rate set to 0.001, which is divided by 2 every 5 epochs. We stop the training after 30 epochs when the convergence is observed. We generate 10 image pairs from each indoor panorama in SUN360 panorama database, resulting in a total of 128,000 image pairs, 108,000 for training, 6000 for validation and 6000 for test. Following [3, 15], each panorama is exclusively used for training, validation or test. The image size is 299×299 pixels [28].

5.1. Model ablation study and analysis

Our DSNet outputs two sets of parameters $\hat{\theta}_f$ and $\hat{\theta}_b$ as illustrated in Figure 1(b). For clarity, we only report $\hat{\theta}_f$. The average error obtained on our test dataset (6000 image pairs) generated from panoramas is shown in Table 1. This comparative analysis shows that correlation is better than concatenation for combining features extracted from the two input images. Thus in the remaining of the paper, the DSNet always adopts correlation for combining features. Moreover, it can be concluded from Table 1 that DSNet significantly outperforms the (single) Siamese network. During the test, averaging the two outputs of the network can further decrease the error of f_1 and ξ_1 by more than 10%.

5.2. Comparison to CNN based approaches

To our best knowledge, our work is the first CNN based approach dedicated to general PTZ camera calibration using image pairs as the input. Therefore, we first compare our results with existing CNN based single image camera calibration methods. Previous work [15] jointly performs horizon estimation and camera calibration, excluding the distortion. However, the focal length can be compared with our approach. The average error for the vertical field of view reported in [15] is about 4° while our method admits an average error of around 1.4°. We conclude that our proposed network outperforms [15] by a large margin. Note that the datasets are generated in a similar way, it is a fair comparison even they are not evaluated on the exact same test dataset. We further propose another comparison against DeepCalib [3]. We utilize their pre-trained model available

Test \ Training	Indoor y_{12} / p_{12} / r_{12} / f_1 / ξ_1	Outdoor y_{12} / p_{12} / r_{12} / f_1 / ξ_1
Indoor	0.374/ 0.377/ 0.172/ 11.431/ 0.076	0.417/ 0.450/ 0.200/ 12.974/ 0.086
Outdoor	0.417/ 0.428/ 0.192/ 12.974/ 0.086	0.407/ 0.443/ 0.183/ 13.049/ 0.085

Table 3. Indoor and outdoor cross test results.

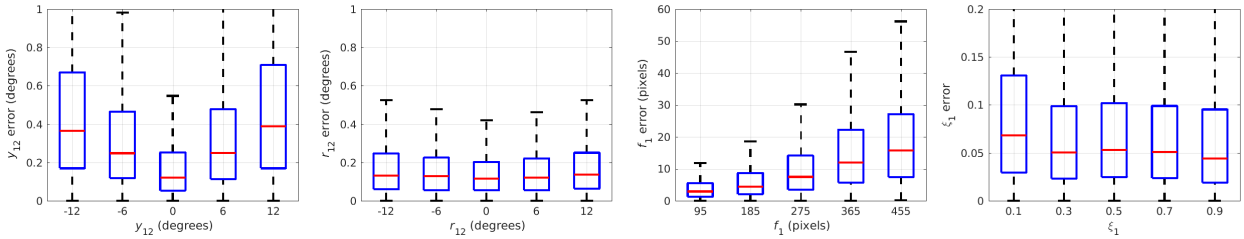

Figure 4. Box-percentile results for yaw angle, roll angle, focal length and distortion.

online[1] and evaluate it on the same test dataset. The results are available in Table 2. Note that the DeepCalib model is trained with millions of images generated in the same way while our model is trained with only 108,000 image pairs. Despite this large difference in terms of training samples, our method outperforms DeepCalib significantly. This performance gap can be explained by the fact that multi-views provide more constraints for an accurate estimation of the camera's parameters. We further compare our proposed approaches with similar CNN architectures used for homography estimation [8] (called DeepHomo). We adapt their network into our task through changing the final fully connected layer to have seven outputs (rotation angles, distortions and focal lengths). We train it with our training dataset and evaluate it on the same test dataset and the results are also available in Table 2.

Additionally, we propose to apply our dual Siamese structure on DeepHomo, which we call DeepHomo-DS. The first four convolutional layers are used for feature extraction and the remaining 4 convolutional layers are used for parameter regression. We note that DeepHomo-DS outperforms DeepHomo with a large margin, which is consistent with the result in Table 1. However, DSNet still outperforms DeepHomo-DS significantly, which empirically shows that the choice of the backbone structure is one important factor that influences the performance.

Overall, our propose DSNet achieves competitive performance and it is an appropriate CNN architecture for performing PTZ camera calibration.

5.3. Comparison to traditional calibration methods

To the best of our knowledge, as discussed in section 2.1, no existing traditional PTZ self-calibration method can estimate the focal lengths and the distortion parameters of both images when the two images have different unknown intrinsic parameters (focal lengths and distortions). Thus, to enable comparison, we intentionally decrease the complexity of the task by either assuming the two images contain distortion but have similar intrinsic parameters or assuming different intrinsic parameters without distortion. We generate two new test sets based on the above assumptions. In this way, we can compare our approach with the 3-point method assuming the two images having the same intrinsic parameters [5] and the Agapito method assuming different intrinsic parameters but without distortion [1]. As a fair comparison metric, we choose cumulative percentage of error instead of the mean average error because the results of traditional methods contain some extreme outliers.

First, we compare our proposed method with the 3-point method [5], see results in Figure 2. For simplicity we choose to compare the yaw angle (as the representative of the three rotation angles) and the focal length. We do not report distortion comparison because their method uses a division distortion model different from our spherical distortion model. The comparison shows that our method achieves significantly better performance. We further compare our method with Agapito method [1], see results in Figure 3. It shows that our method achieves better performance for focal length but worse performance for the yaw angle. Note that the Agapito method can only be applied to PTZ camera without distortion. For the purpose of ablation study, we can make the same assumption for DSNet. We report the performance of another model trained specifically with image pairs with no distortion, indicated as DSNet-NoDistortion as shown in Figure 3. The DSNet-NoDistortion model achieves similar performance as Agapito method for the yaw angle and significantly better performance for the focal length. Overall, we note that the distortion estimation is a challenging issue for both DSNet and traditional methods. Traditional methods can achieve comparable performance as DSNet when the assumption is made that there is

[1] http://vcail.kaist.ac.kr/projects/DeepCalib

Figure 5. Comparison of (a) original image 1 and (b) undistorted image 1; (c) original image 2 and (d) undistorted image 2.

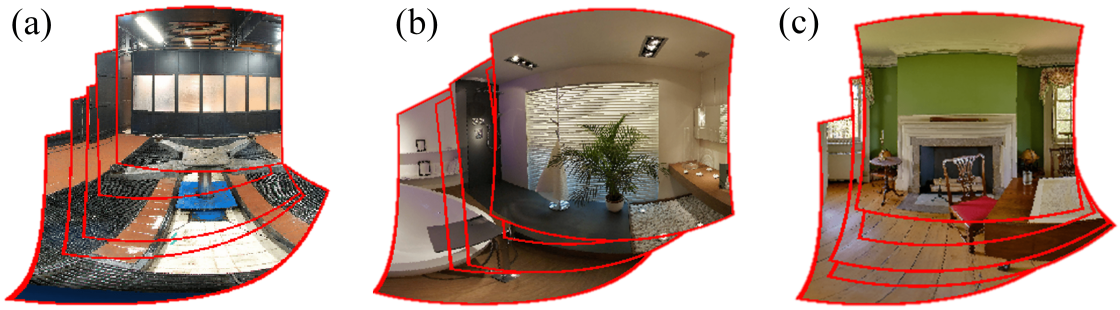

Figure 6. Three representative results of panoramic image stitching by our estimated camera parameters.

no distortion. However, for more challenging scenarios taking distortion into account, DSNet outperforms traditional methods by a large margin.

5.4. Statistical analysis

Section 5.1 presents the average error of the predicted values without any statistical insight of the results. Here in Figure 4, we thus use box-percentile to provide statistical prediction performance at different ranges. Since we find that yaw and pitch follow the same tendency, only the results for yaw angle is provided as reference. We observe a performance decrease with an increase of yaw angle or the focal length. The reason lies in that our approach depends on the correspondences between the two input images. The increase of the yaw angle or the focal length will decrease the overlapping area, which negatively affect the performance of the network. The roll angle has very limited influence on the overlapping area, thus the performance tends to be similar in the whole range. Finally, the distortion error difference along the whole chosen range is also relatively low and stable.

5.5. Generalization capability

In order to evaluate the generalization capability of our approach, we perform a cross test over indoor and outdoor datasets. To conduct this evaluation, we train our network using exclusively indoor data or outdoor data and test the resulting trained networks on both indoor and outdoor datasets (see Table 3). We note that the indoor network shows sensitively better results than its outdoor counterpart when the networks are tested on the same type of data as they have been trained with. In the outdoor dataset many images contain large area of textureless sky, which can be the cause for this performance degradation. We also notice that the model trained on the indoor dataset performs similarly as being trained on the outdoor dataset when tested on the same outdoor dataset. Overall, the performance gap among all the four scenarios is not significantly different, indicating a satisfying generalization capability of our proposed approach. The underlying reason might be that the correlation module in the proposed model mainly depends on correspondence matching not the content/feature itself.

5.6. Applications: Undistortion and image stitching

To provide a qualitative evaluation, we apply our method to two tasks whose visual quality is directly influenced by the accuracy of the camera's parameters: image undistortion and image stitching. A representative result of our method for image undistortion is shown in Figure 5. The corresponding error values of the predicted f_1, ξ_1, f_2, ξ_2 are 12.1, 0.10 , 9.7 and 0.08, respectively. We can notice that our undistortion is visually pleasing and properly straight-

ens the lines in the image.

A sequence of images with varying orientations but taken from one fixed point in space can be mapped into a common reference frame to create a perfectly aligned larger photograph with a wider field of view. Such a task is normally called image stitching [25], which can be fulfilled with the predicted camera parameters. We show three representative examples of image stitching containing five consecutive images in Figure 6. The averages error of the 5 predicted focal lengths and distortion parameters for Figure 6 (a) are 10.906 and 0.093 respectively. For Figure 6 (b) they are 3.96 and 0.052 respectively. For Figure 6 (c) they are 9.693 and 0.046 respectively. Some small stitching inconsistency can still be observed in Figure 6 (a) and (c).

6. Conclusions

We have presented the first deep learning based approach for PTZ camera calibration using image pairs as the input. We have targeted to estimate the focal length and distortion parameters of both images and the rotation between them. For the network design, we have explored two variants of Siamese networks and our DSNet by imposing bidirectional constraints improves the performance by a large margin compared with single Siamese network. The comparison result shows that our proposed approach achieves competitive performance with respect to traditional methods. Our proposed approach is also shown to have good dataset generalization through indoor and outdoor datasets cross test. Our method can be applied to image undistortion and panoramic image stitching. An interesting direction for future research would be the integration of a larger number of images in the self-calibration process.

Acknowledgements This work was funded by Naver Labs. Francois Rameau was supported by Korean Research Fellowship Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2015H1D3A1066564).

References

[1] L. Agapito, E. Hayman, and I. Reid. Self-calibration of rotating and zooming cameras. *Int. J. Comput. Vision*, 45(2):107–127, 2001.

[2] J. P. Barreto. A unifying geometric representation for central projection systems. *Comput. Vis. Image Und.*, 103(3):208–217, 2006.

[3] O. Bogdan, V. Eckstein, F. Rameau, and J.-C. Bazin. DeepCalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *CVMP*, 2018.

[4] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[5] M. Byröd, M. A. Brown, and K. Åström. Minimal solutions for panoramic stitching with radial distortion. In *BMVC*, 2009.

[6] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[7] L. De Agapito, R. I. Hartley, and E. Hayman. Linear self-calibration of a rotating and zooming camera. In *CVPR*, 1999.

[8] D. DeTone, T. Malisiewicz, and A. Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

[9] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2015.

[10] F. Erlik Nowruzi, R. Laganiere, and N. Japkowicz. Homography estimation from image pairs with hierarchical convolutional networks. In *ICCV*, 2017.

[11] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *CVPR*, 2001.

[12] R. Galego, A. Bernardino, and J. Gaspar. Auto-calibration of pan-tilt cameras including radial distortion and zoom. In *ISVC*, 2012.

[13] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[14] R. I. Hartley. Self-calibration from multiple views with a rotating camera. In *ECCV*, 1994.

[15] Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gambaretto, S. Hadap, and J.-F. Lalonde. A perceptual measure for deep single image camera calibration. In *CVPR*, 2018.

[16] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] M. Lalonde, S. Foucher, L. Gagnon, E. Pronovost, M. Derenne, and A. Janelle. A system to automatically track humans and vehicles with a PTZ camera. In *Visual Information Processing XVI*, volume 6575, page 657502, 2007.

[19] H. Li and C. Shen. An LMI approach for reliable PTZ camera self-calibration. In *AVSS*, 2006.

[20] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *CVPR*, 2016.

[21] C. Mei and P. Rives. Single view point omnidirectional camera calibration from planar grids. In *ICRA*, 2007.

[22] F. Rameau, A. Habed, C. Demonceaux, D. Sidibé, and D. Fofi. Self-calibration of a PTZ camera using new lmi constraints. In *ACCV*, 2012.

[23] R. Ranftl and V. Koltun. Deep fundamental matrix estimation. In *ECCV*, 2018.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[25] M. V. S. Sakharkar and S. Gupta. Image stitching techniques-an overview. *Int. J. Comput. Sci. Appl.*, 6:324–330, 2013.

[26] S. N. Sinha and M. Pollefeys. Pan–tilt–zoom camera calibration and high-resolution mosaic generation. *Comput. Vis. Image Und.*, 103(3):170–183, 2006.

[27] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[29] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs. DeepFocal: A method for direct focal length estimation. In *ICIP*, 2015.

[30] Z. Wu and R. J. Radke. Keeping a pan-tilt-zoom camera calibrated. *Pattern Anal. Mach. Intell.*, 35(8):1994–2007, 2013.

[31] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba. Recognizing scene viewpoint using panoramic place representation. In *CVPR*, 2012.

[32] Z. Zhang. A flexible new technique for camera calibration. *Pattern Anal. Mach. Intell.*, 22:1330–1334, 2000.