Pose Estimation for Event Camera Using Charuco Board Based on Image Reconstruction

Ngoc Trung Mai, Ren Komatsu, Hajime Asama, and Atsushi Yamashita

***Abstract*— Event cameras offer attractive properties compared to conventional frame-based cameras, such as high temporal resolution, very high dynamic range, and low power consumption. Thanks to these characteristics, event cameras have a great potential for sensing challenging lighting or high motion conditions in computer vision tasks and robotics applications.**

Traditional patterns such as chessboard and circle grid based methods have been proposed to calibrate or estimate the pose of the event camera. However, these methods are less versatile as they require the entire board to be visible in all images and do not allow occlusion. To overcome these limitations, this paper proposes a new method to estimate the 6DoF pose of an event camera using a Charuco board based on image reconstruction with a deep learning approach. Using images reconstructed from the event streams captured of the Charuco board, it can be successfully estimated the 6DoF pose of the event camera even in the presence of occlusion. Experiments performed in a simulation environment show the effectiveness of the proposed method.

I. INTRODUCTION

In recent years, the demand for 3D sensing technology has increased. Conventional frame-based cameras have typically been used to capture information. A conventional frame-based camera captures a scene by accumulating photons reflected from objects over a period of time to produce an image. Conventional cameras have been used in many studies on 3D sensing due to their advantage of high-resolution images. However, conventional frame-based cameras often suffer from low frame rates, high latency, or poor adjustment to extreme lighting conditions.

Recently, new image sensors, called event cameras, neuromorphic cameras, or dynamic vision sensors offer a revolutionary new paradigm for capturing scenes. Event cameras have received a lot of attention for their potential in robotics and sensing in challenging environments, such as optical flow estimation, object segmentation, and visual odometry in low-light, dynamic scenarios, etc [1]. Inspired by the behavior of a biological retina, event cameras have a hardware setup that is fundamentally different from conventional frame-based cameras. Instead of recording image frames, event cameras record asynchronous sequences of intensity changes per pixel with precise time stamps. It allows capturing data for high temporal resolution, very low power consumption, and high dynamic range. While event cameras have many advantages over conventional cameras, their differences from conventional frame-based cameras prevent the direct use of standard computer vision techniques on event data.

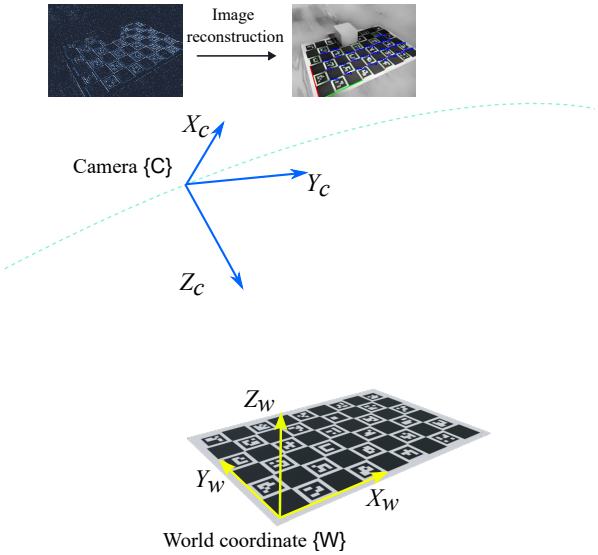


Fig. 1. Overview of event camera pose estimation using Charuco board.

In the case of conventional frame-based cameras, fiducial marker systems such as ARTag and ArUco have been widely researched in computer vision and robotics like augmented reality, and camera calibration [2]-[4]. These systems are very useful for camera calibration, monocular pose estimation, and pose verification. Conversely, in the case of event cameras, there are still few systems for pose verification and pose estimation. Several methods are proposed for the calibration of event cameras. Such as, Huang et al. [5] proposed a dynamic event camera calibration system using a circle grid pattern. Muglikar et al. [6] proposed an event camera calibration framework using a traditional chessboard. These methods can be extended to be also used to estimate camera pose. While these methods can archive high accuracy but only in good conditions, it is lack versatility and robustness since it requires that the entire board must be visible in all image and all corners or circles must be detectable in order to be able to be used and of cause occlusions are also not permitted. This motivated us to use the Charuco board to estimate event camera pose more robustly since the

N. T. Mai, R. Komatsu, and H. Asama are with the Department of Precision Engineering, School of Engineering, The University of Tokyo, 113-8656, Tokyo, Japan. {mai, komatsu, asama}@robot.t.u-tokyo.ac.jp

A. Yamashita is with the Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, 277-8563, Chiba, Japan. yamashita@robot.t.u-tokyo.ac.jp

Charuco board is known to overcome these limitations of classical checkerboards and circle grid boards in the case of conventional frame-based cameras [7].

In this paper, we propose a new method for estimating event camera pose using a Charuco board through an image reconstruction approach. An overview of the proposed method is shown in Fig. 1. The event streams captured of the Charuco board are then used to reconstruct grayscale intensity images. We define the world coordinate system with reference to the Charuco board. Firstly, the poses of the Charuco board in the camera coordinate are estimated using the reconstructed images. Then, coordinate transformations are then utilized to estimate the camera poses. The estimation method is evaluated throughout the simulation.

The potential advantage of the Charuco board is that the Charuco board combines the benefits of the Aruco marker-based approach and chessboard pattern-based approach. All checker pattern is uniquely coded and identifiable. This allows even partially obscured or non-ideal images to be used to estimate camera pose. This is particularly relevant in the case of event cameras since although techniques for reconstructing intensity images from events have been actively studied and it is possible to reconstruct high dynamic range images at a very high frame rate [8]-[11], however, the quality of reconstructed images is still limited and often contains partial artifacts per image [10]. In this paper, we also propose a method that applies some image processing techniques such as denoising and contrast adjustment to improve the quality of the reconstructed intensity images.

The remainder of this paper is organized as follows. Section II briefly presents the principle of an event camera. Section III describes our proposed methodology for the estimation of event cameras using the Charuco board. Section IV presents the experimental results constructed in the simulation environment. Finally, Section V presents conclusions.

II. PRELIMINARIES OF EVENT CAMERA

Event cameras are arrays of pixels that respond to local changes in brightness. Unlike conventional frame-based cameras, which capture images using a shutter, event cameras have pixels that operate independently and asynchronously respond to changes in brightness that occur. Each pixel in event cameras is performed as a continuous logarithmic photoreceptor with asynchronous signal processing $L_(x_k, y_k, t_k)$ [12]. It stores a reference brightness level and continuously compares it to the current brightness level as follows:

$$\Delta L(x_k, y_k, t_k) = L(x_k, y_k, t_k) - L(x_k, y_k, t_k - \Delta t_k), \quad (1)$$

where $\Delta L(x_k, y_k, t_k)$ is the change in brightness at the pixel (x_k, y_k) and at timestamp t_k. If absolute value of $\Delta L(x_k, y_k, t_k)$ exceeding given threshold C, the pixel will respond with an event $e_k(x_k, y_k, t_k, p_k)$, where p_k is a polarity of the k^{th} event that represents an increase or decrease in intensity.

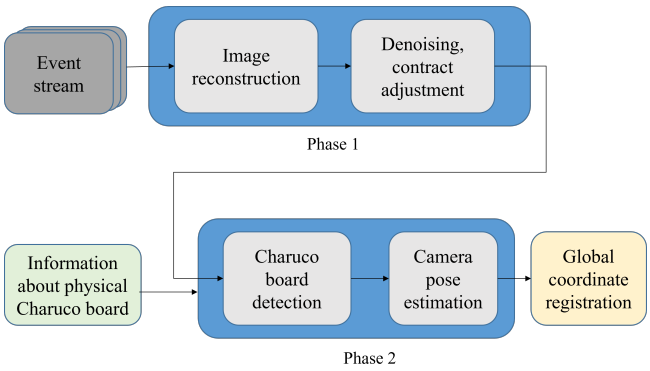


Fig. 2. Schematic representation of the proposed method.

III. POSE ESTIMATION USING CHARUCO BOARD

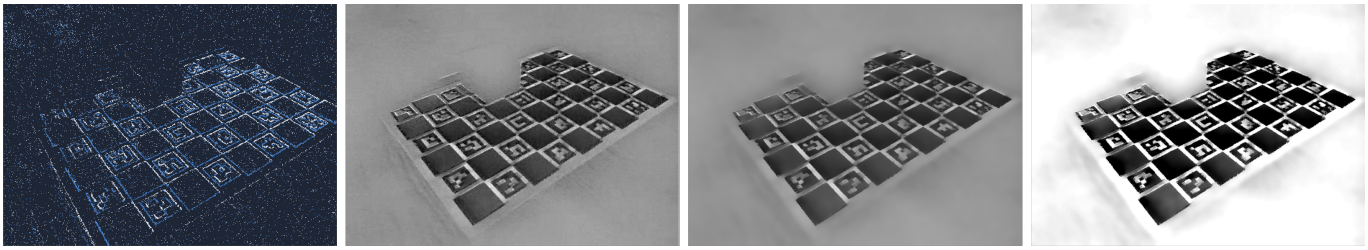
A. Framework overview

A schematic representation of the method proposed in this study is shown in Fig. 2. The input of the algorithm is the event streams, which are captured of the Charuco board by the event camera, and the related information of the Charuco board. The output is the 6DoF poses of the event camera. There are 2 main phases. In phase 1, grayscale intensity images are reconstructed from the event streams using a neural network. Then, in order to improve the quality of reconstructed images, some image processing techniques are applied. In phase 2, from reconstructed images, the Charuco board is detected, then the 6DoF poses of the camera will be estimated based on the detection results.

B. Image reconstruction and pre-processing

Fig. 3 shows the process of image reconstruction and preprocessing of reconstructed images. First, we reconstruct grayscale intensity images (Fig. 3 (b)) using a neural network from the original event streams obtained from the event camera (Fig. 3 (a)). For the image reconstruction process, we utilized the trained network proposed in E2VID [9]. E2VID proposed a convolutional recurrent neural network (C-RNN) that achieves considerable results on image reconstruction from event streams. To be able to use the event streams in the C-RNN, it is neccessary to convert the event streams into event tensors. An event tensor corresponding to the output of one intensity image is generated by grouping the event streams, which are in temporal neighborhoods of a fixed duration time $\Delta t(\mathrm{s})$. This results that the event tensors being generated at a frequency of $1/\Delta t(\mathrm{Hz})$, and the grayscale intensity images are also being reconstructed at the same frequency.

Second, we noticed that reconstructed images usually contain a lot of noise and artifacts and tend to have small dynamic ranges. Therefore, in order to improve the quality of the reconstructed images, processes including denoising and contrast adjustment are applied. A Bilateral filter is utilized to reduce the noise of images. By using the Bilateral filter, it is possible to smooth the image while preserving edges as shown in Fig. 3 (c). Next, to improve the brightness, an



(a) (b) (c) (d)

Fig. 3. Procedure of the image reconstruction and the preprocessing of reconstructed image: (a) original event data, (b) image reconstruction using neural network, (c) using a Bilateral filter to smooth the image while preserving edges, (d) contrast adjustment.

contrast adjustment conversion is applied in order to convert the range of brightness of images to a higher range, resulting in almost [0, 255] for all images. The contrast adjustment process is a linear transformation, which can be written as:

$$g(u,v) = \alpha * f(u,v) + \beta, \tag{2}$$

where $g(u,v)$ and $f(u,v)$ are brightness at pixel (u,v) after and before conversion, respectively. And, α and β are parameters to control contrast and brightness, corresponding to the scale and shift amount respectively. The values of α and β are calculated based on some first reconstructed images and used for the remaining sequence of reconstructed images.

C. Camera pose estimation

In this phase, reconstructed images of the Charuco board are used. A Charuco board is a planar board where the markers are placed inside the white squares of a chessboard. In this phase, individual Aruco markers will be first detected. The detected Aruco markers are used to interpolate the position of the chessboard corners so that it has the versatility of marker boards since it allows occlusions or partial views. Each corner on a Charuco board has a unique identifier assigned.

In the pose estimation step, we define 2 main coordinate systems as illustrated in Fig 1, the world coordinate system W and the camera coordinate system C. The world coordinate is defined with reference to the Charuco board pose. Because the event camera has the same optical system as the frame-based camera, the model in the case of the frame-based camera can be used for the event camera. Therefore, the relation between the 2D point in the reconstructed image and the corresponding 3D point can be written as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_b & \boldsymbol{t}_b \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} X^w \\ Y^w \\ Z^w \\ 1 \end{bmatrix} \tag{3}$$

where $\left[u, v\right]$ represents 2D corner point in the image plane, which detected from Charuco board. K is the camera intrinsic matrix, which can be obtained in advance through the calibration procedure. $\left[X^w, Y^w, Z^w\right]$ represent corresponding 3D corner points with respect to the world coordinate system, which can be calculated from given the marker size and the board size. R_b and $\boldsymbol{t}_b$ are respectively the rotation matrix and the translation vector, which translate $\left[X^w, Y^w, Z^w\right]$ from the world coordinate system to the camera coordinate system.

Given a set of 2D corner points and computable corresponding 3D corner points, the perspective-n-Point (PnP) algorithm can be applied to estimate the pose of the Charuco board relative to the camera coordinate system. To detect the Aruco markers and corners of the chessboard, and to estimate the Charuco board pose from detected corners, we use well-constructed functions in the Aruco module provided in OpenCV [13]. The function also checks if the detected corners are sufficient for performing the estimation. If not enough corners are detected, the function returns false for estimating the pose of the Charco board.

Next, in order to calculate the pose of the camera on the world coordinate, the coordinate transformation is applied. As the world coordinate system is defined with reference to the Charuco board pose, the camera pose can be estimated as follows:

$$R_c = R_b^{-1}, \tag{4}$$

$$\boldsymbol{t}_c = -R_c \boldsymbol{t}_b, \tag{5}$$

where, R_c and $\boldsymbol{t}_c$ are the rotation matrix, the translation vector of the camera with respect to the world coordinate system, respectively. The camera pose is estimated for each frame of the reconstructed image. Therefore, the frequency of camera pose estimation depends on the frequency of the reconstruction of intensity images from the event streams.

IV. EXPERIMENT

In order to verify the proposed method, an experiment has been carried out by using a simulation environment. Blender, an open-source 3D computer graphics software [14], was

TABLE I
SPECIFICATIONS OF CHARUCO BOARD

Number of squares	6×9
Size of board	1.95×2.85 m
Square length	0.3 m
Marker length	0.225 m
ArUco dictionary	DICT_4X4_50

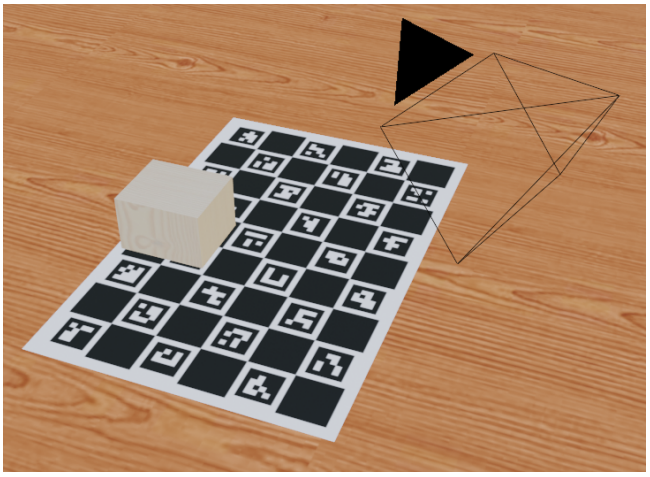


Fig. 4. The simulation experiment setup. The event camera was set up to point at the Charuco board while moving above.

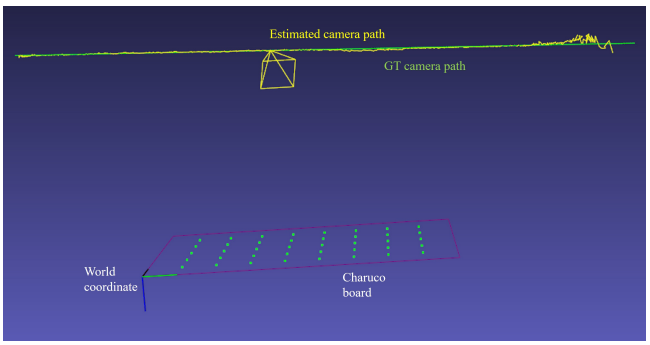


Fig. 5. Results of event camera pose estimation. The green line indicates the GT camera path. The yellow line indicates the estimated camera path of the case duration time Δt being 5 ms

used to create simulations. To simulate event data, we adopted the method of constructing an event camera simulator proposed in [15]. Firstly, we used Blender to create a 3D scene, set camera parameters, and render high frame rate continuous images. Then, the rendered images were used as input to the event camera simulator to generate event streams. Shot noise was also added to the simulation to generate more realistic data.

The experiment setup is shown in Fig. 4. An event camera was set up to always point at the floor and Charuco board while moving above. The specifications of the Charuco board are reported in Table I. The Charuco board (1.85 $\times$ 2.95 m) was placed on the floor and a cube object was placed on the board to create occlusion. The event camera was set about 2 m above the floor and moved along to the Charuco board. In our experiments, the 6 × 9 Charuco board which contains the first 27 elements of the DICT_4X4_50 ArUco dictionaries was used as shown in Fig. 4. From those settings, approximately 4 s of event streams were generated. From event streams, intensity images were reconstructed using the trained C-RNN network. As mentioned in III-B, to use event streams in the trained C-RNN network, it is necessary to transform the event streams into a sequence of event tensors by accumulating events over time intervals Δt.

TABLE II

RESULTS OF THE CAMERA POSE ESTIMATION WITH DIFFERENT DURATION TIMES Δt FOR IMAGE RECONSTRUCTION

Duration time (ms)	5	8	10
Frequency of pose estimation (Hz)	200	125	100
Number of total frames	791	495	396
Valid rate in estimation	0.814	0.844	0.859
Average RMSE (m)	0.0115	0.0113	0.0113

The estimation results of camera pose with the duration time Δt being 5 ms are shown in Fig. 5. We defined the world coordinate frame with reference to the Charuco pose. In other words, one of the corners of the Charuco board was the origin (0, 0, 0) of the coordinate. As shown in Fig. 5, the yellow line indicates the estimated camera path while the green line indicates the GT camera path. Fig. 6 shows reconstructed intensity images and the Charuco board detection results at different frame numbers of the case that the duration time Δt being 5 ms. When the duration time Δt being 5 ms, approximately 800 frames of intensity images were reconstructed from the simulated event streams. i indicates frame numbers. Corresponding event frames were also generated and shown on the left side of reconstructed images. The Charuco board detection results also were redrawn on the event frames for comparison. As shown in Fig. 6, the Charuco board was detected despite the appearance of the obstacle caused by the cube object. The results in Fig. 5 and Fig. 6 show that the proposed methodology can basically successfully detect the Charuco board and accurately estimate the camera pose even though there was the occlusion occurred by the cube object.

Experiments were also performed with different duration times Δt of 5, 8, and 10 ms. This resulted that the intensity images being reconstructed at 200, 125, and 100 Hz, respectively. Table II summarizes the camera pose estimation results for different duration times Δt of image reconstruction. As shown in Table II, the valid rates in the estimation of camera pose were 0.814, 0.844, and 0.859, respectively. And the corresponding average errors of the estimated camera position were 0.0115 m, 0.0113 m, and 0.0113 m, respectively. Increasing the duration time of the event stream used to reconstruct one intensity image slightly improved the valid estimation rate and the error of the estimation. It could be explained that when increasing the duration time, the quality of the reconstructed intensity images was improved and led to a higher valid rate in the estimation results of camera pose.

Fig. 7 presents the error in estimation results of camera position over time. As shown in Fig. 7, the errors in the position estimation are less than 0.02 m at most of the timestamps, especially in the period of 0-3200 ms. During this period, the results of the detection of the Charuco corners were sufficient so it was able to estimate the camera pose with high accuracy. This can be considered to be good enough since the camera was quite far always from the

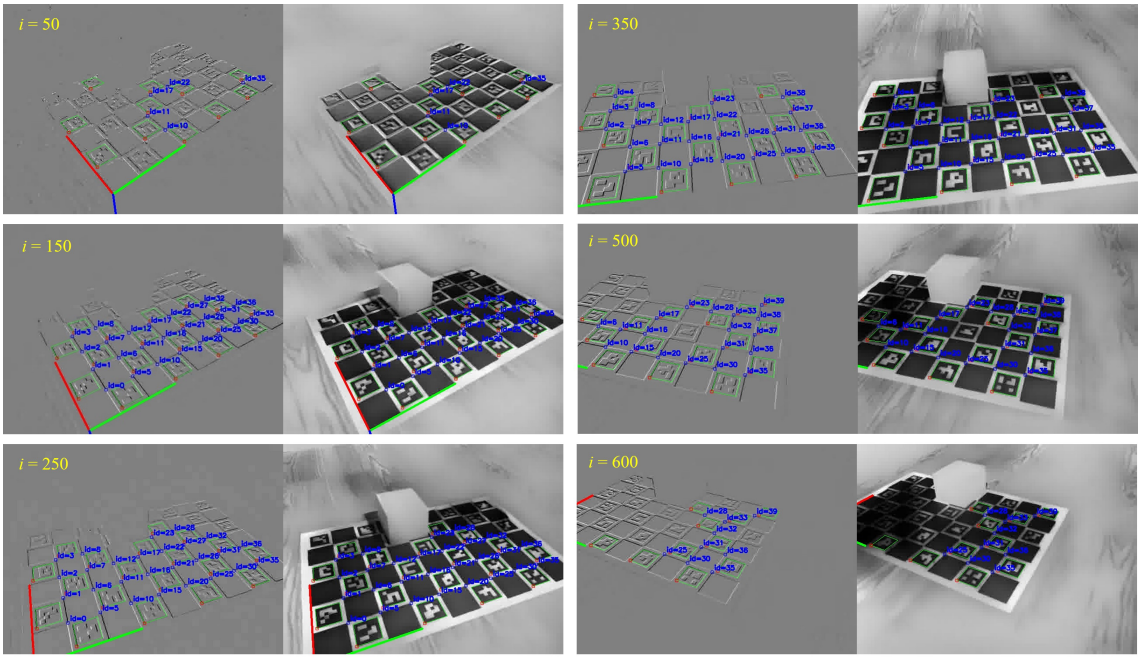


Fig. 6. Charuco board detection results in reconstruction images (2nd and 4th columns) at different frame numbers when the duration time Δt being 5 ms. Corresponding event frames were also generated (1st and 3rd columns). The detection of the Charuco board was performed in the reconstructed images. The detected results were plotted on both the reconstructed images and corresponding the event frame images.

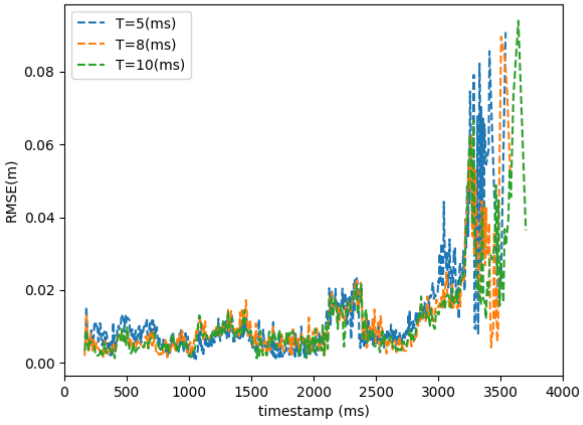


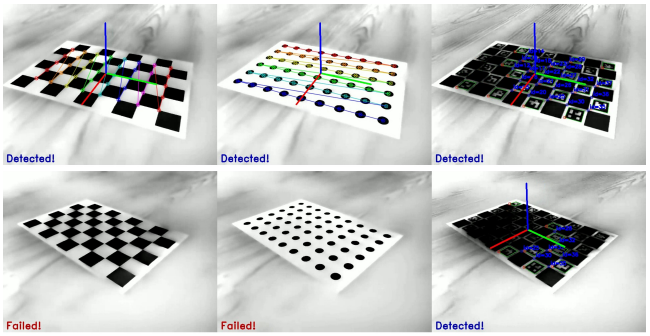
Fig. 7. The error (RMSE) in the results of the estimation of camera position with different duration times Δt(s). The blue dot line, orange dot line, green dot line indicate the case of the duration time $\Delta t = 5$ ms, 8 ms, and 10 ms, respectively.

Charuco board. However, the error tends to become bigger at later timestamps. This can be explained by the camera being farther away and also the reconstructed image quality is degraded at the later timestamps.

Charuco vs chessboard and circle grid board: In order to confirm the versatility of the proposed methods, we also conducted the second experiment to compare with methods that use a chessboard or a circle grid board. Note that to be able to use the chessboard or circle grid board for camera pose estimation, the entire board must always be visible to the camera. Therefore, we set up the camera path so that entire board is always inside the field of view of the camera and the obstacle above the board as shown in Fig. 4 also was removed. All setups and processes for image reconstruction were the same in all the methods. Fig. 8 shows some estimation results of different methods. As shown in Fig. 8, the proposed methods were able to detect the Charuco board even though there were failures in the detection of the chessboard or circle grid board. The world coordinate system was set with reference to the center of the boards. Table III summarizes the camera pose estimation results for different methods. As shown in Table III, even in the case that the camera was set to always capture entire the boards, the valid rate in the estimation of the proposed method was 0.791, much higher than the methods using chessboard (0.467) or circle grid board (0.182). However, even though the proposed method archived quite good results (average RMSE: 0.008 m), the methods using the chessboard or circle grid board showed better results, with average RMSEs were

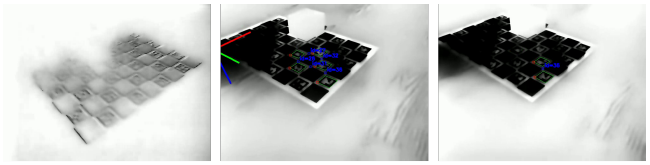
TABLE III
RESULTS OF THE CAMERA POSE ESTIMATION

Methods (ms)	Charuco (Ours)	chess	circle grid
Frequency estimation (Hz)	200	125	100
Number total frames	571	571	571
Valid rate	0.791	0.467	0.182
Average RMSE (m)	0.008	0.003	0.003



(a) Chess board (b) Circle grid board (c) Charuco board

Fig. 8. Some estimation results of different methods using chessboard, circle grid board, or Charuco board.



(a) (b) (c)

Fig. 9. Some failure examples of the Charuco board detection in the sequence of reconstructed images at 200 Hz. The trained C-RNN failed to reconstruct good quality images at some first frames such as frame no. 14 (a). Or some artifacts appeared in the reconstructed images such as frame no. 706 (b) frame no. 777 (c).

approximately 0.003 m in both methods. From those results, we concluded that the chessboard or circle grid board-based methods are preferable to the Charuco board-based method for the calibration task that accuracy is more prioritized. On the other hand, for tasks like camera pose estimation, the proposed method is preferable because it showed good accuracy enough and a much higher valid rate, and also allows the obstacle.

Limitations: Although compelling results have been demonstrated for pose estimation, we acknowledge a few limitations still exist in the proposed approach. Fig. 9 shows some failures in reconstructing intensity images. The trained C-RNN could not reconstruct high enough quality images in some frames, such as early frames like frame no. 14 (Fig. 9 (a)) or some later frames like frame no. 706 (Fig. 9 (b)) and frame no. 777 (Fig. 9 (c)). These issues are considered to be the disadvantages of the proposed method, which is based on the image reconstruction-based approach. The intensity image reconstruction approach is useful as well-constructed algorithms for normal frame-based images can then be utilized. However, because of using the recurrent neural network for intensity image generation from event streams, it also can lead to some common issues which usually being encountered with the recurrent neural network. It is the lack of quality at some initial frames or bad results at one point that will affect the results in later frames.

V. CONCLUSION

In this study, we proposed a novel methodology to estimate 6DoF pose for the event camera based on an image reconstruction approach using a Neural network. By using reconstructed images from event streams that captured a Charuco board, the 6DoF pose of the event camera can be estimated with high accuracy. Experiments performed in simulation environments demonstrate the effectiveness of the proposed method.

In the future, experiments on real event data will be conducted for further evaluation. Moreover, the methodology for improving the quality of reconstructed images will be explored.

REFERENCES

[1] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based Vision: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, No. 1, pp. 154–180, 2020.

[2] M. Fiala, "Artag, A Fiducial Marker System Using Digital Techniques," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 590-–596, 2005.

[3] S. Garrido-Jurado, R. Munoz-Salinas, F. Madrid-Cuevas, and M. Marin-Jimenez, "Automatic Generation and Detection of Highly Reliable Fiducial Markers under Occlusion," Elsevier Pattern Recognition, Vol. 47, No. 6, pp. 2280—2292, 2014.

[4] F. E. Ababsa and M. Malik, "Robust Camera Pose Estimation Using 2d Fiducials Tracking for Real-time Augmented Reality Systems," Proceedings of the ACM SIGGRAPH International Conference on Virtual Reality Continuum and Its Applications in Industry, pp. 431–435, 2004.

[5] K. Huang, Y. Wang, and L. Kneip, "Dynamic Event Camera Calibration," Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 7021–7028, 2021.

[6] M. Muglikar, M. Gehrig, D. Gehrig, and D. Scaramuzza, "How to Calibrate Your Event Camera," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1403–1409, 2021.

[7] D. Hu, D. DeTone, and T. Malisiewicz, "Deep Charuco: Dark Charuco Marker Pose Estimation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8436–8444, 2019.

[8] C. Scheerlinck, B. Nick, and M. Robert, "Continuous-time Intensity Estimation Using Event Cameras," Proceedings of the Asian Conference on Computer Vision, pp. 308–324, 2018.

[9] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High Speed and High Dynamic Range Video with An Event Camera," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 6, pp. 1964–1980, 2019.

[10] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-Video: Bringing Modern Computer Vision to Event Cameras," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3857–3866, 2019.

[11] Y. Zou, Y. Zheng, T. Takatani, and Y. Fu, "Learning to Reconstruct High Speed and High Dynamic Range Videos from Events," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2024–2033, 2021.

[12] C. Posch, D. Matolin, and W. Rainer, "A QVGA 143 dB Dynamic Range Frame-Free PWM Image Sensor with Lossless Pixel-Level Video Compression and Time-Domain CDS," IEEE Journal of Solid-State Circuits, Vol. 46, No. 1, pp. 259–275, 2010.

[13] G. Bradski and A. Kaehler, "The OpenCV Library," Miller Freeman Inc. Dr. Dobb's Journal: Software Tools for the Professional Programmer, Vol. 25, No. 11, pp. 120–123, 2000.

[14] The Blender Foundation, https://www.blender.org, (Accessed 27 Oct. 2022).

[15] D. Gehrig, M. Gehrig, J. Hidalgo-Carrio, and D. Scaramuzza, "Video to Events: Recycling Video Datasets for Event Cameras," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3586–3595, 2020.