IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 June 2023, accepted 19 July 2023, date of publication 28 July 2023, date of current version 2 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299857

RESEARCH ARTICLE

Multiple Projector Camera Calibration by Fiducial Marker Detection

MOONGU SON AND KWANGHEE KO
School of Mechanical Engineering, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea
Corresponding author: Kwanghee Ko (khko@gist.ac.kr)

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Ministry of Science and ICT (MSIT) under Grant 2021-0-00315.

ABSTRACT Projection mapping has been used for various purposes in everyday situations. A key step in projection mapping is to project images and videos without distortion through calibration, which is typically performed manually. Calibration becomes more challenging when multiple projectors are involved. To address this issue, a fully automated calibration method for a multi-projector-camera system is proposed in this paper. The projectors and cameras are assumed to be un-calibrated, and an arbitrary geometric shape of the projection surface is considered. Without using checkboards or user-provided parameters, the proposed method can automatically estimate calibration parameters for the cameras and projectors and generate compensated content for projection without distortion within a reasonable amount of time. The proposed method utilizes AprilTag markers and modified YOLOv8 with deformable convolution for robust marker detection and correspondence estimation between the projectors and cameras, providing an automatic process for completing calibration and distortion correction. Various experiments have demonstrated that the proposed method outperforms existing methods using checkerboards in terms of calibration accuracy and processing time across various camera-projector configurations. The proposed method can minimize the difficulty of projection mapping, allowing it to be used in everyday situations without requiring a certain level of knowledge about projection mapping theory and related hardware.

INDEX TERMS Multiple projector-camera calibration, marker detection, fiducial marker, deformable convolution, distortion correction.

I. INTRODUCTION

Projection mapping, also known as spatial augmented reality (SAR), is used to project images and videos onto the surface of a real object, creating various visual effects for artistic or commercial purposes. It provides several advantages over traditional screen-based augmented reality (AR). SAR can provide immersive experiences for multiple users simultaneously, does not require screen-based devices, such as mobile devices or smart glasses, and is more scalable in size than AR. These advantages, along with rapid advancements of hardware, have made it more attractive in a wide range of areas such as interior design, fine arts, exhibitions, and performances.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

Projection mapping requires projector-camera calibration, which is a process used to align and calibrate a projector and a camera to achieve accurate and precise projections onto complex surfaces. The projector and the camera have different characteristics such as lens distortion, and the geometric shape of the surface is complex. Due to this, the projected images may not match the desired content, resulting in serious artifacts on the projected images. The problem becomes even more challenging when a multiple projector-camera setup is considered.

Structured light has been used to calibrate projectors and cameras. Various structured light patterns are projected onto the target surface, and the projected patterns are then captured by the camera. The projected and captured patterns are then encoded and decoded to establish pixel-wise correspondences between the projection and camera image planes.

In this context, the projectors are treated as inverted cameras, enabling them to capture the target surface. Once the correspondences are obtained, Zhang's method [1] can be employed for calibration. Moreno and Taubin [2] utilized a sequence of encoded gray bar images for accurate calibration. Gray bar images are projected onto the target surface, and the projected images are then captured and processed to establish correspondence. However, this method requires the projection-capture of each image, resulting in a relatively long processing time. Huang et al. [3] employed a single-color grid pattern for projector-camera calibration to simplify the calibration process. However, it faces a challenge in decoding the color-grid pattern when a textured projection surface is used. Additionally, a flat calibration board is still required for calibration.

Calibration based on Zhang's approach requires positioning the calibration board in such a way that a significant portion of the camera's field of view captures the board for high accuracy. For example, a calibration board of at least A0 size should be utilized at two meters. Therefore, this method is not convenient, and achieving a high level of calibration accuracy is challenging [4]. Additionally, the application of this method to calibrate multi projector-camera systems is limited and impractical.

A lot of effort has been put into self-calibration methods that do not use a calibration board. Yamazaki et al. [5] proposed a fully automatic calibration method for projector-camera systems without using a flat calibration board. However, computing the intrinsic and extrinsic parameters of the camera and projector is sensitive to the initial values used in the computation, and only one radial distortion coefficient can be estimated.

Li et al. [6]. utilized prior values for the focal length and principal points during calibration. Structured light is projected to find an unknown principal point using the focus of expansion (FOE). Projected images, generated by changing the zoom, provide a projection of the image points on the optical axis of the projector lens. The center of optical blur between images with transformed zoom values becomes the principal point. The approach is impractical because the user needs to change the zoom level manually. Moreover, its applicability is limited because it requires prior estimates of the focal length, principal points, and the size of target objects. Willi and Grundhöfer [7] proposed a method for automatically calibrating multiple projectors and RGB cameras in a static environment. This method demonstrated good performance overall. However, establishing corresponding points based on structured light is vulnerable to lighting conditions, and a large error may occur during the initial calibration based on the Epipolar geometry. Tehrani et al. [8] utilized the EXIF (Exchangeable Image File Format) information of the images to estimate the internal parameters of the cameras, perform geometric registration of multi-projector systems, and estimate the parameters of the cameras and projectors.

Most of the methods developed so far utilize the Gray-code structured lighting technique to establish correspondence for calibration. However, the method has a drawback: when multiple structured light patterns are used to improve accuracy, the shooting time increases as the hardware configuration becomes more complex, which limits its applicability to systems with multiple projectors and cameras. As one solution to this problem, an image with a repeated single pattern can be considered. This method identifies the spatial relations between patterns based on their arrangement and positions encoded within a local window area. Therefore, if a part of the pattern is not detected within the window, the entire domain cannot be processed.

To solve this problem, markers capable of recognizing the pixel correspondence of the area projected by the projector can be used. Fiala [9] employed ARtag [10] markers of different densities to enable projection mapping in multiple projector systems. However, this method is limited to planar surfaces. Hu et al. [11] proposed a network that estimates posture by detecting patterns of ChArUco [12] markers. By incorporating a network that detects markers and corrects their locations using a feature point extraction network based on SuperPoint [13], it becomes possible to robustly identify the vertices of ChArUco marker patterns even in dark or obscured situations. Liu et al. [14] introduced the Ghost-DeblurGAN network for AprilTags and ArUcos. The marker detection rate was improved using a network that processed blurred marker patterns resulting from the rapid motion of the drone. Zhang et al. [15] estimated the positions and identities of markers through a two-level network. The network estimates probable Regions of interest (ROIs) for rectangular markers, distorts the estimated ROIs to create rectified patches, and detects markers using key-point detectors. They introduced a generic network that detects various markers such as AprilTags [16], Aruco [17], Artoolkitplus [18], etc. However, the method requires calibrated camera parameters. The marker detection network assumes a projected surface and is unable to handle cases where brightness changes, the projected images are obscured or blurred due to non-planar projection surfaces and projector characteristics. Yaldiz et al. [19] proposed a network for marker detection on non-planar shapes using color-encoded markers. However, this method is not robust against variations in the color or texture of the projection surface, lighting conditions, and relies on the performance of the cameras.

In this paper, a fully automatic method for calibrating multiple projector-camera systems is proposed. Given un-calibrated cameras and projectors, as well as arbitrary geometric shapes of the projection surface, the method projects markers onto the surface, captures them, detects the bounding boxes and IDs of the markers, establishes correspondence between the camera and projection image planes, estimates the intrinsic and extrinsic parameters, and distortion coefficients of the cameras and projectors, and reconstructs the 3D geometric shape of the projection surface. Once the

calibration is completed, the method generates images and projects them onto the projection surface with minimized distortion when viewed from a selected direction. The contributions of the proposed method are as follows:

- The proposed method eliminates the need for calibration checkerboards, initial parameters, or additional hardware information.
- The proposed method employs a marker detection technique utilizing deformable convolution, which enhances the robustness of marker detection and ensures precise correspondence computation.
- The proposed method is fully automatic and enables non-experts to effortlessly utilize projection mapping using multiple projectors and cameras in everyday situations, without requiring a specific level of knowledge about projection mapping theory and related hardware.

The paper is structured as follows: in Section II, the overall procedure of the proposed method is presented with illustrative diagrams. In Section III, detailed explanations are provided for preparing markers for training and a modified network for marker detection using deformable convolution. In Section IV, the camera-projector calibration method is introduced, followed by the methods of generating view based contents and edge-blending in Section V. In Section VI, the experiments demonstrating the performance of marker detection and calibration of the proposed method are presented. Finally, Section VII concludes the paper with brief discussions on the limitations of the proposed method and recommendations for future work.

II. OVERVIEW OF THE APPROACH

Consider a system with N_c un-calibrated cameras and M_p un-calibrated projectors, where N_c and M_p are the numbers of un-calibrated cameras and projectors, respectively, and $N_c > 1$ and $M_p > 0$. The projection surface has an arbitrary shape, and the cameras are set up to cover the projected area. Multiple projectors do not have to be configured to overlap.

The overall process of the proposed calibration method consists of three parts: projection and detection of marker patterns, projector-camera calibration and reconstruction, and generation of geometry-corrected pixel correspondence, as shown in Fig. 1. In the first part, an image with markers is projected onto the projection surface and captured using cameras. Next, the captured images are processed to detect the markers using the YOLOv8-based network. In the second part, the correspondence between the projection and the captured images is established using the detected markers. Projector-camera calibration is then performed through bundle adjustment to produce the 3D points of the projection surface, as well as the intrinsic and extrinsic parameters of the projectors and cameras. In the third part, the projection surface is represented by a mesh of triangles through the Delaunay triangulation method, and pixel-wise correspondences between the cameras and projectors are estimated. Finally, an image is adjusted and warped to compensate for distortion and projected onto the projection surface to obtain an image without distortion.

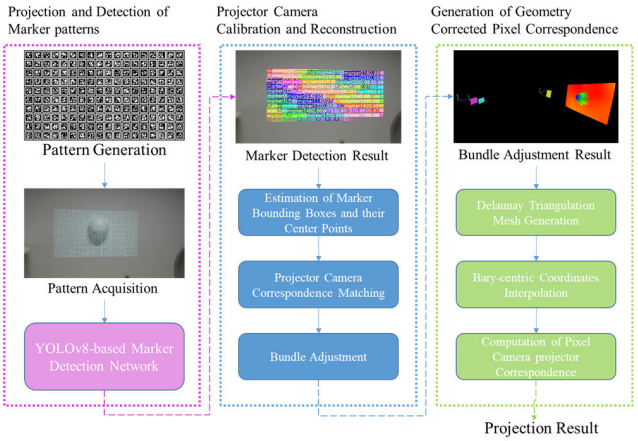

FIGURE 1. Overall process of the proposed calibration method.

III. GENERATION AND DETECTION OF MARKERS

The most crucial step in calibrating a projector-camera system is to establish accurate pixel-wise correspondence between the image planes of projectors and cameras. Methods based on multi-pattern structured light have been widely used to address this problem. However, they often fail to work because changes in light can compromise the quality of captured images, which negatively affects correspondence estimation, and the time required for projection and acquisition increases linearly with the number of devices in the system (typically, around 20 seconds per set of one projector and one camera.) Therefore, they are not suitable for multi-projector-camera systems.

In this work, markers with unique patterns are used to improve correspondence estimation in the calibration process. An image with markers is projected and captured. The markers in the captured image are detected and identified to establish correspondence. The correspondence is then provided as input to subsequent processes.

A. MARKER GENERATION

Fiducial markers are designed to be easily detected and recognized by computer vision algorithms. Since each marker contains unique information, correspondence between the projected and captured images can be established once markers with the same pattern in both images are recognized.

AprilTag, a type of fiducial marker, is recommended for its high accuracy in detection and pose estimation [20], as well as its robustness against blurring phenomena that can frequently occur during projection [21]. In this work, 180 markers with unique IDs were created and used to generate an image of markers in a 10 by 80 matrix, as shown in Fig. 2.

The number of markers needs to be determined based on the resolutions of the camera and projector, the shape of the projection surface, and the computation time. 180 markers

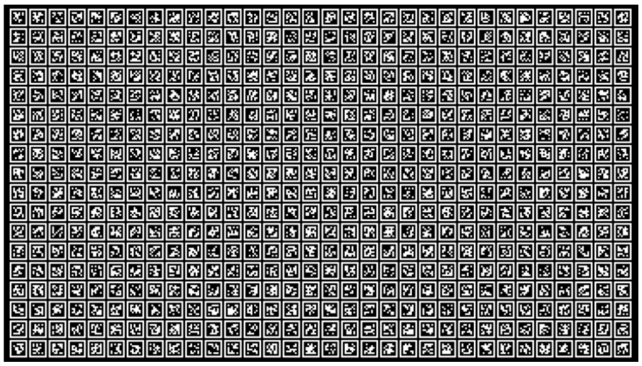

FIGURE 2. AprilTag marker pattern for projection.

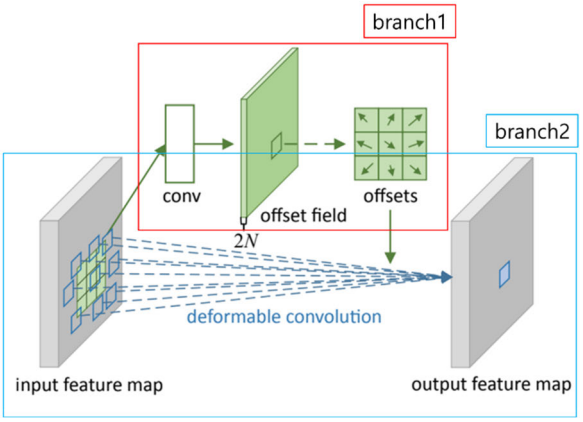

FIGURE 3. Deformable convolution layer structure.

were determined empirically for the proposed system configuration. However, different numbers of markers can be used without modifying the system.

B. MARKER DETECTION

An object detection network based on YOLOv8 [22] is used to improve the performance of marker detection. YOLOv8 is the latest deep neural network specialized for object detection and uses an anchorless model that directly predicts the center of a detected object, rather than an offset from an anchor box enclosing the object. This approach reduces the number of boxes to be predicted and speeds up NMS (Non-Maximum Suppression), which filters out redundant object detections. Additionally, YOLOv8 demonstrates better accuracy than the existing YOLOv5 [23] network in fiducial marker detection.

YOLOv8 is designed for the detection of general objects. Therefore, it may not show the expected performance in detecting specialized targets. YOLOv8 has been modified to improve the performance of detecting fiducial markers. The convolution part of YOLOv8 has been replaced with deformable convolution [24]. Deformable convolution allows for the adaptive adjustment of the receptive field of each convolutional kernel, considering the input data. Thus, it can capture features more effectively that may be deformed or misaligned in the input data compared to traditional convolution operations, which results in improved object detection. It consists of two branches, as shown in Fig. 3. Branch 1 is a convolutional layer that learns offsets to change the positions of pixels. Branch 2 generates a feature map by performing a convolution operation through filters of various shapes created by the offset information.

The existing convolution operation extracts features only from the grid used in the feature map. For instance, a 3×3 filter extracts features solely from a 3×3 region. Consider a specific region in the grid with $n+1$ pixel elements, where the position of i^{th} element is t_i. The position of the center is t_0, and the positions of the neighbouring elements are t_j ($j = 1, \ldots, n$). The weighted feature in this region, $y(t_0)$, considering the features in the neighborhood, is given as Eq. (1)

$$y(t_0) = \sum_{j=1}^{n} w(t_j) \cdot x(t_0 + t_j), \tag{1}$$

where $w(t_i)$ is the weight at t_i, and $x(t_i)$ is the feature at t_i computed by convolution. A deformable convolution introduces an offset Δt_n in Eq. (1) to produce

$$y(t_0) = \sum_{j=1}^{n} w(t_j) \cdot x(t_0 + t_j + \Delta t_n). \tag{2}$$

It allows for the extraction of features from a wider grid area. The offset can also be trained and is typically a small number. The feature at a position with the offset can be estimated through bilinear interpolation.

C. PREPARATION OF TRAINING DATA FOR MARKER DETECTION

A set of training data for fiducial marker detection can be created as follows. First, a fiducial marker pattern is created, as shown in Fig. 2. The pattern is then projected onto surfaces of diverse shapes in various environmental conditions, and the resulting images are captured to produce 2D images. Bounding boxes are then created for each marker in each image, and appropriate IDs are assigned to the bounding boxes. However, this process is time-consuming and labor-intensive and cannot cover a wide range of environmental conditions that may affect the quality of the captured images.

In this work, a method for creating training data is presented. Firstly, an image of a fiducial marker pattern is generated. Then, four augmentation methods are employed to simulate realistic conditions that may arise in a real projection environment. The first method applies Gaussian blur to introduce blurring effects to the image that may occur during projection and capture. The second method introduces various lighting conditions in an indoor environment by adjusting gamma, contrast, and brightness, and incorporating them into the image generation process. The third method simulates geometric distortion of markers when projected onto complex surfaces. It utilizes perspective and piecewise affine transformations applied to the pattern to generate distorted marker images. The last method adds Gaussian noise caused by the camera to the images.

Once a set of images is obtained, each marker is segmented and labelled with its corresponding ID. This step

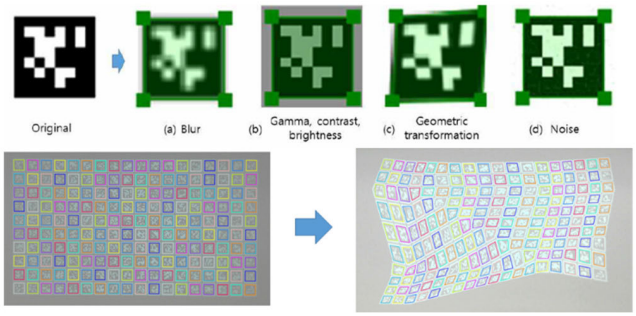

FIGURE 4. Example of the augmented markers processed by the augmentation methods, (a) -(d), and a resulting image generated by data augmentation.

can be performed automatically as pixel-wise correspondence between the images before and after augmentation can be mathematically computed.

A total of 4,000 images were generated using the augmentation methods mentioned above. Out of these, 3,600 images were used for training, and 400 images were used for validation. Fig. 4 shows an example of an augmented image.

IV. PROJECTOR CAMERA CALIBRATION

The calibration of a projector-camera system is important in any application involving projectors and cameras. A camera has an image plane onto which the real world is mapped through lenses, while a projector has a projection plane where an image is displayed and then projected onto the surface of an object. The projector-camera calibration determines the accurate pixel-wise correspondences between the image planes of the cameras and the projection planes of the projectors by estimating the intrinsic and extrinsic parameters of the devices. The intrinsic parameters are the internal characteristics of each device, including their focal length, principal point, and lens distortion. On the other hand, the extrinsic parameters represent the rigid body transformation between the cameras and the projectors, aligning them within a reference coordinate system.

In this work, the theory of a pinhole camera model for calibration is employed. It is a linear model that transforms 3D world coordinates into 2D image coordinates and is used to describe the characteristics of the camera. The projector, on the other hand, can be viewed as an inverse camera that emits light from a lamp, passes it through a lens, and projects it onto a surface. Hence, the same model can be utilized to characterize the intrinsic parameters of the projector.

A. CORRESPONDENCE COMPUTATION

Projector-camera calibration requires pixel-wise correspondence between the image planes of the cameras and the projection planes of the projectors. The process of the correspondence computation is given as follows. An image of N_m markers with unique IDs is mapped onto the projection plane of a projector. Here, N_m is 180 as presented in Section III-A. The coordinates of the pixels corresponding to the center of each marker, u_{pk}, $(k = 1, \ldots, N_m)$, are computed in the projection plane. Then, the image is projected onto a target surface. A camera captures the projected image and maps it onto the camera's image plane. The proposed detection network detects each marker from the captured image, estimates the ID and bounding box of each marker, and determines the pixel coordinates of the center of each box, u_{ci} $(i = 1, \ldots, n_d)$. Here, the number of the detected markers is n_d, and ideally, $n_d = N_m$. Subsequently, the n_d correspondences between the center pixels of the markers in the projection plane and the image plane can be established by matching the pixels with the same IDs in both planes. The same process is applied to the remaining cameras and projectors to establish correspondences between them.

B. PROJECTOR CAMERA MODEL

An ideal pinhole camera model as given in Eq. (3), which does not consider lens distortion, is defined by Hartley and Zisserman [25].

$$\boldsymbol{sp} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{t}]\boldsymbol{P}, \tag{3}$$

where $\mathbf{P} = (X, Y, Z)^T \in \mathbf{R}^3$ is a vector representing a 3D point in the world coordinate system, $\mathbf{p} = (x_p, y_p)^T \in \mathbf{R}^2$ denotes a pixel in the image plane, $\mathbf{K}$ is the 3×3 camera intrinsic matrix, $\mathbf{R}$ is a 3×3 rotation matrix, $\mathbf{t}$ is a translation vector, and s is a scaling factor that does not affect the camera model. The camera intrinsic matrix $\mathbf{K}$ comprises the focal lengths f_x and f_y, the principal point c_x and c_y, and the asymmetric coefficient s_k, as shown in Eq. (4).

$$\mathbf{K} = \begin{bmatrix} f_x & s_k & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

The focal lengths are given as f_x and f_y, to account for the potential variation in the distribution of image sensor cells along the horizontal and vertical directions. However, the difference is typically small, and it is generally accepted to consider $f = f_x = f_y$. The principal point (c_x, c_y) indicates the intersection of the optical axis and the image plane. The asymmetric coefficient s_k accounts for skew errors but is often ignored in practice due to their minor impact. Therefore, the intrinsic matrix for the camera-projector system can be expressed as shown in Eq. (5),

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{f} & \mathbf{0} & \boldsymbol{c_x} \\ \mathbf{0} & \boldsymbol{f} & \boldsymbol{c_y} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix} \tag{5}$$

The rotation matrix $\mathbf{R}$, represented in Euler angles, is a 3×3 matrix, and the translation vector $\mathbf{t}$ is a 3×1 column matrix. Therefore, Eq. (3) can be expressed as in Eq. (6).

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \mathbf{1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f} & \mathbf{0} & \boldsymbol{c_x} \\ \mathbf{0} & \boldsymbol{f} & \boldsymbol{c_y} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix} [\boldsymbol{R}|\boldsymbol{t}] \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \\ \mathbf{1} \end{bmatrix} \tag{6}$$

where $s = 1$.

The intrinsic and extrinsic parameters can be estimated using Eq. (6). However, real projectors and cameras cannot be precisely described by the ideal pinhole camera model. For instance, low-cost webcams with wide view angles or projectors using large lenses can introduce significant distortion. Hence, it is crucial to consider lens distortion during the calibration process. Lens distortion can be classified into two types: radial distortion and tangential distortion. They are modeled using non-linear functions defined by Zhang [1] as given in Eq. (7),

$$\begin{bmatrix} x_{pc} \\ y_{pc} \end{bmatrix} = \left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) \begin{bmatrix} x_p \\ y_p \end{bmatrix} + \begin{bmatrix} 2q_1 x_p y_p + q_2\left(r^2 + 2x_p^2\right) \\ q_1\left(r^2 + 2y_p^2\right) + 2q_2 x_p y_p \end{bmatrix}, \tag{7}$$

where $(x_p, y_p)^T$ and $(x_{pc}, y_{pc})^T$ are the pixels before and after correction, $r \cong \sqrt{x_p^2 + y_p^2}$, and k_1, k_2, k_3, q_1 and q_2 are unknown coefficients to be determined.

C. CALIBRATION BY BUNDLE ADJUSTMENT

The correspondence relations between the projection and camera image planes are utilized to estimate the intrinsic and extrinsic parameters of the projectors and cameras and to reconstruct the 3D points on the projection surface. Consider a setup with one projector and two cameras. The number of correspondences, n_c, would be the same as the number of markers projected onto the surface if all the markers in the image planes were detected. Arbitrary n_c points in 3D space, $(\mathbf{p}_1, \ldots, \mathbf{p}_{nc})$, are generated, and $\mathbf{p}_i$ is assigned ID_i, where ID_i is the ID of the i^{th} marker. Suppose that $\mathbf{p}_i (i = 1, \ldots, n_c)$ are projected onto the image plane of Camera 1. The coordinates of the i^{th} detected marker in the Camera 1's plane are u_{1i}. The coordinates of the projected p_i onto the Camera 1's image plane, $\mathbf{u}^p_{c1i} = (u^p_{c1i}, v^p_{c1i})$, are computed using Eqs. (3) and (7) with the camera's intrinsic $\mathbf{K}_{c1}$ and extrinsic parameters $\mathbf{R}_{c1}$ and $\mathbf{t}_{c1}$, as well as distortion correction for the camera's lens. The re-projection error for a camera, E_{C1}, is defined by

$$E_{C1} = \sum_{i=1}^{n_c} \left|\mathbf{u}_{1i} - \mathbf{u}^p_{c1i}\right|^2. \tag{8}$$

Similarly, the re-projection error for Camera 2, E_{C2}, and the projector, E_p, can be defined by

$$E_{C2} = \sum_{i=1}^{n_c} \left|\mathbf{u}_{2i} - \mathbf{u}^p_{c2i}\right|^2, \tag{9}$$

and

$$E_P = \sum_{i=1}^{n_c} \left|\mathbf{v}_i - \mathbf{v}^p_{pi}\right|^2 \tag{10}$$

where $\mathbf{v}_i$ is the coordinate of the center pixel of the i^{th} marker in the projection plane, $\mathbf{v}^p_{pi}$ is the coordinate of the projected $\mathbf{p}_i$ in the projection plane, which is computed using the projector's intrinsic $\mathbf{K}_p$ and extrinsic parameters $\mathbf{R}_p$ and $\mathbf{t}_p$, and distortion correction for the projector's lens. For a camera, 14 parameters need to be determined, including three intrinsic parameters (one focal length and two principal points), six extrinsic parameters (three components for translation, and three components for rotation), three radial distortion coefficients, and two tangential distortion coefficients. A projector is treated as an inverse camera; therefore, the same number of parameters is considered. The projector-camera setup can be calibrated by finding $\mathbf{P}$, K_{c1}, $\mathbf{R}_{c1}$, $\mathbf{t}_{c1}$, $\mathbf{K}_{c2}$, $\mathbf{R}_{c2}$, $\mathbf{t}_{c2}$, $\mathbf{K}_p$, $\mathbf{R}_p$, and $\mathbf{t}_p$ that minimize $E = E_{C1} + E_{C2} + E_P$. The function E is nonlinear with many parameters, which can be solved using the Levenberg-Marquardt method.

The initial conditions for optimization are set as follows. The Rodrigues rotation 3D vector is employed instead of the rotation matrix, with the default positive z-axis direction of (0, 0, 1). The translation vectors along the three axes are initialized as (0, 0, 0). The initial value for the focal length of the camera or projector is determined as the average of the horizontal and vertical values of the image resolution. The principal points for the camera and projector are initialized to be half the width and height of their respective image resolutions. All radial and tangential distortion coefficients are initially set to 0, and the values of $\mathbf{p}_i$ are initialized as (0, 0, 1). The optimization process performs simultaneous reconstruction of the projection surface and calibration of the camera and projector. Fig. 5 illustrates the intermediate and final results of the optimization process for a setup with two cameras and one projector. The top-left image shows the initial state before optimization. The top-right and bottom-left images display the adjusted positions and orientations of the projector and cameras, as well as the rough shape of the projection surface. The bottom-right image presents the final results of the optimization process, showing the reconstructed shape of the surface and the finalized intrinsic and extrinsic parameters of the cameras and projector. The optimization process in this work can be directly generalized to setups involving multiple projectors and cameras.

V. CAMERA VIEW-BASED CONTENTS GENTRATION

Since calibration is performed based on the center points of each marker, a maximum of 180 pixels are utilized during calibration. Therefore, correspondence relations between all the pixels in the image and projection planes need to be estimated to display contents with no distortion on the projection surface when viewed in the camera's direction.

A. PIXEL-WISE CORRESPONDENCE COMPUTATION

The pixel-wise correspondence between the image and projection planes can be obtained as follows. Firstly, the reconstructed 3D points after calibration are processed to generate a triangular mesh using Delaunay triangulation, which approximates the shape of the projection surface. Consider a mesh element with the vertices **A**, **B**, and **C**, as shown in Fig. 6. $\pi()$ is the projection of a point in 3D space onto a 2D plane.

A point $\mathbf{p}^+$ in 3D space is projected onto the image plane using the camera's parameters and the camera's lens distortion correction model. Suppose that the point in the image

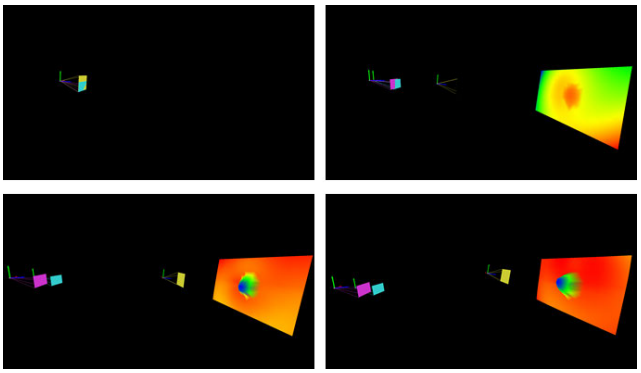

FIGURE 5. Visualization of calibration and reconstruction during optimization for two cameras and one projector.

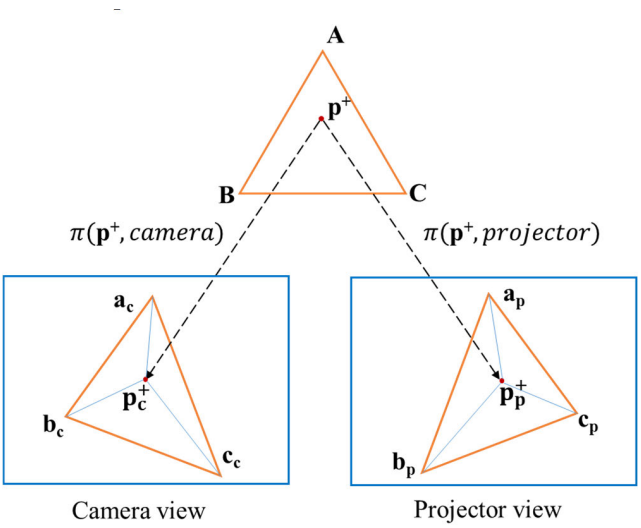

FIGURE 6. Illustration of generating a geometry compensated image.

plane is represented as $\mathbf{p}_c^+$. As the coordinates of **A**, **B**, and **C** in the image plane, denoted as $\mathbf{a}_c$, $\mathbf{b}_c$, and $\mathbf{c}_c$, are computed during calibration, the relation between $\mathbf{p}_c^+$ and $\mathbf{a}_c$, $\mathbf{b}_c$, and $\mathbf{c}_c$ can be established by

$$\mathbf{p}_c^+ = w_a\mathbf{a}_c + w_b\mathbf{b}_c + w_c\mathbf{c}_c, \tag{11}$$

where

$$w_a = \frac{\Delta p_c b_c c_c}{\Delta a_c b_c c_c}, w_b = \frac{\Delta p_c c_c a_c}{\Delta a_c b_c c_c}, w_c = \frac{\Delta p_c a_c b_c}{\Delta a_c b_c c_c},$$
$$0 \le w_a, w_b, w_c \le 1, w_a + w_b + w_c = 1,$$

and $\Delta\alpha\beta\gamma$ is the area of the triangle defined by α, β, and γ points. Then, the point that corresponds to $\mathbf{p}_c^+$ in the projection plane, $\mathbf{p}_p^+$, is estimated by

$$\mathbf{p}_p^+ = w_a\mathbf{a}_p + w_b\mathbf{b}_p + w_c\mathbf{c}_p \tag{12}$$

where $\mathbf{a}_p$, $\mathbf{b}_p$, and $\mathbf{c}_p$ are the points in the projection plane corresponding to **A**, **B**, and **C**. This method can determine the correspondence relations between all the pixels in the image and the projection planes. Consequently, an image to be projected can be adjusted based on these correspondence relations and then projected onto the surface to display a distortion-free image when viewed from the camera.

FIGURE 7. Example of edge blending using four projectors. Before (left) and after (right) edge blending.

B. EDGE BLENDING

When the projected areas of multiple projectors overlap, the overlapping region becomes significantly brighter than the non-overlapping region, leading to visual inconsistency in the projection. To address this issue, Chen's method [26] is employed. This method adjusts the brightness of the images to be projected by applying weights proportional to the distances among the pixels of each projector that correspond to the overlapping region. As a result, a seamless projected image is obtained, as depicted in Fig. 7. This figure shows that the multiple projected images are merged into a single image after edge blending.

VI. EXPERIMENT RESULTS

A. FIDUCIAL MARKER DETECTION AND CENTER PIXEL ESTIMATION

The proposed marker detection network was compared with existing networks (YOLOv5, YOLOv5 anchor free, YOLOv8, YOLOv8 anchor free.) For this test, the validation dataset presented in Section III-C was utilized. The performance was evaluated based on the box loss, the class loss, the *mAP* (mean Average Precision), and the pixel error. The box loss measures the difference between the centers, widths, and heights of the bounding boxes of the detected marker and the corresponding original marker. The classification loss represents the failure rate of ID detection.

mAP is the average of *AP*s (Average Precision). *AP* is computed as follows: Suppose m_1 is the number of IDs used in the process, m_2 is the number of IDs that are correctly predicted, m_3 is the number of IDs that are incorrectly predicted, m_4 is the number of IDs that should be detected but have not been detected, and m_5 is the number of IDs that should not be detected and have not been detected. Here, $m_1 \ge m_2 + m_4$. Prediction and recall are computed as $m_2/(m_2 + m_3)$ and $m_2/(m_2 + m_4)$, respectively. A graph is plotted using the prediction and recall values in the recall-prediction plane, and the area bounded by the graph and the recall-prediction axes becomes an *AP*. The average value of APs becomes the *mAP*, which is used to evaluate the detection algorithm. A high value of *mAP* indicates that a detection network can accurately distinguish the IDs of the detected markers.

TABLE 1. Performance comparison of the proposed and existing networks for marker detection.

Model	Box loss	Classification loss	mAP	pixel error
YOLOv5	0.26768	0.30921	0.905	0.37
YOLOv5 (anchor-free)	0.17928	0.15835	0.967	0.13
YOLOv8 (anchor-free)	0.17102	0.12716	0.995	0.05
YOLOv8 (DeConv)	0.15322	0.10027	0.995	0.04

The pixel error was measured as the average distance between the center points of the markers detected by the existing AprilTag library and the markers predicted by the network.

Table 1 presents a summary of the test results of the proposed and existing networks. The proposed method yielded the smallest values for the box loss and classification loss. Both the proposed and YOLOv8 (anchor free) networks achieved the highest *mAP* value of 0.995. Additionally, the proposed network exhibited the lowest pixel error, scoring a value of 0.04.

The tests reveal that the anchor-free methods (YOLOv5 anchor-free, YOLOv8 and the proposed methods) outperformed the anchor-box based method (YOLOv5). When a marker is deformed, the anchor-based method tends to overestimate multiple boxes containing the marker. On the other hand, the anchor-free methods can generate a tighter bounding box by first estimating the center of the marker and then computing the region enclosing the marker, as illustrated in Fig. 8. The figure shows that the anchor-free method predicts more accurate center points than the anchor-based method.

Additionally, the application of deformable convolution to the network has improved the detection performance. The existing convolution method typically assumes that the input pattern is of a grid shape and geometrically undistorted. Hence, it is not well-suited for distorted patterns as it tends to treat them differently from undistorted ones. However, deformable convolution can overcome this problem by training the shape of an appropriate receptive field for markers from input data while adaptively adjusting the convolution pattern.

The proposed network, trained with augmented data, demonstrated better detection performance for unfavorable lighting conditions and the defects of the projected images, such as blurring, as shown Fig. 9. A total of 180 markers were used in this test. The AprilTag library detected 46.67% of the total markers, as shown in Fig 9(a). YOLOv8 trained without augmented data detected only 5.56% of the markers, as shown in Fig. 9(b). In the same condition, the proposed method detected 85.56% of the markers, as presented in Fig. 9(c).

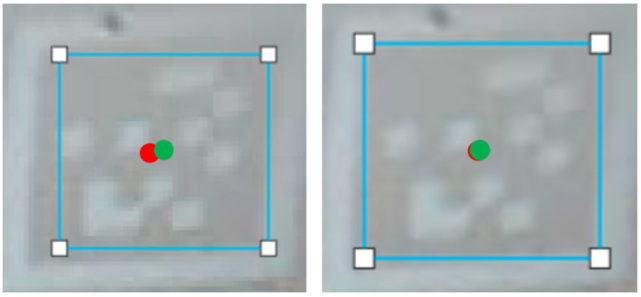

FIGURE 8. Example of the detected centers by YOLOv5 (anchor-based) and YOLOv8 with DeConv (anchor-free) marker detection. The red point is the center of the reference marker, and the green point is the center of the detected marker in each image.

FIGURE 9. Marker detection results under various projection conditions by (a) AprilTag library, (b) YOLOv8 without augmentation data, and (c) the proposed method.

The proposed method demonstrated superior performance on both textured and circular projection surfaces compared to the latest color-based marker method [19], as shown in Fig. 10. A total of 24 markers were used in this test, projected onto a textured flat surface and a half sphere, as depicted in the figure. The color-based marker method detected 13 markers on the textured surface and only one marker on the half sphere. In contrast, the proposed method detected all 24 markers on the textured surface and 13 markers on the half sphere. These results emphasize the effectiveness of the proposed method in challenging projection environments.

The general marker detection method, known as the Deeptag network [15], was compared with the proposed method. A total of 180 markers were utilized in this test, as depicted in Fig. 11. The Deeptag network detected 87 and 101 markers, while the proposed method successfully detected 167 and 171 markers, demonstrating a detection rate of 93.89%. The Deeptag network encountered challenges in identifying marker areas during the ROI detection process, as illustrated in Fig. 11, resulting in a higher number of false positive results.

B. CALIBRATION AND ACCURACY

1) CALIBRATION

The proposed method was compared with the camera calibration toolbox of GML (Graphics and Media Lab) [27] and Moreno's calibration method for estimating the camera's intrinsic parameters. Additionally, it was compared with Moreno's [2] and Willi's [7] methods for estimating the projector's intrinsic parameters. In this test, a projector-camera system with two cameras and two projectors was utilized. Here, for Camera 1 and Camera 2, ABKO APC1000 and

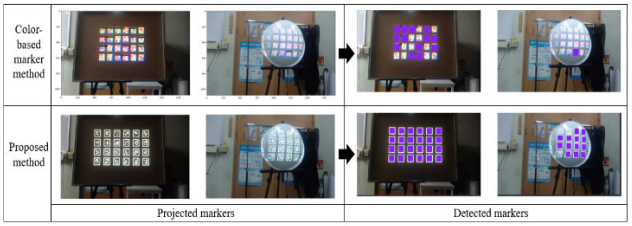

FIGURE 10. Marker detection results on the textured surface and sphere. The top row shows the detection results by color based method [19]. The bottom row shows the results by the proposed method.

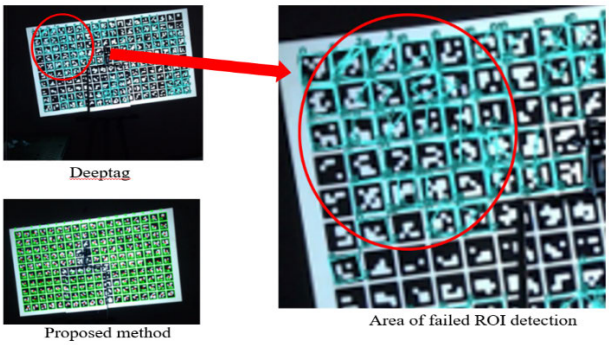

FIGURE 11. Marker detection results. The top image shows the detection results by Deeptag [15], and the bottom image the detection results by the proposed method.

TABLE 2. Camera's intrinsic parameter estimation.

Device	Methods	Focal length	Principal points X	Principal points Y
Camera 0	GML[27]	3320.64	2010.06	1245.88
	MORENO[2]	3212.04	2033.30	1324.86
	PROPOSED	3377.65	2029.98	1268.40
Camera 1	GML[27]	1501.32	958.84	538.21
	MORENO[2]	1478.32	936.13	567.72
	PROPOSED	1499.94	959.85	540.04

Logitech C920 were used. For Projector 0 and Projector 2, LG PW800 and Projector mania PJM500F were used.

Tables 2 and 3 demonstrate that the estimated parameters by the proposed method closely align with those by the existing methods using checkerboards. This indicates that the proposed method performs calibration reasonably well, even without the use of checkboards.

2) ACCURACY

The overall accuracy of the calibration can be measured by the re-projection error. In this test, the proposed method was compared with Moreno's method using a setup involving one projector and two cameras. The re-projection errors for the camera and the projector were summarized in Table 4.

TABLE 3. Peojector's intrinsic parameter estimation.

Device	Methods	Focal length	Principal points X	Principal points Y
Projector 0	MORENO[2]	*1463.24*	*922.89*	*523.51*
	WILLI[7]	*1464.32*	*927.26*	*531.26*
	PROPOSED	*1497.68*	*931.12*	*537.29*
Projector 1	MORENO[2]	*1583.09*	*922.89*	*561.34*
	WILLI[7]	*1496.32*	*937.54*	*573.21*
	PROPOSED	*1539.68*	*944.26*	*573.52*

TABLE 4. Re-projection errors for a projector and two camera setup.

Methods	Camera	Projector	Stereo
Moreno [2]	0.587	0.436	0.453
Proposed	0.326	0.173	0.281

TABLE 5. Average re-projection errors on multiple projector-camera setups.

Devices configuration	Methods	Re-projection error
2 cameras, 4 projector	Willi[7]	0.5
	PROPOSED	0.37
2 cameras, 3 projector	Tehrani[8]	0.44
	PROPOSED	0.31

The test demonstrated that the proposed method yielded smaller re-projection errors than Moreno's method. Furthermore, the proposed method was compared with Willi's and Tehrani's methods in the context of a multi-projector camera system, and the re-projection errors were summarized in Table 5.

The tests show that the proposed method achieved smaller re-projection errors for the multi-projector and camera systems than the existing methods. Fig. 12 illustrates the result of geometric correction for a non-flat surface. The figure shows that the proposed method maintains the regular shape of the checkerboard by correcting distortions, as indicated by the circles.

Fig. 13 shows the projection results on a wall with two corners before and after using the proposed method. It can be seen that the images are warped to produce a distortion-corrected image from the camera's point of view.

In these experiments, a laptop computer with an intel core i7 CPU without GPU acceleration was used. The program was implemented using Python.

The computation time of the proposed method is roughly linearly related to the number of projectors and cameras. For

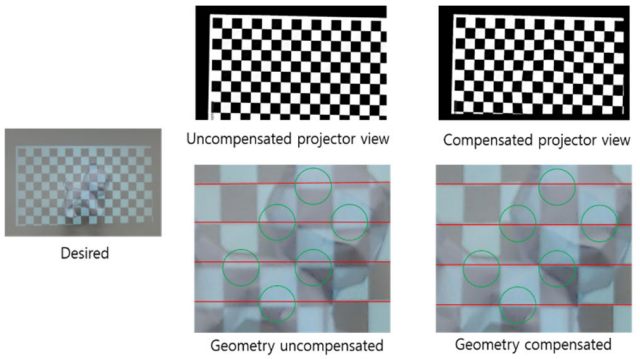

FIGURE 12. Example of geometric correction. The sides of squares and the horizontal lines are aligned after distortion correction as indicated by the circles.

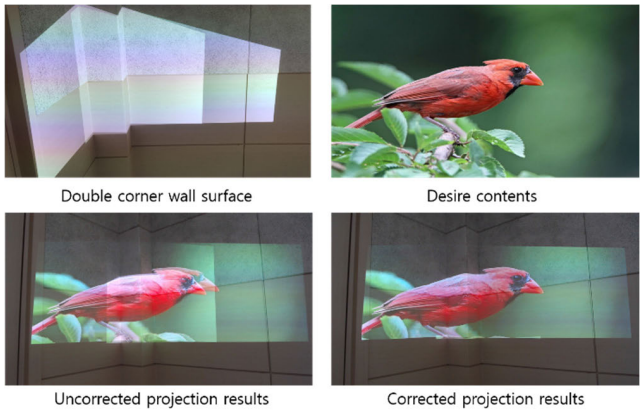

FIGURE 13. Projection result of the distortion correction for the corner surface.

a system consisting of two cameras and one projector, the proposed method completed calibration within 20 seconds: two seconds for projection and acquisition of images, one second for marker pattern detection, 10 seconds for calibration, and five seconds for creating pixel correspondence. In a setup with two cameras and four projectors, the proposed method took about one minute and 40 seconds, which is faster than Willi's method that required about two minutes and 30 seconds. Additionally, the proposed method outperformed Tehrani's method in a system with two cameras and three projectors by completing the process in about one minute, which is approximately 30 seconds faster than Tehrani's method.

Willi's method utilizes structured light, and its processing time significantly increases as the system configuration becomes more complex. On the other hand, Tehrani's method requires prior knowledge of the camera's focal length, necessitating an additional step for estimation. However, the proposed does not encounter such issues and can be extended to general configurations while maintaining processing time within a reasonable range.

VII. CONCLUSION

In this work, a fully automated calibration method for a multi-projector-camera system is proposed. This method does not require checkboards for calibration and can handle both planar and non-planar projection surfaces. The proposed method utilizes AprilTag markers and modified YOLOv8 with deformable convolution for robust marker detection and correspondence estimation, leading to improved calibration performance without the need for user-defined parameters or any hardware information. Given a projector-camera system, the proposed method can automatically estimate calibration parameters and generate distortion-compensated content for projection without distortion within a reasonable amount of time. Furthermore, the proposed method can be easily extended to multi-projector-camera systems. Various experiments demonstrated that the proposed method outperformed existing methods using checkerboards in terms of calibration accuracy and processing time across various camera-projector configurations.

The proposed method has several limitations that need to be addressed. Firstly, although the proposed method demonstrated robust detection performance under various conditions during tests, there still exist cases where lighting conditions and the reflection properties of projection surfaces negatively affect the detection performance. Therefore, a systematic method to create more extensive training data covering possible conditions and additional image processing methods, which help the network detect markers, is required. Secondly, the proposed method does not include adjustments for color in the projected image. As a result, the projected color may be affected by various factors, compromising the quality of the projected content. Since cameras are used to capture the projected image, the color of the image needs to be adjusted, and the adjusted content should be projected on the projection surface. An automatic color compensation method using this process is necessary to solve this problem. Lastly, the proposed system has not been tested for outdoor environments where problems different from indoor cases are expected. More challenging lighting and reflection properties and demanding hardware requirements are needed, which should be considered in the calibration process. These issues must be addressed to achieve the ultimate projection quality while providing an easy-to-use method in various conditions and environments. These areas are recommended for future research and development.

ACKNOWLEDGMENT

ChatGPT was used for checking the grammar of part of the manuscript.

REFERENCES

[1] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000, doi: 10.1109/34.888718.

[2] D. Moreno and G. Taubin, "Simple, accurate, and robust projector-camera calibration," in *Proc. 2nd Int. Conf. 3D Imag., Model., Process., Visualizat. Transmiss.*, Zurich, Switzerland, Oct. 2012, pp. 464–471, doi: 10.1109/3DIMPVT.2012.77.

[3] B. Huang, S. Ozdemir, Y. Tang, C. Liao, and H. Ling, "A single-shot-per-pose camera-projector calibration system for imperfect planar targets," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Munich, Germany, Oct. 2018, pp. 15–20, doi: 10.1109/ISMAR-Adjunct.2018.00023.

[4] L. Yang, J.-M. Normand, and G. Moreau, "Practical and precise projector-camera calibration," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Merida, Mexico, Sep. 2016, pp. 63–70, doi: 10.1109/ISMAR.2016.22.

[5] S. Yamazaki, M. Mochimaru, and T. Kanade, "Simultaneous self-calibration of a projector and a camera using structured light," in *Proc. CVPR Workshops*, Colorado Springs, CO, USA, Jun. 2011, pp. 60–67, doi: 10.1109/CVPRW.2011.5981781.

[6] F. Li, H. Sekkati, J. Deglint, C. Scharfenberger, M. Lamm, D. Clausi, J. Zelek, and A. Wong, "Simultaneous projector-camera self-calibration for three-dimensional reconstruction and projection mapping," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 74–83, Mar. 2017, doi: 10.1109/TCI.2017.2652844.

[7] S. Willi and A. Grundhöfer, "Robust geometric self-calibration of generic multi-projector camera systems," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Nantes, France, Oct. 2017, pp. 42–51, doi: 10.1109/ISMAR.2017.21.

[8] M. A. Tehrani, M. Gopi, and A. Majumder, "Automated geometric registration for multi-projector displays on arbitrary 3D shapes using uncalibrated devices," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 4, pp. 2265–2279, Apr. 2021, doi: 10.1109/TVCG.2019.2950942.

[9] M. Fiala, "Automatic projector calibration using self-identifying patterns," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 2005, p. 113, doi: 10.1109/CVPR.2005.416.

[10] M. Fiala, "ARTag, a fiducial marker system using digital techniques," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 2005, pp. 590–596, doi: 10.1109/CVPR.2005.74.

[11] D. Hu, D. DeTone, and T. Malisiewicz, "Deep ChArUco: Dark ChArUco marker pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 8428–8436, doi: 10.1109/CVPR.2019.00863.

[12] G. Bradski, "The OpenCV library," *Dr. Dobb's J., Softw. Tools Prof. Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[13] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 337–33712, doi: 10.1109/CVPRW.2018.00060.

[14] Y. Liu, A. Haridevan, H. Schofield, and J. Shan, "Application of ghost-DeblurGAN to fiducial marker detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Kyoto, Japan, Oct. 2022, pp. 6827–6832, doi: 10.1109/IROS47612.2022.9981701.

[15] Z. Zhang, Y. Hu, G. Yu, and J. Dai, "DeepTag: A general framework for fiducial marker design and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2931–2944, Mar. 2023, doi: 10.1109/TPAMI.2022.3174603.

[16] M. Krogius, A. Haggenmiller, and E. Olson, "Flexible layouts for fiducial tags," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Macau, China, Nov. 2019, pp. 1898–1903, doi: 10.1109/IROS40897.2019.8967787.

[17] F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, "Speeded up detection of squared fiducial markers," *Image Vis. Comput.*, vol. 76, pp. 38–47, Aug. 2018, doi: 10.1016/j.imavis.2018.05.004.

[18] H. Kato and M. Billinghurst, "Marker tracking and HMD calibration for a video-based augmented reality conferencing system," in *Proc. 2nd IEEE ACM Int. Workshop Augmented Reality (IWAR99)*, San Francisco, CA, USA, Oct. 1999, pp. 85–94, doi: 10.1109/IWAR.1999.803809.

[19] M. B. Yaldiz, A. Meuleman, H. Jang, H. Ha, and M. H. Kim, "DeepFormableTag: End-to-end generation and recognition of deformable fiducial markers," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–14, Aug. 2021.

[20] M. Kalaitzakis, B. Cain, S. Carroll, A. Ambrosi, C. Whitehead, and N. Vitzilaios, "Fiducial markers for pose estimation," *J. Intell. Robotic Syst.*, vol. 101, no. 4, Mar. 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10846-020-01307-9#citeas, doi: 10.1007/s10846-020-01307-9.

[21] D. B. D. S. Cesar, C. Gaudig, M. Fritsche, M. A. dos Reis, and F. Kirchner, "An evaluation of artificial fiducial markers in underwater environments," in *Proc. OCEANS*, Genova, Italy, May 2015, pp. 1–6, doi: 10.1109/OCEANS-Genova.2015.7271491.

[22] *Ultralytics/Ultralytics*. Accessed: Mar. 20, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[23] *Ultralytics/YOLOv5*. Accessed: Mar. 20, 2023. [Online]. Available: https://github.com/ultralytics/yolov5

[24] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9300–9308, doi: 10.1109/CVPR.2019.00953.

[25] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. New York, NY, USA: Cambridge Univ. Press, 2017.

[26] R. Chen and J.-M. Xue, "Seamless multi-projector displays using nonlinear edge blending," *Appl. Math.*, vol. 9, no. 6, pp. 764–778, 2018, doi: 10.4236/am.2018.96053.

[27] (2023). *Vuforia.com*. Accessed: Jun. 22, 2023. [Online]. Available: https://library.vuforia.com/sites/default/files/vuforia-library/docs/camera/GML_CameraCalibrationInstall_0.75.exe.zip

MOONGU SON received the B.S. degree in computer science and engineering from the University of Seoul (UOS), Seoul, South Korea, in 2015. He is currently pursuing the Ph.D. degree with the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea. His research interests include localization and augmented reality.

KWANGHEE KO received the B.S. degree in naval architecture and ocean engineering from Seoul National University, in 1995, and the M.S. degree in mechanical and ocean engineering and the Ph.D. degree in ocean engineering from MIT, in 2001 and 2003, respectively. From 2003 to 2004, he was a Postdoctoral Researcher with the Seagrant College Program, MIT. From 2004 to 2005, he was a Research Associate with the Stevens Institute of Technology. He joined the Gwangju Institute of Science and Technology, in 2006. He was an assistant professor, from 2006 to 2010, an associate professor, from 2010 to 2016, and has been a professor, since 2016. He is currently a Professor with the Gwangju Institute of Science and Technology, Republic of Korea. His research interests include CAD/CAM/CAE, geometric modeling, digital twin application, projection mapping, virtual reality, and augmented reality.