

CylinderTag: An Accurate and Flexible Marker for Cylinder-Shape Objects Pose Estimation Based on Projective Invariants

Shaoan Wang, Mingzhu Zhu, Yaoqing Hu, Dongyue Li, Fusong Yuan, and Junzhi Yu, *Fellow, IEEE*

Abstract—High-precision pose estimation based on visual markers has been a thriving research topic in the field of computer vision. However, the suitability of traditional flat markers on curved objects is limited due to the diverse shapes of curved surfaces, which hinders the development of high-precision pose estimation for curved objects. Therefore, this paper proposes a novel visual marker called CylinderTag, which is designed for developable curved surfaces such as cylindrical surfaces. CylinderTag is a cyclic marker that can be firmly attached to objects with a cylindrical shape. Leveraging the manifold assumption, the cross-ratio in projective invariance is utilized for encoding in the direction of zero curvature on the surface. Additionally, to facilitate the usage of CylinderTag, we propose a heuristic search-based marker generator and a high-performance recognizer as well. Moreover, an all-encompassing evaluation of CylinderTag properties is conducted by means of extensive experimentation, covering detection rate, detection speed, dictionary size, localization jitter, and pose estimation accuracy. CylinderTag showcases superior detection performance from varying view angles in comparison to traditional visual markers, accompanied by higher localization accuracy. Furthermore, CylinderTag boasts real-time detection capability and an extensive marker dictionary, offering enhanced versatility and practicality in a wide range of applications. Experimental results demonstrate that the CylinderTag is a highly promising visual marker for use on cylindrical-like surfaces, thus offering important guidance for future research on high-precision visual localization of cylinder-shaped objects. The code is available at: <https://github.com/wsakobe/CylinderTag>.

Index Terms—CylinderTag, cylindrical object, fiducial marker, projective invariants, pose estimation.

1 INTRODUCTION

POSE estimation plays a crucial role in the vision-driven interacting tasks ranging from VR/AR [1], human-robot interaction (HRI) [2], to motion capture (MoCap) [3]. As more advanced applications emerge, there is a growing demand for more accurate, convenient, and environment-adaptive 6-DoF object pose estimators. Traditional methods focus on designing fiducial markers with encoded code within the pattern on the marker, enabling the recognition algorithm to detect the pattern and decode the unique information stored in different markers. These markers are rigidly mounted to objects whose poses are to be estimated and are used to provide high-precision features (e.g., corners, lines, and circles) to estimate the relative 6-DoF poses between the objects frame and the camera frame.

With the continuous development of advanced technology, the demand for high-precision pose estimation is

increasing in complex scenarios, particularly those involving curved objects. Square markers are the most widely used type of visual marker, but their recognition algorithm depends on the structural properties of the square, which can result in recognition failure due to surface deformation. As a result, these markers can only be mounted with low adhesion on curved objects, making them difficult to decode from most viewing angles. Circular markers can usually withstand a certain degree of deformation, which means they can be used on surfaces with some curvature, however, their localization relies on the fitting of circular features, and surface deformation can lead to the deformation of circular features, which can make localization inaccurate. Topological markers, which do not require canonical modeling, can mitigate the effect of surface deformation. However, they are not cyclic and require a limited viewpoint for full recognition.

Several recently developed marker systems have recognized the effects of deformation and are utilizing emerging technologies, such as deep learning, to address this issue. Through end-to-end training, a large number of images simulating marker deformation can be generated, which enhances the detection performance in the presence of deformation. Despite the significant improvement in recognition performance for deformation, existing deep learning-based marker systems are still limited to square shapes and do not account for the nature of the deformed surface, resulting in low localization accuracy.

In this work, a fiducial marker system towards the cylindrical-like surfaces called CylinderTag is proposed.

S. Wang, Y. Hu, D. Li, and J. Yu are with the State Key Laboratory for Turbulence and Complex Systems, Department of Advanced Manufacturing and Robotics, College of Engineering, Peking University, Beijing 100871, China (e-mail: wangshaoan@stu.pku.edu.cn; 2101111894@stu.pku.edu.cn; 2001111648@stu.pku.edu.cn; junzhi.yu@ia.ac.cn).

M. Zhu is with the Department of Mechanical Engineering, Fuzhou University, Fuzhou 350000, China (e-mail: mzz@fzu.edu.cn).

F. Yuan is with the National Engineering Laboratory for Digital and Material Technology of Stomatology, Center of Digital Dentistry, Peking University School and Hospital of Stomatology, Beijing 100190, China (e-mail: yuanfusong@bjmu.edu.cn).

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1312800, in part by the National Natural Science Foundation of China under Grant 62003007, and in part by the Peking University Clinical Scientist Training Program under Grant BMU2023PYJH019. (Corresponding author: Junzhi Yu.)

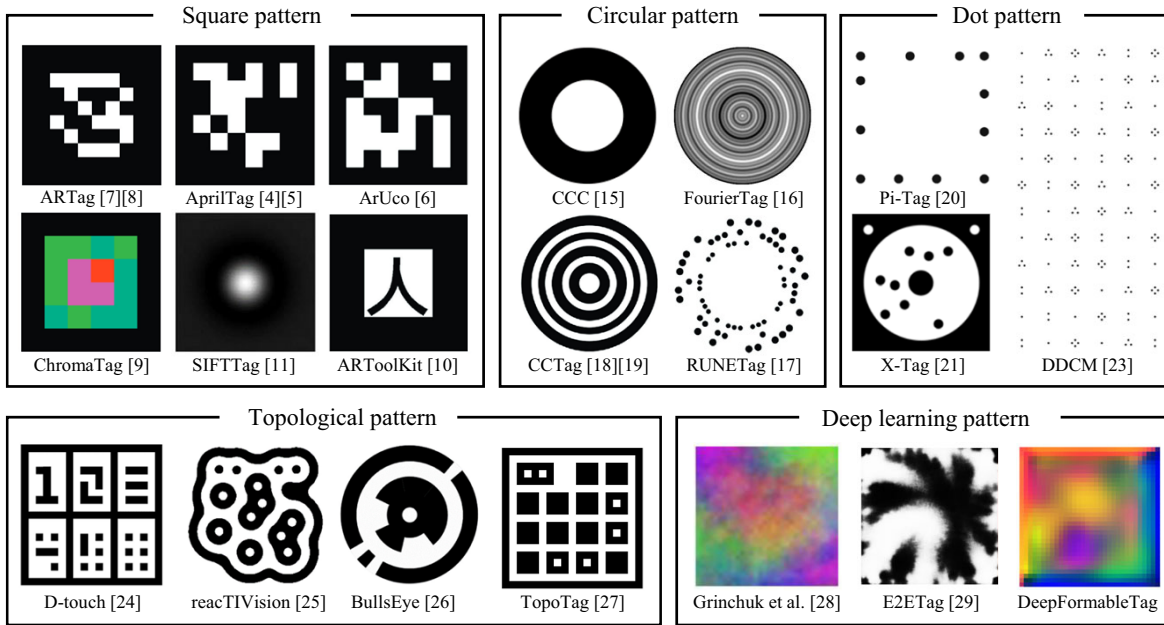


Fig. 1. Various types of visual markers, classified into square, circular, dot, topological, and deep learning patterns.

CylinderTag incorporates a long strip feature, which is based on the manifold assumption of cylindrical-like surfaces, and employs the cross-ratio as the encoding method. We propose both the marker generator and recognizer, and conduct experiments to verify the advantages of this marker in recognition and localization compared to traditional markers. The main contributions of this work are elaborated as follows:

- 1) A novel visual marker system called CylinderTag is proposed for cylindrical-like surface design, with an encoding principle based on projective invariants.
- 2) A CylinderTag marker generator based on heuristic search is proposed to efficiently generate a sufficient number of markers.
- 3) A real-time CylinderTag marker recognizer is proposed. Extensive experiments are conducted to verify its advantages in detection and localization accuracy.

The remainder of this paper is organized as follows. Section 2 describes the literature survey on existing marker systems. Section 3 provides a detailed explanation of the CylinderTag system, which comprises the marker generator and recognizer. Several experiments were conducted in Section 4 to evaluate the design, detection, localization jitter, and pose estimation performance of CylinderTag compared to traditional markers, as well as to showcase its versatility in various application scenarios. Section 5 provides the conclusions and future work of this paper.

2 RELATED WORKS

Since the invention and widespread use of 2D barcodes such as Data Matrix, Maxi code, and QR code in the late 1980s, more research has focused on developing more generalized, robust, and stable fiducial markers. Nowadays, various fiducial markers with different applications or advantages have been developed, as shown in Fig. 1. These markers can be categorized as follows:

Square pattern The square pattern usually comprises a high-contrast black square border and an internal coded pattern. AprilTag [4], [5], ArUco [6], and ARTag [7], [8] are the most commonly used marker systems in robot navigation and AR/VR today. Their internal patterns are similar by creating tight black/white regions to encode binary codes. In addition, for system robustness, the Cyclic Redundancy Check (CRC) or Reed-Solomon codes which are commonly used in the communication field are introduced to improve the capability of the error correlation mechanism. ChromaTag [9] introduces LAB colorspace that replaces black/white region into different colors over AprilTag to improve marker detection speed. ARToolKit [10] uses image cross-correlation to obtain marker information, which means the internal pattern can be arbitrary except for symmetric ones. Tanaka *et al.* designed LentiMark [11] and its improved version [12] based on VMP and FDP in order to solve the problem of pose ambiguity of planar markers. Tateno *et al.* designed Nested Marker [13] by introducing a recursive layered structure and thus realizing robust marker recognition at different scales. Schweiger *et al.* [14] proposed markers that are specifically designed for scale-invariant feature detectors SIFT and SURF. These two detectors can obtain the highest response scores, resulting in very high detection rates at different scales. Herout *et al.* [15] used De Bruijn tori for information encoding, which means that only a subregion on the marker needs to be observed to decode the corresponding information. Garrido-Jurado *et al.* [16] also focused on marker generation tasks, using mixed integer programming to accelerate the generation of ArUco. Zhu *et al.* [17], [18] proposed the binary wave function collapse (bWFC) algorithm that can arbitrarily set the shape of the pattern and the subregion with a fast-generating speed.

Circular pattern The circular pattern has also received significant attention due to the property that the quadratic

form remains invariant under projection transformation, which means that the circle can remain quadratic in most views. This property helps the detector to easily segment out the marked area. In the earliest work, Concentric Contrast Circles (CCC) [19] was first introduced in target design with a circular pattern. Subsequent work was based on this by using multiple rings of different colors. In addition, FourierTag [20] utilizes frequency images as features for adequate identification. RUNETag [21] arranges the points uniformly into concentric circles and introduces a generalized BCH code for the strongest error correction and anti-obscuration capability. CCTag [22], [23] is an improvement of CCC. Its main contribution is the robustness of motion blur.

Dot pattern Because of the excellent noise resistance of the dot feature, patterns with dots also raise some attention. Pi-Tag [24] is composed of equal-sized dots distributed around the edges of the square. It adopted cross-ratio, which is one of the projective invariants, as the encoder. Consisting of multiple dots randomly distributed in the middle of two concentric circles, X-Tag [25] similarly utilized cross-ratios and intersection preservation constraints as a reference for decoding. In addition, Watanabe *et al.* [26] proposed Extended Dot Cluster Marker (EDCM), which is an improved version of Deformable Dot Cluster Marker (DDCM) [27]. These two markers are encoded by using the number of dots in small areas as features to obtain a strong resistance to occlusion.

Topological pattern Another widely developed category of the marker is encoded based on the topological information, which demonstrates the ability to improve robustness against noise and brightness. ReacTIVision [29] provided unique identities purely with the topological structure by building a left heavy depth sequence of the region adjacency graph. BullsEye [30] consists of a central white dot surrounded by a solid black ring and one or more data rings again surrounded by a solid white ring inside a black ring with three white studs. Although this type of marker can be decoded without the limitation of shape, its localization performance is also affected by the lack of sufficient features. TopoTag [31] utilized the structure of the square pattern and only used the topological pattern as the encoder, which offered enough corner features to locate.

Deep learning pattern With the rapid development of deep learning, many scholars have pondered whether deep neural networks can be used instead of humans to design more robust and stable markers. A feasible solution was first proposed by Grinchuk *et al.* [32], which used a synthesizer network to generate patterns that are difficult for humans to understand. E2ETag [33] is an attempt to build the first end-to-end trainable fiducial marker generator, while the very small size of the code dictionary and the slow detecting speed both limit its wide application. Recently, DeepFormableTag [34] brought sufficient attention to this field. By proposing a differentiable image simulator that can render images including close-to-realistic reflection, color alteration, imaging noise, and shape deformation of markers, a large image dataset containing the majority of real-world situations is generated, resulting in the most robust detector to date. Currently, using deep learning to generate new types of markers is still in the preliminary stage, and although it has advantages that are difficult to compare with

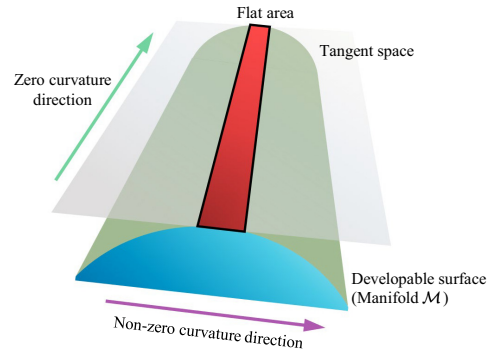


Fig. 2. **Schematic diagram of the manifold assumption.** The neighborhood of a point on the manifold can be approximated by its tangent space.

artificial markers, it is still limited to the traditional square marker shape.

3 CYLINDER TAG SYSTEM

CylinderTag is dedicated to designing a marker system with no additional accessories (like the infrared sphere-based optical MoCap system [35]) and uniformity in any viewing angle for high-precision pose estimation applications of cylindrical objects. For markers that are tightly attached to cylindrical objects, a problem that needs to be overcome is the deformation of the surface perpendicular to the cylindrical axis. However, the planes passing through the cylindrical axis intersect the cylindrical surface as lines, revealing that the cross-ratio in these lines satisfies the perspective invariants. Furthermore, under the manifold assumption, any narrow enough strip of surface in a cylinder parallel to the direction of the cylindrical axis can be regarded as a plane, providing ideas for the design of the features in CylinderTag.

3.1 Marker Design & Generator

In this subsection, the design principle and basic structure of CylinderTag are introduced in detail. Based on simulation experiments, the nature of the cross-ratio as a projection invariant is analyzed, and the design of CylinderTag is guided by the analysis results. In addition, a CylinderTag generator based on heuristic search is proposed, which can heuristically select the optimal solution according to the potential probabilities of different solutions in the current solution space, thus significantly improving the speed and efficiency of marker generation.

3.1.1 Manifold Assumption

For developable surfaces, such as cylindrical and conical surfaces, the property of zero curvature in a given direction is satisfied. That is, a straight line in the direction of zero curvature can be kept straight. Therefore, some projective invariants associated with lines can be considered to participate in the design of the marker. Cross-ratio (cr) is one of the most common projective invariants in projective geometry and is also widely used for feature encoding of visual markers [36], [37], [38]. It describes the proportional invariance of four points that are invariant in a line under a

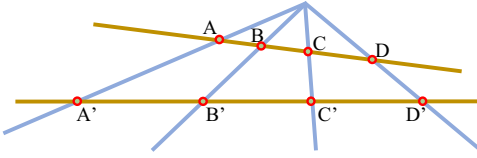


Fig. 3. **Schematic diagram of the cross-ratio.** Under the projective transformation, the cross-ratios corresponding to the four points on the same line remain unchanged.

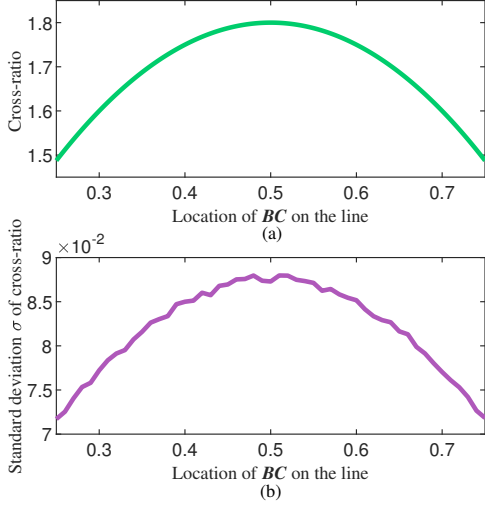


Fig. 4. **The nature of the cross-ratio.** (a) The value of the cross-ratio with respect to the location of BC . (b) The standard deviation of the cross-ratio with respect to the location of BC .

projective transformation. This property can be well applied to the zero curvature direction of the developable surface described above. For the other non-zero curvature direction, it can be conceptualized as a manifold \mathcal{M} in \mathbb{R}^3 where the local neighborhood of \mathcal{M} can be treated as a tangent space of \mathcal{M} and satisfying the planar property, as shown in Fig. 2.

3.1.2 The Nature of Cross-ratio

For four fixed points A, B, C, D located on a straight line and four points A', B', C', D' under projective transformation, as shown in Fig. 3, the cross-ratio is formulated as:

$$(A, B; C, D) = (A', B'; C', D') = \frac{AC}{CB} / \frac{AD}{DB} \quad (1)$$

In order to make better use of the cross-ratio, the properties it holds deserve to be studied and analyzed. The nonlinear nature of the cross-ratios is often discussed [39], and as Fig. 4a shows, the cross-ratios change more slowly near the midpoint. Next, if the cross-ratio is used as the encoding principle for CylinderTag, it is also necessary to analyze the error of the cross-ratio in different cases. Here, we perform simulation experiments for the general case. For four points A, B, C, D on a two-dimensional straight line of fixed length, we fix the positions of points A and D and fix the distance BC . An i.i.d. 2D Gaussian noise is applied to the positions of these points to simulate the corner localization errors. Calculate the standard deviation σ of the cross-ratio calculation when the midpoint of BC is located at different positions on the line. As illustrated

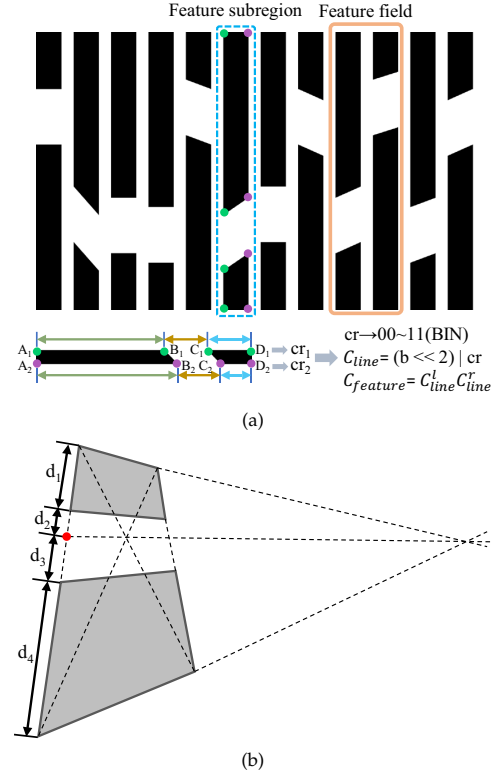


Fig. 5. **The structure of CylinderTag.** (a) Schematic diagram of the pattern and encoding principle of CylinderTag (12c2f). (b) Schematic diagram of the method to distinguish the symmetry of the cross-ratio.

in Fig. 4b, the standard deviation of the cross-ratio under Gaussian noise exhibits a trend that bears resemblance to that of the cross-ratio itself. Notably, the standard deviation is significantly higher near the midpoint compared to the ends. This indicates that under the same noise level, the standard deviation of the cross-ratio near the midpoint is larger, while the derivative of the cross-ratio near the midpoint is smaller, which together lead to a lower decoding success rate in this region. Hence, the regions closer to the ends are chosen as the encoding regions, and the regions near the midpoints are ignored, thus guaranteeing the decoding performance.

3.1.3 The Structure of CylinderTag

CylinderTag is composed of several “feature subregions”. Each “feature subregion” consists of two co-elongated quads with equal lengths of short edges. Fig. 5a illustrates the structure of CylinderTag (12c2f), where “12c” means the number of the columns (feature subregions) and “2f” means the length of the feature field. For each “feature subregion”, we assume that the direction of the long edges aligns with the direction of zero curvature on the surface. Intuitively, in this case, the four corners corresponding to the long edges remain co-linear in different poses, and their corresponding cross-ratio should also remain constant. We categorize the cross-ratio as the ID of each “feature subregion”, which can be known by detecting the corner positions on each long side and computing their corresponding cross-ratio. The cross-ratio of a “feature subregion” is thus determined by the lengths of the long sides of the two quadrangles only. In order to minimize the calculation error of the cross-ratio

due to undesirable marker pasting and to ensure a high decoding success rate, a large number of simulation and physical experiments are carried out to obtain the best cross-ratio categorization. As a result, the cross-ratio of the long side of CylinderTag is divided into four categories: 1.47, 1.54, 1.61, and 1.68. Users can adjust these values according to their needs. Unfortunately, the cross-ratio is of the nature of symmetry, which means that swapping two rectangles in the feature subregion yields the same result. It is worth noting that “feature subregions” under the same marker will always remain in the same direction, so there will only be two possible encodings of a marker in either direction. The symmetry of the cross-ratio can be dealt with by encoding the distribution of short and long edges

$$\delta = \begin{cases} 0, & \frac{d_2}{d_3} > \frac{d_1+d_2}{d_3+d_4} \\ 1, & \frac{d_2}{d_3} < \frac{d_1+d_2}{d_3+d_4} \end{cases} \quad (2)$$

where δ is the indicator of the distribution of short and long edges and $d_1 \cdots d_4$ are the distances of feature corners shown in Fig. 5b. In addition, the cross-ratios of the two long edges of the feature subregion are independent and the two cross-ratios participate in the encoding together, thus increasing the number of codes. The encoding of each feature subregion is formulated as follows

$$\begin{aligned} \{cr\} &\rightarrow \{00, 01, 10, 11\} \\ C_{line} &= (\delta \ll 2) | cr \\ C_{feature} &= (C_{line}^l \ll 1) | C_{line}^r \end{aligned} \quad (3)$$

where C_{line} is the code of a long edge, each feature subregion corresponds to the two codes C_{line}^l, C_{line}^r , and the feature code $C_{feature}$ is obtained according to these two codes. However, the code count of individual “feature subregions” makes it difficult to support large-scale scenario applications. Therefore, similar to the idea of marker fields [15], [17], CylinderTag introduces “feature fields” that promise that k consecutive encodings of feature subregions can only decode a unique marker ID, thus exponentially increasing the size of the marker dictionary.

CylinderTag offers an additional advantage of providing a larger number of features for localization as compared to traditional 4-corner quad markers. As described in [40], the increased number of features provides a better estimate of the object pose result for a variety of PnP algorithms.

3.1.4 Generation Algorithm

A “feature field” based marker generator can often be implemented using a search algorithm that takes an arbitrary feature as a node and iteratively generates a marker by judging the conflict between the marker and the current feature. However, traditional brute-force searching suffers from low efficiency and a tendency to fall into local backtracking for large coding numbers. We introduce the heuristic idea to optimize the search process. Specifically, at each step of updating markers, the number of potential solutions for each legal feature is counted, i.e., the number of legal update solutions that exist after selecting the current feature, and the feature with the highest number of potential solutions is selected for updating to maximize the feature utilization and success rate.

Algorithm 1 Heuristic Search Based Marker Generator

Input: Conflict library $\{C\}$, marker size n , marker length l , feature length f

Output: Marker ID dictionary $\{M\}$

- 1: **while** $\text{size}(\{M\}) < n$ **do**
- 2: **if** $\text{size}(M') == l$ **then**
- 3: $\{M\} \leftarrow \{M\} + M'$
- 4: **break**
- 5: **else if** $M' == \emptyset$ **then**
- 6: $\text{code} \leftarrow \text{rand}(64^f)$
- 7: **while** $\text{code} \notin \{C\}$ **do**
- 8: $\text{code} \leftarrow \text{rand}(64^f)$
- 9: **end while**
- 10: $M' \leftarrow \text{Convert code into octal}$
- 11: **else if** $\text{size}(M') < n - 1$ **then**
- 12: **for** $i \in 0, \dots, 64$ **do**
- 13: **if** $\text{Legal}(i)$ **then**
- 14: $b_i \leftarrow \text{judgeConflict}(M', i)$
- 15: **end if**
- 16: **end for**
- 17: $M' \leftarrow M' + \arg \max_k b_k$
- 18: **else**
- 19: **while** $\text{Legal}(c)$ and $\text{CyclicConflict}(M', c)$ **do**
- 20: $c \leftarrow \text{rand}(64)$
- 21: **end while**
- 22: $M' \leftarrow M' + c$
- 23: **end if**
- 24: **end while**

Marker generation is often a time-consuming operation. Many works have proposed various types of marker generation algorithms to obtain marker dictionaries in a fast and robust manner. Although marker generation only needs to be done offline, requiring less efficient generation algorithms, slow and inefficient generation algorithms still affect users. As mentioned in [6], ArUco takes about 8, 20, and 90 minutes for dictionaries of size 10, 100, and 1000 respectively, while for AprilTag, it can take several days to generate a 36-bit marker. Our marker generator enables efficient mass generation of markers thanks to a large pool of codes and heuristic searching concepts. In our use, it takes only 20 seconds to generate over 1000 markers.

3.2 Marker Recognizer

Fig. 6 shows the pipeline of the marker recognizer. To reduce the effects of noise or image blurring, the input grayscale image is downsampled to a certain scale depending on the original image size. Test results show that this procedure gives better recognition performance and a larger speed benefit compared to the original image.

3.2.1 Adaptive Threshold

For features consisting of high-contrast regions, adaptive threshold provides good segmentation of feature regions from the background by changing the input grayscale image into the binary image. Inspired by AprilTag 2, the input image is divided into tiles of 5×5 pixels to reduce computational cost compared with traditional methods. For each tile, the extrema are computed and stored. To avoid the influence

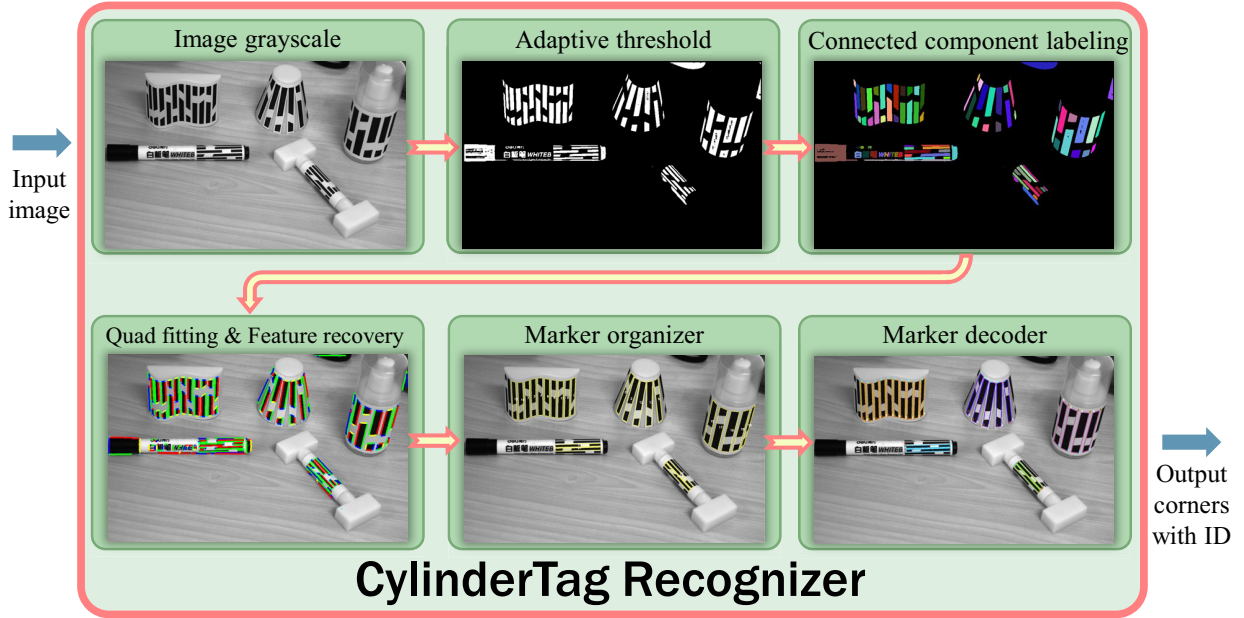


Fig. 6. The pipeline of CylinderTag recognizer.

of extreme pixels occurring in tile boundaries, the extrema pixel values p_{\max} and p_{\min} are defined as the extrema over 3×3 neighbor tiles. Thus, each pixel is given black or white in accordance with the threshold value defined as $(p_{\max} + p_{\min})/2$.

3.2.2 Connected-component Labeling

After obtaining the binary image, the feature subregions are set to black, whereas the surrounding background is expected to become white. Therefore, each feature subregion should be divided into two isolated quads. Here, a fast and accurate connected components labeling algorithm named BBDT proposed by Grana *et al.* [41] is introduced. By establishing this algorithm, all the connected regions are congregated with different labels. The size of features is taken into account so that connected areas with a disproportionately large or small proportion of pixels are automatically filtered out.

3.2.3 Quad Fitting

To recover the feature subregions, it is crucial to extract reasonable quads while excluding connected regions. However, traditional quad-fitting algorithms often suffer a high time complexity and are not suitable for use in CylinderTag, which is composed of a large number of quadrilaterals. To enhance the performance of the marker recognizer, we propose a fast quad-fitting mechanism called Extended Ramer-Douglas-Peucker (ERDP) algorithm. Ramer-Douglas-Peucker (RDP) algorithm is an iterative algorithm widely used in the GIS field for trajectory data compression. Its idea of iterative search is applied to the proposed quad-fitting algorithm. First, the border pixels of all connected regions are extracted by the 2D ray casting algorithm as shown in Fig. 7. Next, we rearrange the border pixels in connected order. A randomly selected border pixel is used

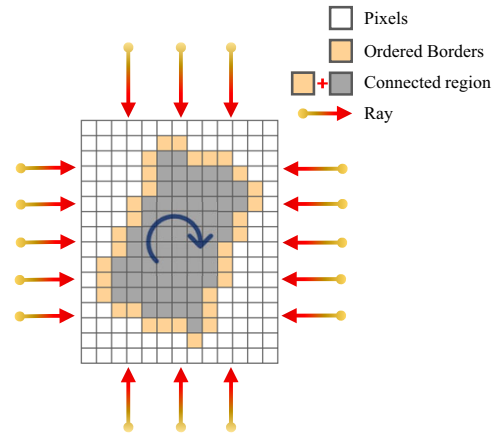


Fig. 7. 2D ray casting schematic. Each grid represents a pixel, the yellow and gray grids are the connected regions acquired in the CCL algorithm, and the yellow grids are the acquired ordered border pixels.

as the starting point, and the pixels that are at the border are arranged in order using depth-first search (DFS).

As the algorithm shows, the border closest to the center is selected as the starting point of the ordered border, and the $cost = \frac{\|\mathbf{p}_s + \mathbf{p}_{s+2} - 2\mathbf{p}_{s+1}\|_2}{\|\mathbf{p}_s - \mathbf{p}_{s+2}\|_2}$ of the triplets connected to the current starting point is cyclically judged. The point that satisfies $cost \leq T_{cost}$ is considered the starting point. The point with the farthest distance from the ordinal number is selected as the endpoint, and the distance between the middle point and the line connecting these two points is obtained. If the maximum distance does not exceed the threshold T_{line} , it is considered a legal border, and the expansion module is entered to expand this border until both ends; if the maximum distance exceeds T_{line} , the border corresponding to the maximum distance is treated as the new endpoint, and the process is repeated until four legal

Algorithm 2 Extended Ramer-Douglas-Peucker Algorithm**Input:** Ordered border pixel cluster $\{P\}$ **Output:** Four line functions $\{L\}$

```

1: Calculate center  $c$  of  $\{P\}$ 
2:  $n \leftarrow$  the size of  $\{P\}$ ,  $s \leftarrow 0$ 
3: while size $\{L\} \neq 4$  do
4:   while  $cost > T_{cost}$  do
5:     Calculate  $cost \leftarrow \frac{\|p_s + p_{s+2} - 2p_{s+1}\|_2}{\|p_s - p_{s+2}\|_2}$ 
6:     if  $cost > T_{cost}$  then
7:        $n \leftarrow n + 1$ 
8:     end if
9:   end while
10:  while True do
11:     $e \leftarrow (s + n/2) \% n$ 
12:    Calculate line  $l \leftarrow$  Fit line( $p_s, p_e$ )
13:    Calculate distances  $d_i, i=s+1, \dots, e-1$  to line  $l$ 
14:    if  $\max d_i < T_{line}$  then
15:       $\{P'\} \leftarrow$  Expand line( $p_s, p_e$ )
16:       $\{P\} \leftarrow \{P\} \setminus \{P'\}$ 
17:       $\{L\} \leftarrow \{L\} +$  Fit line( $\{P'\}$ )
18:    break
19:    else
20:       $e \leftarrow \arg \max_s d_s$ 
21:    end if
22:  end while
23: end while

```

borders are obtained.

Finally, for the candidate regions where the border pixel cluster is successfully recovered, we fit lines to the pixels of the four clusters to obtain four line functions. Thus the intersections of the lines can be calculated and considered as feature corners. Moreover, an indicator called Regional Area Coverage ($RAC = |S - Card(\{p_i\})|/S$, where S refers to the area of the quad and $Card(\{p_i\})$ refers to the number of pixels belonging to this candidate region) is introduced, candidate regions with RAC higher than the threshold T_{RAC} are considered to deviate substantially from the quads and are thus rejected.

3.2.4 Feature Recovery

At this point, we have obtained all the candidate quad corners, and the next step is to select the quad pairs that belong to the same feature subregion. Geometrically there exist several conditions as follows to assist in filtering two quads Q_i and Q_j of the same feature subregion. First, the difference between the corresponding long side angles θ_l of the two quads should not exceed a threshold T_θ , and both should be similar to the angle of the line connecting the centers of these two quads θ_c . Second, since the short sides of the two quads are close to each other, the two short side angles θ_s should remain parallel. In addition, the difference between the long sides l_l of two quads in a standard feature subregion and the distance between the quads d_{gap} , i.e., the width of the middle gap area, should be kept under a certain ratio α_{gap} . Finally, the difference between the lengths of the two short edges l_s and the lengths of the two long edges of the standard feature subregion should be kept under a certain ratio α_{len} , and the difference between the lengths of

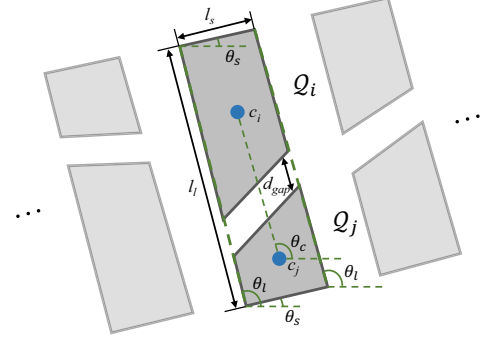


Fig. 8. Schematic diagram of the variables used in feature recovery.

the two short edges should be lower than the ratio α_s . Fig. 8 provides a detailed representation of the specific meanings of the aforementioned variables. The above criteria can be summarized as follows:

$$\begin{cases} |\Delta_{\theta_l}| \leq T_\theta \\ |\Delta_{\theta_s}| \leq T_\theta \\ |\delta_{\theta_i}| \leq T_\theta \\ |\delta_{\theta_j}| \leq T_\theta \\ d_{gap} \leq \alpha_{gap} * \sigma_l \\ \sigma_s \leq \alpha_{len} * \sigma_l \\ |\Delta_{l_s}| \leq \alpha_s * \min(l_{s_i}, l_{s_j}) \end{cases} \quad (4)$$

where $\delta_{\theta_i} = \theta_{l_i} - \theta_c$, $\delta_{\theta_j} = \theta_{l_j} - \theta_c$, and $\sigma_l = l_i + l_j$. Therefore, by filtering the above conditions, each quad can be enumerated to obtain a unique corresponding quad, thus forming a feature subregion as illustrated in Fig. 6.

3.2.5 Edge Refinement

To further improve the localization accuracy, a gradient-based double-weighted edge refinement method is implemented to obtain higher corner position accuracy. Due to the manifold assumption of a cylindrical surface, the quad of the feature subregion should lie approximately on a plane, which means the edges of the quad should remain straight. Since the inside of the quad of the feature subregion is uniformly black and the outside is white. For each corner, we uniformly sample its two adjacent edges and sample the image gradient along the normal edge to find the location of the largest image gradient in the current sampling point. However, poor placement when pasting the marker may cause the long edges of CylinderTag to be affected by curved surface deformations, making it difficult to maintain the straight-line assumption. A very intuitive idea is that edges close to corners are more likely to carry more accurate edge information, so the discrete point with the largest gradient is double-weighted according to the distance from the current corner and the magnitude of the gradient. The edge functions after refinement are obtained using PCA, and the refined feature corners are obtained based on the intersection of these edge functions.

3.2.6 Marker Organization & Feature Extraction

Feature subregions belonging to the same marker should have close area size and angle, and the neighboring feature

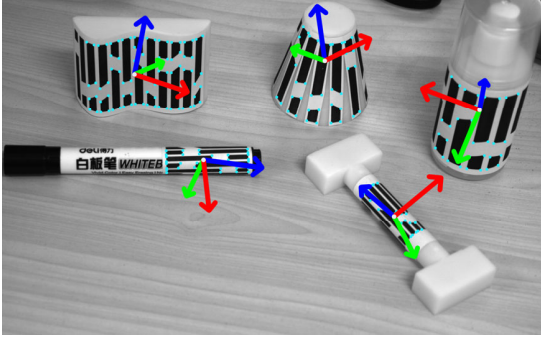


Fig. 9. Schematic diagram of curved object pose estimation.

subregions should maintain the property of parallel long edges due to continuity. For two feature subregions f_i and f_j , they are considered to belong to the same marker when the following properties are satisfied:

$$\begin{cases} |\Delta_{\theta_f}| \leq T_{\theta} \\ \vec{l}_i \cdot \vec{c}_i \vec{c}_j^T \leq T_{ver} \\ \vec{l}_j \cdot \vec{c}_i \vec{c}_j^T \leq T_{ver} \end{cases} \quad (5)$$

where θ_f refers to the long-edge angle of the feature subregion, \vec{l}_i and \vec{l}_j refer to the long-edge vectors of the two feature subregions, and c_i and c_j refer to the center of the two feature subregions.

Therefore, we introduce a marker organization method based on the Union-find algorithm to search for feature subregions that satisfy the above properties and cluster them. Once the markers are organized, we recover the cross-ratio of each feature subregion by extracting the edge lengths to obtain the corresponding ID. The corresponding subfigure in Fig. 6 shows the results of marker organization and feature extraction.

3.2.7 Marker Decoder

For any marker, the center positions of the feature subregions are sorted and the encoding vector \mathbf{c}^* is recovered according to the ratio of the center distance and the length of the short edge of the feature subregion. The encoding vector is matched with the marker ID dictionary C for maximum coverage in both directions, and the dictionary position with the highest matching degree is the marker decoding result as shown in Fig. 6.

3.3 3D Pose Recovery

3.3.1 Model Reconstruction

CylinderTag is mainly oriented to the estimation of the pose of curved objects, so compared with the a priori knowledge that the traditional marker is located on a plane, this system often needs first to obtain the object model $\{\mathbf{P}_k^W \in \mathbb{R}^3 \mid k = 1, \dots, m\}$ corresponding to the current marker with m feature corners and recover the object pose according to its model. This 3D reconstruction process can be accomplished using many classical SfM algorithms or software such as COLMAP. Here, we introduce an iterative 3D reconstruction pipeline [42]. First, a calibrated

TABLE 1
False Positive of Different Marker Systems in Indoor Scene Recognition Dataset

Marker	False number	FP
AprilTag (16h5)	6631	42.452%
AprilTag (25h9)	79	0.506%
AprilTag (36h11)	0	0%
ArUco_ORIGINAL_5×5	704	4.507%
ArUco_4×4_250	699	4.475%
ArUco_5×5_100	5	0.032%
ArUco_6×6_250	1	0.006%
RUNE-Tag	0	0%
TopoTag (3×3)	2	0.013%
TopoTag (4×4)	2	0.013%
CylinderTag (12c2f)	0	0%
CylinderTag (15c3f)	0	0%

stereo vision system is used to take a surround shot of the object to be reconstructed, and N pairs of stereo images $\{(I_{li}, I_{ri}) \mid i = 1, \dots, N\}$ are obtained. Next, an iterative optimization algorithm is introduced to minimize the following reprojection error function

$$\mathbf{P}^{W*} = \arg \min_{\mathbf{P}^W} \sum_{i \in \{(I_{li}, I_{ri})\}}^{\mathbf{p}_{ik} \in C_{li} \cup C_{ri}} \left\| \mathbf{p}_{ik} - \frac{1}{z_{ik}} \mathbf{K} \mathbf{T}_i \mathbf{P}_k^W \right\|_2^2 \quad (6)$$

where \mathbf{p}_{ik} refers to the corners in (I_{li}, I_{ri}) , z_{ik} refers to the depth of \mathbf{p}_{ik} , and \mathbf{K} is the intrinsic matrix.

3.3.2 Pose Estimation

After obtaining the 3D model $\{\mathbf{P}_k^W\}$ of the object, for any camera with known intrinsic parameters, we can obtain the matching relationship between the feature points in the captured image and the spatial location of the object model based on the decoding result of CylinderTag in the image, so as to optimize the following equations using the PnP algorithm and solve the poses \mathbf{T}^* of the object in the camera coordinate.

$$\mathbf{T}^* = \arg \min_{\mathbf{T}} \frac{1}{2} \sum_{i=1}^n \left\| \mathbf{p}_i - \frac{1}{z_i} \mathbf{K} \mathbf{T} \mathbf{P}_i^W \right\|_2^2 \quad (7)$$

4 EXPERIMENTAL VALIDATION

This section gives evaluations of CylinderTag versus traditional marker systems and shows several application scenarios for CylinderTag. The cameras used are HikVision MV-CA023-10GM industrial cameras (monochrome, 1920×1200 pixels). The computer is equipped with Intel I7-11700K (2.50 GHz). Also, the hyperparameters for the recognizer of CylinderTag are set as follows: $T_{cost} = 1.05$, $T_{line} = 1.8$, $T_{RAC} = 0.3$, $T_{\theta} = 5^\circ$, $\alpha_{gap} = 0.067$, $\alpha_{len} = 15$, $\alpha_s = 0.33$, and $T_{ver} = 0.5$.

4.1 Detection Accuracy

Detection accuracy is the most important metric to evaluate the effectiveness of a marker system in use. In this experiment, we compared the detection accuracy of CylinderTag with four traditional planar marker systems (AprilTag 3 [4], ArUco [6], TopoTag [31], and RUNETag [21]) under different viewpoints. True positive (TP) is defined as a correct detection of the marker, which contains both a

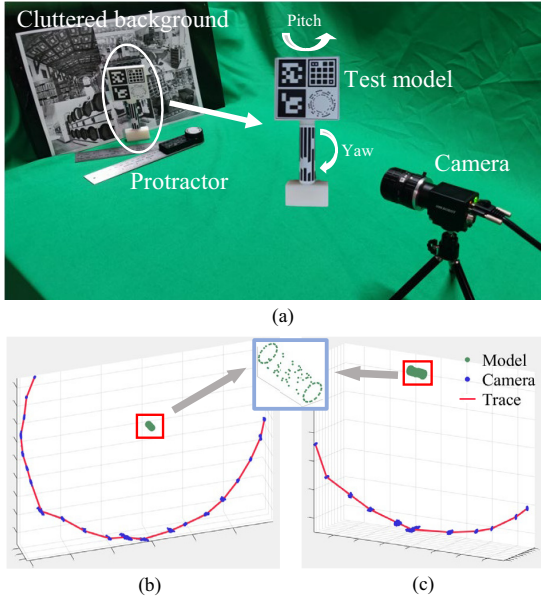


Fig. 10. **Experiment setup of detection accuracy.** (a) Real-world image acquisition process. (b) Camera pose and movement track for yaw-view experiments. (c) Camera pose and movement track for pitch-view experiments.

successful location of the marker and a correct decoding of the ID. False positive (FP) is defined as a detection result returned by the detection algorithm that does not correctly identify the location or ID. False negative (FN) is defined as the presence of a marker at a location that is not correctly identified with any marker. Furthermore, precision is defined as $\frac{TP}{TP+FP}$ while recall is defined as $\frac{TP}{TP+FN}$.

First, the Indoor Scene Recognition dataset [43] with 15,620 images from 67 categories to represent unmarked scenes is introduced to measure the FP of CylinderTag versus traditional marker systems. Table 1 shows that both Apriltag and ArUco rely heavily on their error correction capabilities for false rejection. Therefore, the FP decreases significantly as the Hamming distance increases. TopoTag detects only 2 false markers due to its tight internal topological structure arrangement. No false marker is detected by RUNETag, which is attributed to its extremely strong error correction capability. CylinderTag does not detect any false positives either due to the unique design of the features and the encoding concept of “feature field”.

Further, a 3D printed model attached with traditional markers and CylinderTag is used to test the detection accuracy under cluttered backgrounds with different view-points. Here, we tested the precision and recall of each marker system under different viewing angles corresponding to the two axes of pitch and yaw with a fixed camera, and Fig. 10 plots the position and movement track of the camera relative to the model during the experiment. The specific test procedure is: after fixing the model to a fixed viewing angle, the camera is moved to the vertical direction in the visual range while acquiring around 200 frames, and the detection algorithm is used to calculate the precision and recall of each marker system.

For cylindrical-like objects, a similar viewing area tends to be maintained on the yaw axis. For planar markers, one

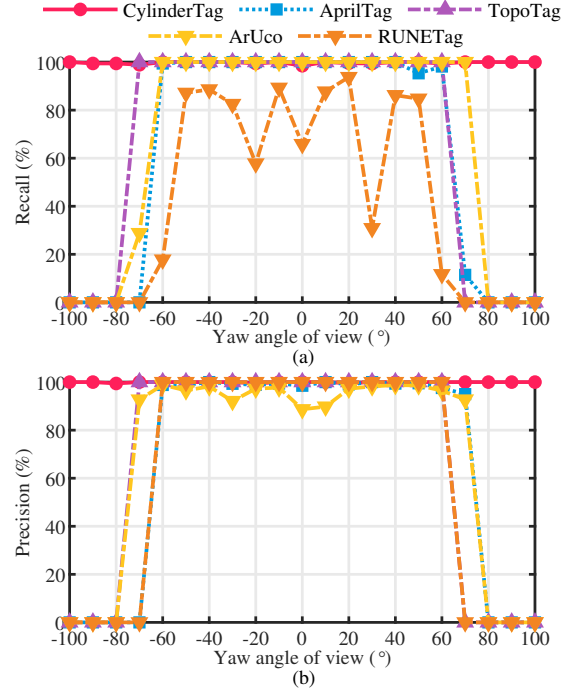


Fig. 11. Precision and recall of different marker systems at different yaw angles of view.

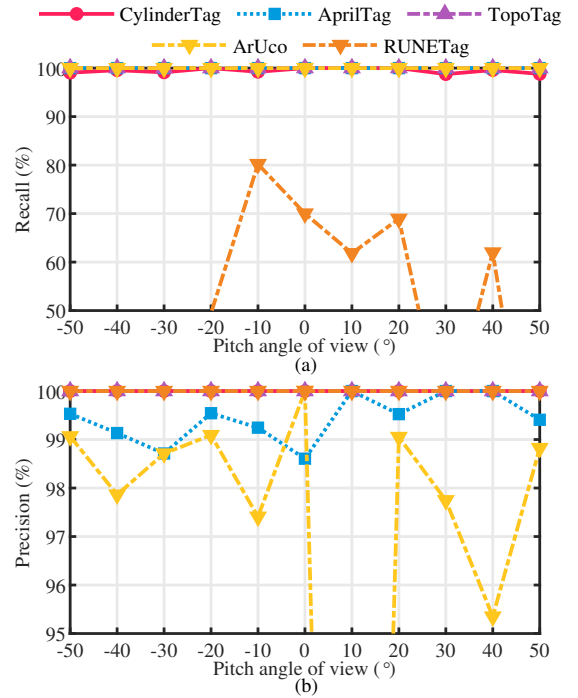


Fig. 12. Precision and recall of different marker systems at different pitch angles of view.

of the biggest problems is that a larger viewing angle can lead to a significant reduction in the visible area, thus reducing the detection performance, as Fig. 11 demonstrates. For quadrilateral markers such as AprilTag, ArUco, and TopoTag, the detection is not stable when the yaw axis angle is greater than 60°; for RUNETag, the precision is highly unstable because its features are made up of dots, which

TABLE 2
Comparison of Detection Speed and Dictionary Size

Marker	Detection speed (ms)	Type	Dictionary size
CylinderTag	35.21	12c2f	41
		15c3f	1092
		18c4f	29098
AprilTag	33.49	16h5	30
		25h9	35
		36h11	587
TopoTag	121.40	3X3	128
		4X4	16384
RUNE-Tag	648.82	RUNE-129	17000
ArUco	32.03	16h3	250
		25h7	100
		36h12	250

are difficult to be detected at small scales; CylinderTag, as a marker that tightly fits on the curved surface, can maintain a stable detection performance on the yaw axis, which reflects its detection advantages under different viewpoints.

In addition, experiments on the detection performance of the markers at different pitch angles are required. For planar markers, the yaw angle and pitch angle can be regarded as similar and can be reflected in Fig. 11 and Fig. 12. For CylinderTag, due to the projection invariance of the quadrilateral feature, its stable detection performance can still be maintained under a certain pitch angle, which verifies its high detection rate under different viewing angles.

The results similar to Table 1 are also reflected in the precision results of various markers in this experiment. The precision of CylinderTag, TopoTag, and RUNETag all remain close to 100%. In contrast, AprilTag and ArUco both had some amount of false positives.

4.2 Detection Speed & Dictionary Size

Detection speed and dictionary size are also two important metrics to measure the performance of the marker system. Table 2 lists the detection speed and dictionary size of different markers. CylinderTag, AprilTag, and ArUco all maintain ~ 30 FPS real-time detection capability, while TopoTag maintains ~ 8 FPS owing to the segmentation part; RUNETag achieves only ~ 1.5 FPS due to the extensive and time-consuming ellipse fitting operation.

In addition, the application of large scenes poses a challenge to the dictionary size of the marker system. In order to compromise sufficient error correction capability and recognition performance, the dictionary size of AprilTag and ArUco cannot exceed one thousand; RUNETag can support the generation of 17000 markers due to its heptad expansion; TopoTag achieves direct mapping from binary to ID due to its topological structure, thus obtaining sufficient dictionary size as well; CylinderTag uses cross-ratio as the encoding method, and introduces the idea of “feature field” to expand the encoding, thus providing a large enough encoding amount. In summary, compared with the problems of traditional marker systems, CylinderTag can support real-time applications in large scenarios.

4.3 Localization Jitter Evaluation

The localization accuracy of CylinderTag in real-world scenarios was further evaluated by introducing jitter error as

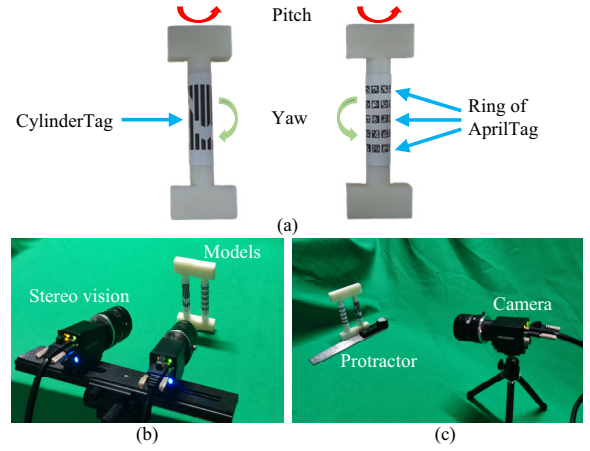


Fig. 13. **Experiment setup of localization accuracy.** (a) Schematic diagram of the 3D printed model. (b) 3D reconstruction process. (c) Real-world experimental acquisition process.

TABLE 3
Comparison of Input Images and Reprojection Errors in 3D Reconstruction

Evaluation metrics	CylinderTag		AprilTag
	w/ slope	w/o slope	
Input images	20		57
Triangulation RPE	0.236	0.217	0.245
Left-view BA RPE	0.403	0.121	0.122
Right-view BA RPE	0.288	0.092	0.106

a metric to evaluate the marker system. In fields such as AR and VR, the violent jitter of the marker can lead to an unpleasant “shaking” effect on 3D objects that are rendered on top of the marker. To measure the localization accuracy of the marker, a cylindrical-shaped 3D model was designed to test the jitter error under different viewing angles and distances. Here, with a focus on detection success rates on the curved surface, we designed a ring of 8×5 AprilTag as a comparison model.

4.3.1 Reconstruction

First, we performed 3D reconstructions described in Section 3.3 of the two marker systems and compared the reprojection errors (RPE) of these two reconstructed models. As demonstrated in Table 3, the reprojection error of the reconstruction model using CylinderTag is larger than that of AprilTag. Nevertheless, in subsequent experiments, we utilized the model to verify that its localization accuracy is significantly superior to that of AprilTag. This finding indicates that reprojection error may not be an accurate indicator of localization accuracy. Upon further examination of the CylinderTag model, we discovered that the reprojection error of corners corresponding to sloping edges was notably high. However, after eliminating these corners, the reprojection errors of the reconstructed model were all lower than those of AprilTag. This provides evidence that in the same conditions, the reconstruction accuracy of CylinderTag is superior. In addition, the compact structure of the CylinderTag allows for a convenient 3D reconstruction process which requires significantly fewer input images than required by the AprilTag.

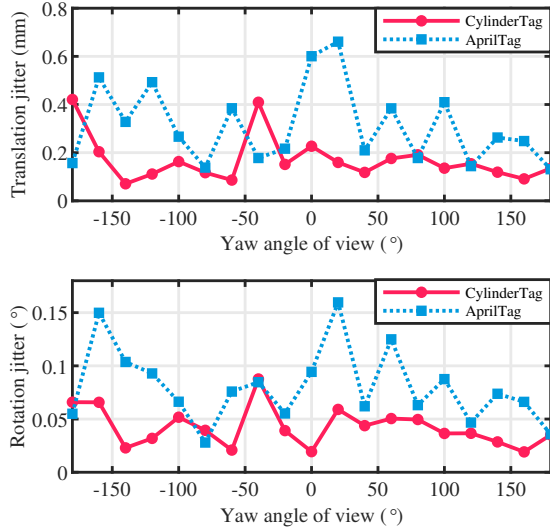


Fig. 14. Rotation and translation jitter errors at different yaw angles of view.

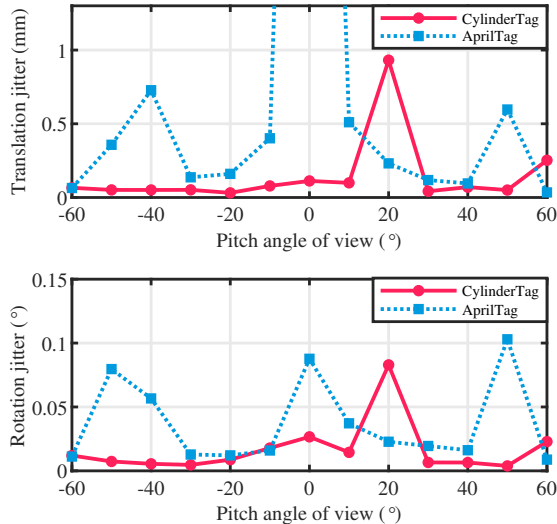


Fig. 15. Rotation and translation jitter errors at different pitch angles of view.

4.3.2 Under Different Viewing Yaw Angles

The standard deviation (STD) metric was employed to evaluate the translation and rotation jitter for each yaw angle. Fig. 14 depicts that CylinderTag exhibits significantly lower jitter errors than AprilTag. Furthermore, two-sample Kolmogorov-Smirnov tests indicate that CylinderTag surpasses AprilTag significantly in both rotation jitter ($p = 0.006$) and translation jitter ($p = 0.000$).

4.3.3 Under Different Viewing Pitch Angles

When dealing with markers on curved surfaces, the pitch angle and yaw angle cannot be considered interchangeable. Therefore, the assessment of translation and rotation jitter errors for the pitch angle was also carried out using the same evaluation process as that of the yaw angle. As shown in Fig. 15, CylinderTag displays substantially lower jitter errors than AprilTag. Additionally, two-sample Kolmogorov-Smirnov tests reveal that CylinderTag outperforms AprilTag

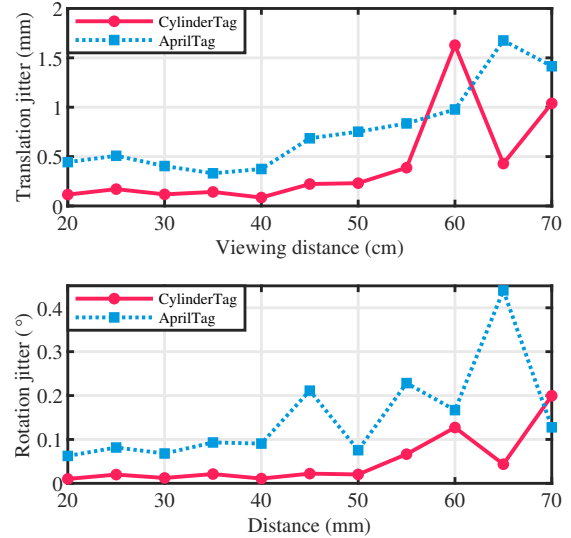


Fig. 16. Rotation and translation jitter errors at different viewing distances.

significantly in both rotation jitter ($p = 0.028$) and translation jitter ($p = 0.008$).

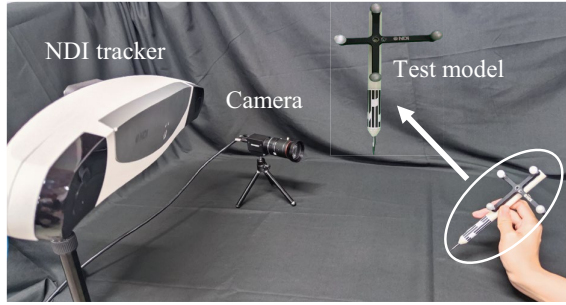
4.3.4 Under Different Distances

Different distances also have a significant impact on localization accuracy, and therefore, we assessed the jitter error for these two models at various distances. As depicted in Fig. 16, CylinderTag displays significantly lower jitter errors than AprilTag. Moreover, two-sample Kolmogorov-Smirnov tests indicate that CylinderTag outperforms AprilTag significantly in both rotation jitter ($p = 0.002$) and translation jitter ($p = 0.012$).

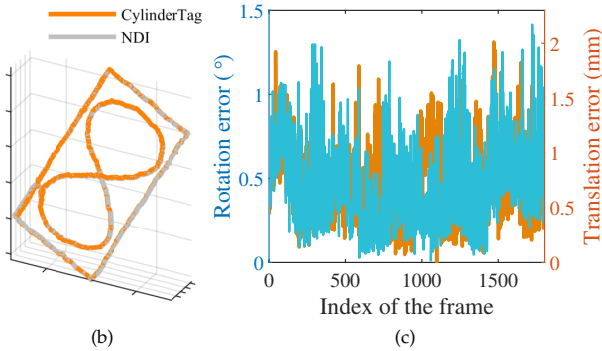
4.4 Pose Estimation Accuracy

For a more thorough evaluation of the pose estimation capability of CylinderTag, it is imperative to assess the accuracy of pose tracking for the CylinderTag attached to an object during motion. Here, similar to the methodology described in [44], we employed the NDI tracker (Polaris Vega tracker, Northern Digital, Inc.) to acquire ground truth pose information. The experiment setup for the pose estimation accuracy is shown in Fig. 17a. First, a 3D-printed cylindrical model with a tip mounted on its end was designed for evaluation. Next, the transformation relationship of the NDI tracker w.r.t the camera was obtained by tip registration and hand-eye calibration. During the experiment, the tip pose detected by the monocular camera through CylinderTag and the tip pose measured by the NDI tracker were recorded simultaneously. 2000 consecutive frames were recorded and the translation and rotation errors of the estimated pose by CylinderTag with those obtained by the NDI were calculated.

Fig. 17b shows the plot of the test model tip trajectory obtained via CylinderTag against the reference trajectory obtained by the NDI tracker, and Fig. 17c shows the errors in translation and rotation of the corresponding estimated pose for each frame. The average pose estimation error was 0.69 mm with a standard deviation of 0.38 mm for the translation



(a)



(b)

(c)

Fig. 17. **Experiment setup and results of pose estimation accuracy.** (a) Setup diagram for the pose estimation experiment. (b) Test model tip trajectories measured with CylinderTag versus ground truth trajectories. (c) Translation and rotation errors for estimated poses.

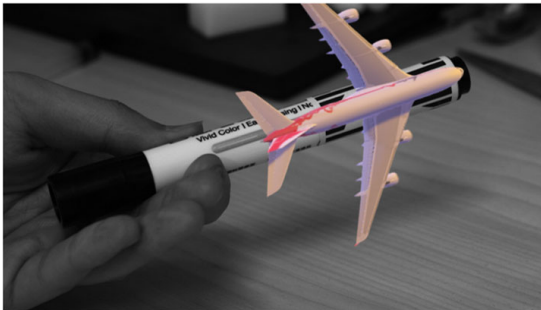


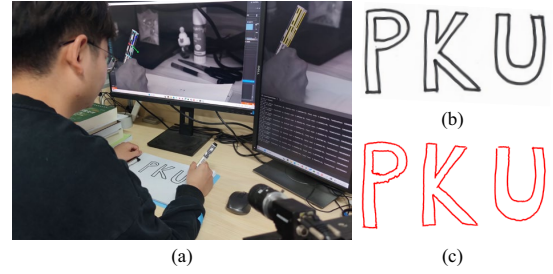
Fig. 18. **AR applications.** Acquire the pose of the pen by detecting the CylinderTag and display a model of the airplane in real time.

and 0.46° with a standard deviation of 0.20° for the rotation. Notably, the pose estimation error of CylinderTag is significantly smaller than the results reported in [44], where the best translation error was 1.43 ± 1.09 mm and the best rotation error was $0.55 \pm 0.38^\circ$.

4.5 Applications

4.5.1 Augmented Reality

Today's visual marker-based AR applications are still stuck on a flat surface, which greatly limits the interaction capability and application scenarios of AR. As shown in Fig. 18, CylinderTag provides feasibility for the implementation of the AR domain on cylindrical-like curved objects. By attaching CylinderTag to its surface, it can be photographed with a calibrated camera or cell phone camera to obtain its positional information, thus enabling more advanced augmented reality applications.



(a)

(c)

Fig. 19. **Digital pen applications.** (a) The digital pen turns a flat surface into a digital drawing surface. (b) Real handwriting obtained by the scanner. (c) Handwriting generated by the digital pen.

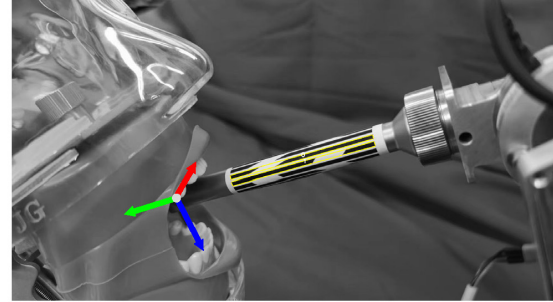


Fig. 20. **Surgical robot navigation applications.** Attaching CylinderTag tightly to the end of a surgical robot or instrument enables compact visual navigation.

4.5.2 Digital Pen

The use of vision-based digital pens has significantly advanced the fields of VR/AR. These pens are equipped with visual marker modules that allow for precise and accurate 6-DoF tracking. As a result, they are widely used in applications such as VR gaming, 2D drawing, and 3D modeling. However, today's digital pens often require customization or additional physical modules [45], which is inconvenient for users to use. With CylinderTag, users can quickly transform their pen into a digital pen that can track their posture in real-time, thus "digitizing any pen." Fig. 19 illustrates the process of transforming an ordinary pen into a digital pen with end-tracking capabilities.

4.5.3 Surgical Robot Navigation

Visual navigation is an essential technology for achieving automated surgical robots [44], [46]. Traditional visual navigation system is mainly based on an external infrared ball joint, which is unsuitable for surgeries with limited operating space, such as oral and pharyngeal surgeries, as well as minimally invasive surgeries due to their large space occupation. CylinderTag can be tightly attached to the cylindrical end-effector of the surgical robot, enabling high-precision visual navigation without occupying additional space shown in Fig. 20.

4.6 Shortcomings & Extensions

CylinderTag relies on the cross-ratio corresponding to the intersection of sub-pixel edges for encoding. However, when there is significant motion blur, as shown in Fig. 21,

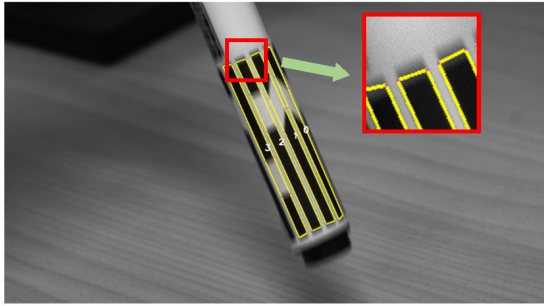


Fig. 21. **A failure case.** A CylinderTag encountering violent motion blur may cause the marker to be decoded incorrectly.

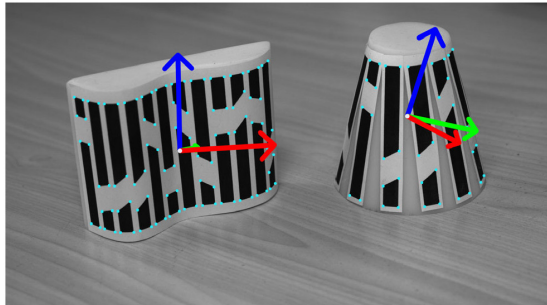


Fig. 22. **An extension case.** CylinderTag can be used to estimate poses for developable surfaces such as conical surfaces and some profiled surfaces.

the edge equation extracted by the recognizer may be biased, resulting in incorrect decoding of the cross-ratio and ultimately leading to the wrong marker ID being recovered.

As CylinderTag is primarily crafted for the mentioned developable surfaces, with cylindrical surfaces being the most prevalent, the marker can also be extended to various other surface types, including domed surfaces, conical surfaces, and some profiled surfaces as shown in Fig. 22, which significantly broadens the application scope of CylinderTag.

5 CONCLUSION

High-precision localization of curved objects offers a promising future for AR, VR, and robotics. In this paper, we propose CylinderTag, a visual marker for cylindrical-like objects, which is based on the manifold assumption and encoded using cross-ratio, and the concept of feature field is introduced to expand the encoding capacity. In addition, a marker recognizer for real-time detection and recognition is proposed, and 3D reconstruction of curved objects is realized. Various experiments, including detection rate, detection speed, dictionary size, localization jitter evaluation, and pose estimation accuracy, have demonstrated the outstanding performance of CylinderTag. It is currently the most advanced visual marker that can be applied to curved objects with developable surfaces.

In the future, we will focus on enhancing the versatility of CylinderTag by developing a more robust recognizer to identify markers on various curved surfaces. We believe that CylinderTag can serve as a source of inspiration for the development of tight visual markers for arbitrary objects.

REFERENCES

- [1] H. Chen, L. Wei, H. Liu, B. Shi, G. Zhang, and H. Zha, "MOUNT: Learning 6DoF motion prediction based on uncertainty estimation for delayed AR rendering," *IEEE Trans. Vis. Comput. Graph.*, early access, 2022, doi: 10.1109/TVCG.2022.3228807.
- [2] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1677–1734, Aug. 2021.
- [3] N. Lv, Z. Jiang, Y. Huang, X. Meng, G. Meenakshisundaram, and J. Peng, "Generic content-based retrieval of marker-based motion capture data," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 6, pp. 1969–1982, Jun. 2018.
- [4] K. Maximilian, H. Acshi, and O. Edwin, "Flexible layouts for fiducial tags," in *Proc. IEEE/RSJ Int. Conf. Intell. Rob. Syst.*, Macau, China, Nov. 2019, pp. 1898–1903.
- [5] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Daejeon, Korea, Oct. 2016, pp. 4193–4198.
- [6] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognit.*, vol. 47, no. 6, pp. 2280–2292, Jun. 2014.
- [7] M. Fiala, "ARTag, a fiducial marker system using digital techniques," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, USA, Jun. 2005, pp. 590–596.
- [8] M. Fiala, "Designing highly reliable fiducial markers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1317–1324, Jul. 2010.
- [9] J. DeGol, T. Bretl, and D. Hoiem, "ChromaTag: A colored marker and fast detection algorithm," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 1472–1481.
- [10] D. Wagner, "ARToolKitPlus for pose tracking on mobile devices," in *Proc. Comput. Vis. Winter Workshop*, St. Lambrecht, Austria, Feb. 2007, pp. 139–146.
- [11] H. Tanaka, Y. Sumi, and Y. Matsumoto, "A Visual Marker for Precise Pose Estimation based on Lenticular Lenses," in *Proc. IEEE Int. Conf. Robot. Automat.*, Saint Paul, USA, Jun. 2012, pp. 5222–5227.
- [12] H. Tanaka, K. Ogata, and Y. Matsumoto, "Solving pose ambiguity of planar visual marker by wavelike two-tone patterns," in *Proc. IEEE/RSJ Int. Conf. Intell. Rob. Syst.*, Vancouver, Canada, Sep. 2017, pp. 568–573.
- [13] K. Tateno, I. Kitahara, and Y. Ohta, "A Nested Marker for Augmented Reality," in *Proc. IEEE Virtual Reality Conf.*, Charlotte, USA, Apr. 2007, pp. 259–262.
- [14] F. Schweiger, B. Zeisl, P. Georgel, G. Schroth, E. Steinbach, and N. Navab, "Maximum detector response markers for sift and surf," in *Proc. Vis. Model. Vis. Workshop*, Braunschweig, Germany, Nov. 2009, pp. 145–154.
- [15] A. Herout, I. Szentandrás, M. Zacharias, M. Dubska, and R. Kajan, "Five shades of grey for fast and reliable camera pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, USA, Jun. 2013, pp. 1384–1390.
- [16] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, "Generation of fiducial marker dictionaries using mixed integer linear programming," *Pattern Recognit.*, vol. 51, pp. 481–491, Mar. 2016.
- [17] M. Zhu, B. He, J. Yu, F. Yuan, and J. Liu, "HydraMarker: Efficient, flexible, and multifold marker field generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5849–5861, May 2023.
- [18] S. Wang, M. Zhu, Y. Hu, D. Li, F. Yuan, and J. Yu, "Accurate Detection and Localization of Curved Checkerboard-Like Marker Based on Quadratic Form," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, Aug. 2022.
- [19] L. B. Gatrell, W. A. Hoff, and C. W. Sklair, "Robust image features: Concentric contrasting circles and their image extraction," in *Proc. Cooperative Intell. Robot. Space II*, Boston, USA, Mar. 1992, pp. 235–245.
- [20] A. Xu and G. Dudek, "Fourier tag: A smoothly degradable fiducial marker system with configurable payload capacity," in *Proc. Can. Conf. Comput. Robot Vis.*, St. John's, Canada, May 2011, pp. 40–47.
- [21] F. Bergamasco, A. Albarelli, L. Cosmo, E. Rodola, and A. Torsello, "An accurate and robust artificial marker based on cyclic codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2359–2373, Dec. 2016.

- [22] L. Calvet, P. Gurdjos, and V. Charvillat, "Camera tracking using concentric circle markers: Paradigms and algorithms," in *Proc. IEEE Int. Conf. Image Process.*, Orlando, USA, Oct. 2012, pp. 1361–1364.
- [23] L. Calvet, P. Gurdjos, C. Griwodz, and S. Gasparini, "Detection and accurate localization of circular fiducials under highly challenging conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, USA, Jun. 2016, pp. 562–570.
- [24] F. Bergamasco, A. Albarelli, and A. Torsello, "Pi-Tag: A fast image-space marker design based on projective invariants," *Mach. Vis. Appl.*, vol. 24, no. 6, pp. 1295–1310, Aug. 2013.
- [25] T. Birdal, I. Dobryden, and S. Ilic, "X-Tag: A fiducial tag for flexible and accurate bundle adjustment," in *Proc. Int. Conf. 3D Vis.*, Stanford, USA, Oct. 2016, pp. 556–564.
- [26] Y. Watanabe, T. Kato, and I. Masatoshi, "Extended dot cluster marker for high-speed 3d tracking in dynamic projection mapping," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, Nantes, France, Oct. 2017, pp. 52–61.
- [27] G. Narita, Y. Watanabe, and M. Ishikawa, "Dynamic projection mapping onto deforming non-rigid surface using deformable dot cluster marker," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 3, pp. 1235–1248, Mar. 2016.
- [28] E. Costanza, "D-touch: A consumer-grade tangible interface module and musical applications," in *Proc. Conf. Hum.-Comput. Interact.*, Crete, Greece, Sep. 2003, pp. 175–178.
- [29] M. Kaltenbrunner and R. Bencina, "reacTIVision: A computer-vision framework for table-based tangible interaction," in *Proc. Int. Conf. Tangible Embedded Interact.*, Baton Rouge, USA, Feb. 2007, pp. 69–74.
- [30] C. N. Klokmoose, J. B. Kristensen, R. Bagge, and K. Halskov, "BullsEye: High-precision fiducial tracking for table-based tangible interaction," in *Proc. ACM Int. Conf. Interact. Tabletops Surfaces*, Dresden, Germany, Nov. 2014, pp. 269–278.
- [31] G. Yu, Y. Hu, and J. Dai, "Topotag: A robust and scalable topological fiducial marker system," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 9, pp. 3769–3780, Sep. 2021.
- [32] O. Grinchuk, V. Lebedev, and V. Lempitsky, "Learnable visual markers," in *Proc. Adv. Neural Inform. Process. Syst. Conf.*, Barcelona, Spain, Dec. 2016, pp. 4143–4151.
- [33] J. B. Peace, E. Psota, Y. Liu, and L. C. Perez, "E2Etag: An end-to-end trainable method for generating and detecting fiducial markers," in *Proc. Brit. Mach. Vis. Conf.*, Manchester, UK, Sep. 2020, pp. 52–61.
- [34] 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.
- [35] J. Pan, P. Zhang, J. Wang, M. Liu, and J. Yu, "Learning for depth control of a robotic penguin: A data-driven model predictive control approach," *IEEE Trans. Ind. Electron.*, vol. 70, no. 11, pp. 11422–11432, Nov. 2023.
- [36] V. S. Tsonisp, K. V. Ch, and P. E. Trahaniaslj, "Landmark-based navigation using projective invariants," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Victoria, Canada, Oct. 1998, pp. 342–347.
- [37] R. V. Liere and J. D. Mulder, "Optical tracking using projective invariant marker pattern properties," in *Proc. IEEE Virtual Reality Conf.*, Los Angeles, USA, Apr. 2003, pp. 191–198.
- [38] F. L. Coutinho and C. H. Morimoto, "Improving head movement tolerance of cross-ratio based eye trackers," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 459–481, Jun. 2012.
- [39] D. Q. Huynh, "The cross ratio: A revisit to its probability density function," in *Proc. Brit. Mach. Vis. Conf.*, Bristol, UK, Sep. 2000, pp. 1–10.
- [40] T. Collins and A. Bartoli, "Infinitesimal plane-based pose estimation," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 252–286, Sep. 2014.
- [41] C. Grana, D. Borghesani, and R. Cucchiara, "Optimized block-based connected components labeling with decision trees," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1596–1609, Jun. 2010.
- [42] Y. Hu, M. Zhu, S. Wang, D. Li, F. Yuan, and J. Yu, "A novel lightweight navigation system for oral and maxillofacial surgery using an external curved self-identifying checkerboard," *IEEE Trans. Automat. Sci. Eng.*, early access, 2023, doi: 10.1109/TASE.2023.3241325.
- [43] A. Quattoni and A. Torralba, "Recognizing indoor scenes," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, USA, Aug. 2009, pp. 413–420.

- [44] L. Zhang, M. Ye, P. L. Chan, and G. Z. Yang, "Real-time surgical tool tracking and pose estimation using a hybrid cylindrical marker," *Int. J. Comput. Assisted Radiol. Surgery*, vol. 12, no. 6, pp. 921–930, Jun. 2017.
- [45] P. C. Wu, R. Wang, K. Kin, C. D. Twigg, S. Han, M. H. Yang, and S. Chien, "DodecaPen: Accurate 6DoF tracking of a passive stylus," in *Proc. Annu. ACM Symp. User Interface Softw. Technol.*, Quebec City, Canada, Oct. 2017, pp. 365–374.
- [46] Z. Fan, G. Chen, J. Wang, and H. Liao, "Spatial position measurement system for surgical navigation using 3-D image marker-based tracking tools with compact volume," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 2, pp. 378–389, Nov. 2017.



Shaoan Wang received the B.E. degree in mechanical engineering from the School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree in general mechanics and foundation of mechanics with the College of Engineering, Peking University, Beijing, China. His current research interests include robot vision and visual localization.



Mingzhu Zhu received the B.E., M.E., and D.E. degrees in the School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China, in 2012, 2015, and 2019, respectively. From 2019 to 2021, he was a Postdoctoral Research Fellow with BIC-ESAT, College of Engineering, Peking University, Beijing, China. In 2021, he joined Fuzhou University, as an Associate Scientist. His current research interests include computer vision and image processing.



Yaoqing Hu received the B.E. degree from University of Science and Technology Beijing, China, in 2018, and the M.E. degree from University of Science and Technology Beijing, China, in 2021. He is currently pursuing the Ph.D. degree in general mechanics and foundation of mechanics with the College of Engineering, Peking University, Beijing, China. His research interests include computer vision and robotics.



Dongyue Li received the B.E. degree from the School of Aerospace, Beijing Institute of Technology, Beijing, China, in 2020. He is currently pursuing the Ph.D. degree in general mechanics and foundation of mechanics with the College of Engineering, Peking University, Beijing, China. His current research interests include robot vision and surgical robot.



Fusong Yuan received the B.A. degree in Dentistry from Weifang Medical University, Shandong, China, in 2008, and the M.E. degree and Ph.D. degree in Prosthodontics from Peking University, Beijing, China, in 2011 and 2014, respectively.

In 2014, he joined the Peking university school and hospital of Stomatology, as a dentist, teacher, and researcher. In 2021, he was promoted to associate chief physician. His current research interest includes the application research of robotic and femtosecond laser technology in stomatology.



Junzhi Yu (Fellow, IEEE) received the B.E. degree in safety engineering and the M.E. degree in precision instruments and mechnology from the North University of China, Taiyuan, China, in 1998 and 2001, respectively, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003.

From 2004 to 2006, he was a Postdoctoral Research Fellow with the Center for Systems and Control, Peking University, Beijing. In 2006, he was an Associate Professor with the Institute of Automation, Chinese Academy of Sciences, where he became a Full Professor in 2012. In 2018, he joined the College of Engineering, Peking University, as a Tenured Full Professor. His current research interests include intelligent robots, motion control, and intelligent mechatronic systems.