2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Homography-Based Adaptive Robot Visual Tracking With Camera Parameter Convergence

Beixian Lai[2], Yongping Pan[1], Zhiwen Li[2], and Changyun Wen[3]

Abstract— **Parametric uncertainties are ubiquitous in vision-based robotic systems. However, existing adaptive visual servoing methods can not simultaneously achieve six-degree-of-freedom (6-DoF) robot pose control in the three-dimensional (3D) space and accurate camera parameter estimation. This paper considers robot manipulators with eye-to-hand monocular cameras under unknown extrinsic parameters and proposes a passivity-based kinematic control method called homography-based visual servoing with composite learning (CL-HBVS) to achieve 6-DoF robot pose tracking. The convergence of camera extrinsic parameters is achieved by designing a composite learning mechanism without the stringent condition of persistent excitation, which ensures the exact estimation of the time-varying depth and mitigates singularity in the estimated rotation matrix. The proposed method eliminates the need to calibrate camera extrinsic parameters and measure the depths of reference feature points. Experiments on a 7-DoF robot manipulator have verified the effectiveness of the proposed CL-HBVS method.**

I. Introduction

Visual servoing, typically including position-based visual servoing (PBVS) and image-based visual servoing (IBVS), utilizes visual feedback to control robot motion [1]. PBVS relies on known scene geometry to reconstruct the six-degree-of-freedom (6-DoF) pose (i.e., position and orientation) from two-dimensional (2D) images, which is then used for robot control within the feedback loop. IBVS directly employs 2D images in the feedback loop with no prior knowledge about scenes but has several drawbacks, such as local minima and image Jacobian singularity. Homography-based visual servoing (HBVS) combines PBVS and IBVS to reconstruct a 6-DoF pose using only a pair of images, which greatly alleviates the drawbacks of both PBVS and IBVS [2].

Parametric uncertainties are common in vision-based robotic systems, where the time-varying depth of feature points and uncalibrated camera parameters are two typical types of parametric uncertainties [3]. Different adaptive control strategies have been proposed to handle these parameter uncertainties [4]–[9]. IBVS methods with passivity-based adaptive laws were developed for robot manipulators in [4]–[6], but they solely achieve pixel error convergence without considering 6-DoF pose control. Note that in the above methods, the time-varying depth of image features can be expressed as a function of unknown parameters such that it can be accurately estimated if parameter estimates converge to their true values. Despite considering the time-varying depth, depth error convergence can not be achieved in these adaptive IBVS methods owing to the requirement of a stringent condition known as persistent excitation (PE) for parameter convergence. Adaptive HBVS methods were developed for robot manipulators in [7]–[9], but they require the reference extrinsic parameters of the camera to be precisely known a priori. Besides, they often assume a constant depth and rarely consider a time-varying depth.

This paper considers robotic manipulators under eye-to-hand (ETH) monocular cameras with unknown extrinsic parameters and proposes a passivity-based kinematic control solution named composite learning HBVS (CL-HBVS) to achieve 6-DoF pose tracking. First, the rotation matrix and depth ratio are extracted from homography decomposition; second, a linearly parameterized camera model is built by extracting extrinsic parameters and representing the time-varying depth via these parameters; third, an HBVS control law is proposed for 6-DoF pose tracking; lastly, a composite learning law is developed to estimate extrinsic parameters exactly, implying the exact estimation of the time-varying depth, under a condition of interval excitation (IE) that weakens PE [10]. Compared to existing adaptive HBVS methods, the distinctive feature of the proposed method is that it does not need to calibrate camera extrinsic parameters and measure the depths of reference feature points. Compared with our existing results on composite learning visual servoing [11]–[14], the current study has two distinctive features: 1) It is developed for visual tracking rather than visual regulation; 2) it is a kinematics-based design that facilitates industrial applications but takes the inner tracking error into consideration for performance enhancement.

Throughout this article, $\mathbb{R}$, $\mathbb{R}^+$, $\mathbb{R}^n$, and $\mathbb{R}^{m\times n}$ are the spaces of real numbers, positive real numbers, real n-vectors, and real $m \times n$-matrices, respectively, $\mathbb{N}$ is the set of natural numbers, $\max\{\cdot\}$ is the maximum operator, $\arccos(x)$ is the arc-cosine function, $\det(A)$ is the determinant of A, $\|\boldsymbol{x}\|$ is the Euclidean norm, $\ln(x)$ is the natural logarithm, $\mathrm{diag}(x_1, x_2, \cdots, x_n)$ is a diagonal matrix with diagonal elements x_1 to x_n, and x_i is the ith element of $\boldsymbol{x}$ with $i = 1$ to n, where $x \in \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$, and $n, m, i \in \mathbb{N}$.

II. Homography-Based Camera Model

A. Camera Projection and Euclidean Reconstruction

A robot manipulator for visual servoing tasks under an ETH monocular camera is depicted in Fig. 1. The setup comprises three Cartesian spatial frames: A world frame $\{B\}$, a camera

*This work was supported in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University, China, under Grant 23lgzy004 (*Corresponding author: Yongping Pan*).

[1]Yongping Pan is with the School of Advanced Manufacturing, Sun Yat-sen University, Shenzhen 518100, China panyongp@mail.sysu.edu.cn

[2]Beixian Lai and Zhiwen Li are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China {laibx, lizhw63}@mail2.sysu.edu.cn

[3]Changyun Wen is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 ecywen@ntu.edu.sg

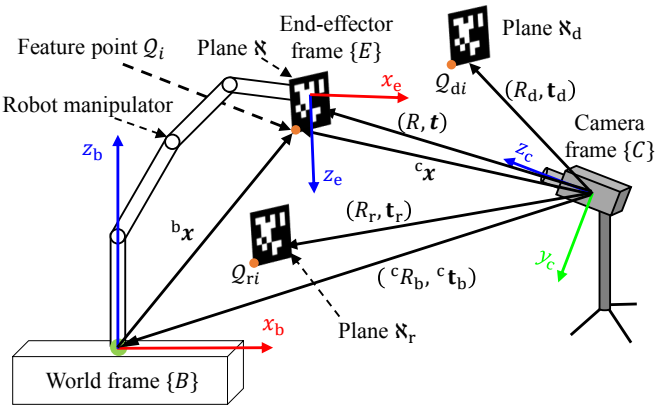

Fig. 1. A robot manipulator with an ETH monocular camera for HBVS.

frame $\{C\}$, and an end-effector frame $\{E\}$. It is assumed that: 1) The camera always captures a plane denoted as $\aleph$, which remains fixed in $\{E\}$; 2) there exist $N \geq 4$ ($\in \mathbb{N}$) non-collinear feature points $\mathcal{Q}_i$ ($i = 1$ to N) lying on $\aleph$; 3) the origin of $\{E\}$ corresponds to the location of $\mathcal{Q}_1$; 4) a fixed and constant reference plane, denoted as $\aleph_r$, contains N reference feature points $\mathcal{Q}_{ri}$; 5) a time-varying desired plane named $\aleph_d$ consists of N desired feature points $\mathcal{Q}_{di}$.

Let $R(t)$, R_r, $R_d(t) \in \mathbb{R}^{3\times3}$ denote the current, reference and desired orientations corresponding to $\aleph$, $\aleph_r$ and $\aleph_d$ in $\{C\}$, respectively, and $\boldsymbol{t}(t)$, $\boldsymbol{t}_r$, $\boldsymbol{t}_d(t) \in \mathbb{R}^3$ denote their positions in $\{C\}$, respectively. Then, let the Euclidean coordinates of $\mathcal{Q}_i$, $\mathcal{Q}_{ri}$ and $\mathcal{Q}_{di}$ in $\{C\}$ be ${}^c\boldsymbol{x}_i(t) = [x_i(t), y_i(t), z_i(t)]^T \in \mathbb{R}^3$, ${}^c\boldsymbol{x}_{ri} = [x_{ri}, y_{ri}, z_{ri}]^T \in \mathbb{R}^3$ and ${}^c\boldsymbol{x}_{di}(t) = [x_{di}(t), y_{di}(t), z_{di}(t)]^T \in \mathbb{R}^3$, and their normalized coordinates be

$$\boldsymbol{x}_i(t) := {}^c\boldsymbol{x}_i/z_i,\ \boldsymbol{x}_{ri} := {}^c\boldsymbol{x}_{ri}/z_{ri},\ \boldsymbol{x}_{di}(t) := {}^c\boldsymbol{x}_{di}/z_{di}$$

respectively. From the Euclidean geometry in Fig. 1, one gets the following relationships among $\boldsymbol{x}_i$, $\boldsymbol{x}_{di}$ and $\boldsymbol{x}_{ri}$ [2]:

$$\boldsymbol{x}_i = \overbrace{(z_{ri}/z_i)}^{\alpha_i} \overbrace{(\tilde{R} + (\tilde{\boldsymbol{t}}/d_r)\boldsymbol{n}_r^T)}^{H} \boldsymbol{x}_{ri}, \tag{1}$$

$$\boldsymbol{x}_{di} = \overbrace{(z_{ri}/z_{di})}^{\alpha_{di}} \overbrace{(\tilde{R}_d + (\tilde{\boldsymbol{t}}_{di}/d_r)\boldsymbol{n}_r^T)}^{H_d} \boldsymbol{x}_{ri} \tag{2}$$

in which $\alpha_i(t) \in \mathbb{R}^+$ and $\alpha_{di}(t) \in \mathbb{R}^+$ are the current and desired depth ratios, respectively, $H(t) \in \mathbb{R}^{3\times3}$ and $H_d(t) \in \mathbb{R}^{3\times3}$ are the current and desired Euclidean homography matrices, respectively, $\tilde{R}(t) := R(t)R_r^T \in \mathbb{R}^{3\times3}$ and $\tilde{\boldsymbol{t}}(t) := \boldsymbol{t}(t) - \tilde{R}(t)\boldsymbol{t}_r \in \mathbb{R}^3$ are the current mismatch rotation matrix and translation vector between $\aleph$ and $\aleph_r$, respectively, $\tilde{R}_d(t) := R_d(t)R_r^T \in \mathbb{R}^{3\times3}$ and $\tilde{\boldsymbol{t}}_d(t) := \boldsymbol{t}_d(t) - \tilde{R}_d(t)\boldsymbol{t}_r \in \mathbb{R}^3$ are the desired mismatch rotation matrix and translation vector between $\aleph_d$ and $\aleph_r$, respectively, $\boldsymbol{n}_r \in \mathbb{R}^3$ is a constant unit normal of $\aleph_r$ in $\{C\}$, and $d_r \in \mathbb{R}^+$ is a constant distance between $\{C\}$ and $\aleph_r$ along $\boldsymbol{n}_r$. Let the homogeneous pixel coordinates of $\mathcal{Q}_i$, $\mathcal{Q}_{ri}$ and $\mathcal{Q}_{di}$ in the image plane be $\boldsymbol{p}_i(t) = [u_i(t), v_i(t), 1]^T \in \mathbb{R}^3$, $\boldsymbol{p}_{ri} = [u_{ri}, v_{ri}, 1]^T \in \mathbb{R}^3$ and $\boldsymbol{p}_{di}(t) = [u_{di}(t), v_{di}(t), 1]^T \in \mathbb{R}^3$, respectively. Using the perspective projection model of pinhole cameras, one obtains [15]

$$\boldsymbol{p}_i = K\boldsymbol{x}_i,\ \boldsymbol{p}_{ri} = K\boldsymbol{x}_{ri},\ \boldsymbol{p}_{di} = K\boldsymbol{x}_{di} \tag{3}$$

where $K \in \mathbb{R}^{3\times3}$ is an intrinsic parameter matrix. Combining (3) with (1) and (2), one obtains

$$\boldsymbol{p}_i = \alpha_i \underbrace{KHK^{-1}}_{G} \boldsymbol{p}_{ri}, \tag{4}$$

$$\boldsymbol{p}_{di} = \alpha_{di} \underbrace{KH_dK^{-1}}_{G_d} \boldsymbol{p}_{ri} \tag{5}$$

where $G(t)$, $G_d(t) \in \mathbb{R}^{3\times3}$ denote the current and desired projective homography matrices, respectively. The current mismatch rotation matrix $\tilde{R}$ and depth ratio α_i can be obtained from G by homography decomposition [14]. In addition, the desired ones $\tilde{R}_d$ and α_{di} can be obtained from G_d.

B. Linear Parameterization of Camera Models

Let ${}^cR_b \in \mathbb{R}^{3\times3}$ and ${}^c\boldsymbol{t}_b \in \mathbb{R}^3$ denote a constant rotation matrix and a translation vector, respectively, which describe the transformation between the frames $\{C\}$ and $\{B\}$. Let ${}^b\boldsymbol{x}_i(\boldsymbol{q}) \in \mathbb{R}^4$ be a homogeneous Cartesian coordinate of the feature point $\mathcal{Q}_i$ in $\{B\}$, depending on the joint position $\boldsymbol{q}(t) \in \mathbb{R}^n$ based on the robot forward kinematics [16], where $n \in \mathbb{N}$ is the number of DoFs. Then, the perspective projection relationship in (3) can be rewritten as follows [14], [17]:

$$\boldsymbol{p}_i = KD{}^b\boldsymbol{x}_i(\boldsymbol{q})/z_i(\boldsymbol{q}) \tag{6}$$

in which $D := [{}^cR_b, {}^c\boldsymbol{t}_b] \in \mathbb{R}^{3\times4}$ is an extrinsic parameter matrix, and $z_i(\boldsymbol{q}) \in \mathbb{R}^+$ is the depth of $\mathcal{Q}_i$ in $\{C\}$ given by

$$z_i(\boldsymbol{q}) = \boldsymbol{d}_3^T{}^b\boldsymbol{x}_i(\boldsymbol{q}) \tag{7}$$

with $\boldsymbol{d}_3^T$ being the 3rd row of D [6]. Multiplying each side of (6) by $z_i(\boldsymbol{q})$ and substituting (7) into the result yield

$$\boldsymbol{p}_i\boldsymbol{d}_3^T{}^b\boldsymbol{x}_i(\boldsymbol{q}) - KD{}^b\boldsymbol{x}_i(\boldsymbol{q}) = \boldsymbol{0}. \tag{8}$$

The following key property based on (8) from [18] is useful for the adaptive control design.

Property 1: Let $\boldsymbol{\theta} := [d_{11}, d_{12}, \cdots, d_{32}, d_{33}]^T \in \mathbb{R}^{11}$ de a camera's extrinsic parameter vector that is unknown but constant, where $d_{ij} \in \mathbb{R}$ is the ijth element of D. If $d_{34} \in \mathbb{R}^+$ is known, for any given $\boldsymbol{\eta} \in \mathbb{R}^3$ and $\boldsymbol{\gamma} \in \mathbb{R}^4$, $\boldsymbol{\eta}\boldsymbol{d}_3^T\boldsymbol{\gamma} - KD\boldsymbol{\gamma}$ can be linearly parameterized by

$$\boldsymbol{\eta}\boldsymbol{d}_3^T\boldsymbol{\gamma} - KD\boldsymbol{\gamma} = \Phi^T(\boldsymbol{\eta}, \boldsymbol{\gamma})\boldsymbol{\theta} - \boldsymbol{y}(\boldsymbol{\eta}, \boldsymbol{\gamma}) \tag{9}$$

with $\Phi(\boldsymbol{\eta}, \boldsymbol{\gamma}) := B(\boldsymbol{\gamma})(\boldsymbol{k}\boldsymbol{\eta}^T - K^T) \in \mathbb{R}^{11\times3}$, $\boldsymbol{y}(\boldsymbol{\eta}, \boldsymbol{\gamma}) := d_{34}\gamma_4(K\boldsymbol{k} - \boldsymbol{\eta}) \in \mathbb{R}^3$, $\boldsymbol{k} = [0, 0, 1]^T \in \mathbb{R}^3$, and

$$B(\boldsymbol{\gamma}) := \begin{bmatrix} \boldsymbol{\gamma} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\gamma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\gamma}_{[1:3]} \end{bmatrix} \in \mathbb{R}^{11\times3} \tag{10}$$

where $\boldsymbol{\gamma}_{[1:3]} \in \mathbb{R}^3$ contains the first 3 elements of $\boldsymbol{\gamma}$. Then, substituting $\boldsymbol{\eta} = \boldsymbol{p}_i$ and $\boldsymbol{\gamma} = {}^b\boldsymbol{x}_i$ into (9), one rewrites (8) into a linearly parameterized camera model

$$\boldsymbol{y}(\boldsymbol{p}_i, {}^b\boldsymbol{x}_i) = \Phi^T(\boldsymbol{p}_i, {}^b\boldsymbol{x}_i)\boldsymbol{\theta},\ i = 1 \text{ to } N. \tag{11}$$

III. Composite Learning Visual Servoing

It is assumed that the point $\mathcal{Q}_1$, denoted by $\mathcal{Q}_\mathrm{e}$, is the origin of the end-effector frame $\{E\}$. Therefore, $\mathcal{Q}_1$, $\mathcal{Q}_{\mathrm{r}1}$ and $\mathcal{Q}_{\mathrm{d}1}$ are replaced by $\mathcal{Q}_\mathrm{e}$, $\mathcal{Q}_\mathrm{re}$ and $\mathcal{Q}_\mathrm{de}$, respectively. The control objective can be described by

$$\tilde{R} \to \tilde{R}_\mathrm{d}, \alpha \to \alpha_\mathrm{d}, \boldsymbol{p} \to \boldsymbol{p}_\mathrm{d} \ \text{ as } \ t \to \infty. \tag{12}$$

With the consideration that the plane $\aleph$ remains fixed in $\{E\}$, the orientation of the robot end-effector is represented by R. Consequently, the triplet $(R, z, \boldsymbol{p})$ uniquely determines the pose of the end-effector in the camera frame $\{C\}$. When $\tilde{R} = \tilde{R}_\mathrm{d}$ and $\alpha = \alpha_\mathrm{d}$, one obtains $R = R_\mathrm{d}$ and $z = z_\mathrm{d}$. Utilizing $z\boldsymbol{p} = K{}^\mathrm{c}\boldsymbol{x}$, if $\boldsymbol{p} = \boldsymbol{p}_\mathrm{d}$, one obtains ${}^\mathrm{c}\boldsymbol{x} = {}^\mathrm{c}\boldsymbol{x}_\mathrm{d}$, indicating that the end-effector reaches its desired pose in $\{C\}$.

A. Pose Error Definition

For control synthesis, definite a rotation error [2]

$$\boldsymbol{e}_\omega(t) := \boldsymbol{\vartheta}(t) - \boldsymbol{\vartheta}_\mathrm{d}(t) \tag{13}$$

where $\boldsymbol{\vartheta}(t)$, $\boldsymbol{\vartheta}_\mathrm{d}(t) \in \mathbb{R}^3$ are the angle-axis representations of $\tilde{R}$ and $\tilde{R}_\mathrm{d}$, respectively, given by

$$\boldsymbol{\vartheta}(t) := \boldsymbol{\mu}(t)\phi(t), \ \boldsymbol{\vartheta}_\mathrm{d}(t) := \boldsymbol{\mu}_\mathrm{d}(t)\phi_\mathrm{d}(t) \tag{14}$$

in which $\phi(t)$, $\phi_\mathrm{d}(t) \in \mathbb{R}$ are the rotation angles around unit axes $\boldsymbol{\mu}(t), \boldsymbol{\mu}_\mathrm{d}(t) \in \mathbb{R}^3$, respectively, confined to

$$-\pi < \phi(t) < \pi, \ -\pi < \phi_\mathrm{d}(t) < \pi.$$

The solutions for ϕ and $\boldsymbol{\mu}$ can be determined by

$$\begin{cases} \phi(t) = \arccos\left(\frac{1}{2}(\mathrm{tr}(\tilde{R}) - 1)\right), \\ [\boldsymbol{\mu}(t)]_\times = \frac{\tilde{R} - \tilde{R}^T}{2\sin(\phi)} \end{cases} \tag{15}$$

where $[\cdot]_\times$ is a skew symmetry operator [14]. Likewise, with $\tilde{R}_\mathrm{d}$ being obtained from homography decomposition, the solutions for ϕ_d and $\boldsymbol{\mu}_\mathrm{d}$ can also be determined. Differentiating $\boldsymbol{e}_\omega$ with respect to time t yields

$$\dot{\boldsymbol{e}}_\omega(t) = L_\omega(\boldsymbol{\vartheta}){}^\mathrm{c}R_\mathrm{b}\,{}^\mathrm{b}\boldsymbol{\omega} - \dot{\boldsymbol{\vartheta}}_\mathrm{d} \tag{16}$$

in which ${}^\mathrm{b}\boldsymbol{\omega}(t) \in \mathbb{R}^3$ is an angular velocity of $\mathcal{Q}_\mathrm{e}$ in $\{B\}$, and $L_\omega(\boldsymbol{\vartheta}) \in \mathbb{R}^{3\times3}$ is a Jacobian-like matrix given by

$$L_\omega(\boldsymbol{\vartheta}) := I - \frac{\phi}{2}[\boldsymbol{\mu}]_\times + \left(1 - \frac{\mathrm{sinc}(\phi)}{\mathrm{sinc}^2(\phi/2)}\right)[\boldsymbol{\mu}]_\times^2$$

with $\mathrm{sinc}(\phi) := \sin(\phi)/\phi$ and $\mathrm{sinc}(0) = 1$.

To control the translation of the feature point $\mathcal{Q}_\mathrm{e}$, define current and desired extended translation vectors

$$\begin{cases} \boldsymbol{p}_\mathrm{e}(t) := [u(t), v(t), -\ln(\alpha(t))]^T, \\ \boldsymbol{p}_\mathrm{ed}(t) := [u_\mathrm{d}(t), v_\mathrm{d}(t), -\ln(\alpha_\mathrm{d}(t))]^T \end{cases} \tag{17}$$

respectively. Then, define an extended translation error

$$\boldsymbol{e}_\mathrm{v}(t) := \boldsymbol{p}_\mathrm{e}(t) - \boldsymbol{p}_\mathrm{ed}(t). \tag{18}$$

Differentiating $z(\boldsymbol{q})$ in (7) with respect to time t yields

$$\dot{z}(\boldsymbol{q}) = \boldsymbol{d}_3^T{}^\mathrm{b}\boldsymbol{v}, \tag{19}$$

where ${}^\mathrm{b}\boldsymbol{v} := {}^\mathrm{b}\dot{\boldsymbol{x}}(\boldsymbol{q}) \in \mathbb{R}^4$ is a linear velocity of $\mathcal{Q}_\mathrm{e}$ in $\{B\}$. From (6), (7) and (19), the time derivative of $\boldsymbol{e}_\mathrm{v}$ becomes

$$\dot{\boldsymbol{e}}_\mathrm{v} = A_\mathrm{e}(\boldsymbol{p}){}^\mathrm{b}\boldsymbol{v}/z(\boldsymbol{q}) - \dot{\boldsymbol{p}}_\mathrm{ed} \tag{20}$$

in which $A_\mathrm{e}(\boldsymbol{p}) := KD - \boldsymbol{p}_0\boldsymbol{d}_3^T \in \mathbb{R}^{3\times4}$ is an extended interaction matrix with $\boldsymbol{p}_0(t) := [u(t), v(t), 0]^T \in \mathbb{R}^3$.

Based on the robot forward kinematics [16], the velocities ${}^\mathrm{b}\boldsymbol{v}$ and ${}^\mathrm{b}\boldsymbol{\omega}$ are calculated by

$${}^\mathrm{b}\boldsymbol{v} = J_\mathrm{v}(\boldsymbol{q})\dot{\boldsymbol{q}}, \ {}^\mathrm{b}\boldsymbol{\omega} = J_\omega(\boldsymbol{q})\dot{\boldsymbol{q}} \tag{21}$$

where $J_\mathrm{v}(\boldsymbol{q}) \in \mathbb{R}^{4\times n}$ and $J_\omega(\boldsymbol{q}) \in \mathbb{R}^{3\times n}$ are the translational and rotational Jacobian matrices, respectively, which can be obtained from the robot's forward kinematics. Inspired by [5], multiplying each side of (20) by $z(\boldsymbol{q})$ and substituting (19) and (21) into the obtained result and (20), one obtains an overall kinematic system for 6-DoF pose control as follows:

$$\begin{bmatrix} z(\boldsymbol{q})\dot{\boldsymbol{e}}_\mathrm{v} + \frac{1}{2}\dot{z}(\boldsymbol{q})\boldsymbol{e}_\mathrm{v} \\ \dot{\boldsymbol{e}}_\omega \end{bmatrix} = J_\mathrm{p}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{e}_\mathrm{v}, \boldsymbol{\vartheta})\dot{\boldsymbol{q}} - \begin{bmatrix} z(\boldsymbol{q})\dot{\boldsymbol{p}}_\mathrm{ed} \\ \dot{\boldsymbol{\vartheta}}_\mathrm{d} \end{bmatrix} \tag{22}$$

with $J_\mathrm{p}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{e}_\mathrm{v}, \boldsymbol{\vartheta}) := [J_\mathrm{pv}^T(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{e}_\mathrm{v}), J_\mathrm{p\omega}^T(\boldsymbol{q}, \boldsymbol{\vartheta})]^T \in \mathbb{R}^{6\times n}$ being a Jacobian matrix that maps the joint velocity $\dot{\boldsymbol{q}}$ to the velocity pairs $(z(\boldsymbol{q})\dot{\boldsymbol{p}}_\mathrm{e} + \frac{1}{2}\dot{z}(\boldsymbol{q})\boldsymbol{e}_\mathrm{v}, \dot{\boldsymbol{\vartheta}})$. Here, $J_\mathrm{pv}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{e}_\mathrm{v}) := Q(\boldsymbol{p}, \boldsymbol{e}_\mathrm{v})J_\mathrm{v}(\boldsymbol{q}) \in \mathbb{R}^{3\times n}$ with $Q(\boldsymbol{p}, \boldsymbol{e}_\mathrm{v}) := A_\mathrm{e}(\boldsymbol{p}) + \frac{1}{2}\boldsymbol{e}_\mathrm{v}\boldsymbol{d}_3^T \in \mathbb{R}^{3\times4}$ is the translational mapping component, and $J_\mathrm{p\omega}(\boldsymbol{q}, \boldsymbol{\vartheta}) := L_\omega(\boldsymbol{\vartheta}){}^\mathrm{c}R_\mathrm{b}J_\omega(\boldsymbol{q}) \in \mathbb{R}^{3\times n}$ is the rotational mapping component. To simplify notation, define two auxiliary variables $\boldsymbol{u}_\mathrm{v} := J_\mathrm{pv}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{e}_\mathrm{v})\dot{\boldsymbol{q}} - z(\boldsymbol{q})\dot{\boldsymbol{p}}_\mathrm{ed} \in \mathbb{R}^3$ and $\boldsymbol{u}_\omega := J_\mathrm{p\omega}(\boldsymbol{q}, \boldsymbol{\vartheta})\dot{\boldsymbol{q}} - \dot{\boldsymbol{\vartheta}}_\mathrm{d} \in \mathbb{R}^3$. Then, it follows from [19] that (22) is passive concerning input-output pairs $(\boldsymbol{u}_\mathrm{v}, \boldsymbol{e}_\mathrm{v})$ and $(\boldsymbol{u}_\omega, \boldsymbol{e}_\omega)$ for the translational and rotational components, respectively, and the corresponding storage functions are $V_\mathrm{v} = z(\boldsymbol{q})\boldsymbol{e}_\mathrm{v}^T\boldsymbol{e}_\mathrm{v}/2$ and $V_\omega = \boldsymbol{e}_\omega^T\boldsymbol{e}_\omega/2$, respectively. With the translational and rotational errors $\boldsymbol{e}_\omega$ and $\boldsymbol{e}_\mathrm{v}$, the control objective (12) becomes

$$\boldsymbol{e}_\omega \to \boldsymbol{0}, \boldsymbol{e}_\mathrm{v} \to \boldsymbol{0} \ \text{ as } \ t \to \infty. \tag{23}$$

B. Homography-Base Visual Servoing

Let $(\boldsymbol{p}_\mathrm{ed}, \dot{\boldsymbol{p}}_\mathrm{ed}, \ddot{\boldsymbol{p}}_\mathrm{ed})$ and $(\boldsymbol{\vartheta}_\mathrm{d}, \dot{\boldsymbol{\vartheta}}_\mathrm{d}, \ddot{\boldsymbol{\vartheta}}_\mathrm{d})$ be desired extended translational and rotational trajectories of the feature point $\mathcal{Q}_\mathrm{ed}$, respectively. Then, introduce nominal reference translational and rotational trajectories as follows:

$$\dot{\boldsymbol{p}}_\mathrm{er} := \dot{\boldsymbol{p}}_\mathrm{ed} - \lambda_\mathrm{v}\boldsymbol{e}_\mathrm{v}, \ \dot{\boldsymbol{\vartheta}}_\mathrm{r} := \dot{\boldsymbol{\vartheta}}_\mathrm{d} - \lambda_\omega\boldsymbol{e}_\omega \tag{24}$$

respectively, where $\dot{\boldsymbol{p}}_\mathrm{er}$, $\dot{\boldsymbol{\vartheta}}_\mathrm{r} \in \mathbb{R}^3$ denote the reference translation and rotation velocities, respectively, and $\lambda_\mathrm{v}, \lambda_\omega \in \mathbb{R}^+$ are certain constants. Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{11}$ be an estimate of $\boldsymbol{\theta}$. Substituting $\hat{\boldsymbol{\theta}}$ into $J_\mathrm{p}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\vartheta}, \boldsymbol{e}_\mathrm{v})$, $Q(\boldsymbol{p}, \boldsymbol{e}_\mathrm{v})$, ${}^\mathrm{c}R_\mathrm{b}$, and $\boldsymbol{d}_3$, one gets

$$\hat{J}_\mathrm{p} := J_\mathrm{p}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \ \hat{Q} := Q|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \ {}^\mathrm{c}\hat{R}_\mathrm{b} := {}^\mathrm{c}R_\mathrm{b}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \ \hat{z} := z|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

with $\hat{J}_\mathrm{p}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\vartheta}, \boldsymbol{e}_\mathrm{v}) \in \mathbb{R}^{6\times n}$, $\hat{Q}(\boldsymbol{p}, \boldsymbol{e}_\mathrm{v}) \in \mathbb{R}^{3\times4}$, ${}^\mathrm{c}\hat{R}_\mathrm{b} \in \mathbb{R}^{3\times3}$ and $\hat{z} \in \mathbb{R}$. Then, define estimation errors

$$\tilde{J}_\mathrm{p} := \hat{J}_\mathrm{p} - J_\mathrm{p}, \ \tilde{Q} := \hat{Q} - Q, \ {}^\mathrm{c}\tilde{R}_\mathrm{b} := {}^\mathrm{c}\hat{R}_\mathrm{b} - {}^\mathrm{c}R_\mathrm{b}, \ \tilde{z} := \hat{z} - z.$$

Design an HBVS-based joint velocity control law

$$\begin{aligned} \dot{\boldsymbol{q}}_\mathrm{c} =& \hat{J}_\mathrm{p}^+(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\vartheta}, \boldsymbol{e}_\mathrm{v}) \begin{bmatrix} \hat{z}\dot{\boldsymbol{p}}_\mathrm{er} \\ \dot{\boldsymbol{\vartheta}}_\mathrm{r} \end{bmatrix} - J_\mathrm{v}^T(\boldsymbol{q})\hat{Q}^T(\boldsymbol{p}, \boldsymbol{e}_\mathrm{v})K_\mathrm{p}\boldsymbol{e}_\mathrm{v} \\ & - J_\omega^T(\boldsymbol{q}){}^\mathrm{c}\hat{R}_\mathrm{b}^T L_\omega^T(\boldsymbol{\vartheta})K_\omega\boldsymbol{e}_\omega \end{aligned} \tag{25}$$

where $K_{\rm p}, K_\omega \in \mathbb{R}^{3\times3}$ are positive-definite diagonal matrices of control gains, and $\hat{J}_{\rm p}^{+}(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\vartheta},\boldsymbol{e}_{\rm v}) \in \mathbb{R}^{n\times6}$ is the pseudo inverse of $\hat{J}_{\rm p}(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\vartheta},\boldsymbol{e}_{\rm v})$. It is observed from (25) that the estimated matrices $\hat{J}_{\rm p}^{+}$, $\hat{Q}$ and ${}^{\rm c}\hat{R}_{\rm b}$ determine how the errors $\boldsymbol{e}_{\rm v}$, $\boldsymbol{e}_\omega$, and velocities $\dot{\boldsymbol{p}}_{\rm er}$ and $\dot{\boldsymbol{\vartheta}}_{\rm r}$ are projected into the joint space, which affects the performance of visual servoing. Then, define a velocity tracking error in the joint space

$$\dot{\boldsymbol{e}}(t) := \dot{\boldsymbol{q}}(t) - \dot{\boldsymbol{q}}_{\rm c}(t) \tag{26}$$

and two auxiliary matrices $\Phi_{{\rm y}i} := \Phi(\boldsymbol{p}_0 - \boldsymbol{e}_v/2, \boldsymbol{J}_{{\rm v}i}) \in \mathbb{R}^{11\times3}$ and $\Phi_{{\rm r}i} := B(\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=[\boldsymbol{J}_{\omega i}^T,0]^T} \in \mathbb{R}^{11\times3}$, where $\boldsymbol{J}_{{\rm v}i} \in \mathbb{R}^4$ and $\boldsymbol{J}_{\omega i} \in \mathbb{R}^3$ are the ith columns of $J_{\rm v}(\boldsymbol{q})$ and $J_\omega(\boldsymbol{q})$, respectively. Then, one has two linearly parameterized models.

Property 2: If d_{34} is known, then for any given $\boldsymbol{\eta}_1$, one gets a linearly parameterized model

$$J_{\rm v}^T(\boldsymbol{q})\hat{Q}^T(\boldsymbol{p},\boldsymbol{e}_{\rm v})\boldsymbol{\eta}_1 + J_\omega^T(\boldsymbol{q}){}^{\rm c}\hat{R}_{\rm b}^T\boldsymbol{\eta}_2 = Y_1^T(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\hat{\boldsymbol{\theta}} \tag{27}$$

with $Y_1(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) := [\Phi_{\rm r1}\boldsymbol{\eta}_2 - \Phi_{\rm y1}\boldsymbol{\eta}_1, \Phi_{\rm r2}\boldsymbol{\eta}_2 - \Phi_{\rm y2}\boldsymbol{\eta}_1, \ldots, \Phi_{{\rm r}n}\boldsymbol{\eta}_2 - \Phi_{{\rm y}n}\boldsymbol{\eta}_1] \in \mathbb{R}^{11\times n}$.

Property 3: If d_{34} is known, then for any given $\boldsymbol{\eta} \in \mathbb{R}^3$, one gets a linearly parameterized model

$$\hat{z}\boldsymbol{\eta} = Y_2^T(\boldsymbol{q},\boldsymbol{\eta})\hat{\boldsymbol{\theta}} + \boldsymbol{y}_2(\boldsymbol{\eta}) \tag{28}$$

with $\boldsymbol{y}_2(\boldsymbol{\eta}) := d_{34}\boldsymbol{\eta} \in \mathbb{R}^3$, $Y_2(\boldsymbol{q},\boldsymbol{\eta}) := B({}^{\rm b}\boldsymbol{x}(\boldsymbol{q}))\boldsymbol{k}\boldsymbol{\eta}^T \in \mathbb{R}^{11\times3}$, and $B(\cdot)$ defined in (10).

Replace the tuple $(\hat{z}, \hat{\boldsymbol{\theta}}, \boldsymbol{\eta})$ in (28) by $(z(\boldsymbol{q}), \boldsymbol{\theta}, \dot{\boldsymbol{p}}_{\rm er})$ yields

$$(\hat{z} - z(\boldsymbol{q}))\dot{\boldsymbol{p}}_{\rm er} = Y_2^T(\boldsymbol{q},\dot{\boldsymbol{p}}_{\rm er})\tilde{\boldsymbol{\theta}} \tag{29}$$

where $\tilde{\boldsymbol{\theta}}(t) := \hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta} \in \mathbb{R}^{11}$ is a parameter estimation error. By using (27), (25) can be rewritten into

$$\dot{\boldsymbol{q}}_{\rm c} = \hat{J}_{\rm p}^{+}(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\vartheta},\boldsymbol{e}_{\rm v})\begin{bmatrix}\hat{z}\dot{\boldsymbol{p}}_{\rm er}\\ \dot{\boldsymbol{\vartheta}}_{\rm r}\end{bmatrix} - Y_1^T(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\hat{\boldsymbol{\theta}} \tag{30}$$

with $\boldsymbol{\eta}_1 = K_{\rm p}\boldsymbol{e}_{\rm v}$ and $\boldsymbol{\eta}_2 = L_\omega^T K_\omega \boldsymbol{e}_\omega$. Applying (24), (26) and (30) to (22), yields the closed-loop overall kinematics

$$\begin{bmatrix}z\dot{\boldsymbol{e}}_{\rm v} + \frac{1}{2}\dot{z}\boldsymbol{e}_{\rm v}\\ \dot{\boldsymbol{e}}_\omega\end{bmatrix} = \hat{J}_{\rm p}\dot{\boldsymbol{e}} - \hat{J}_{\rm p}Y_1^T\hat{\boldsymbol{\theta}} - \tilde{J}_{\rm p}\dot{\boldsymbol{e}} + \tilde{J}_{\rm p}Y_1^T\hat{\boldsymbol{\theta}} - \tilde{J}_{\rm p}\hat{J}_{\rm p}^{+}\begin{bmatrix}\hat{z}\dot{\boldsymbol{p}}_{\rm er}\\ \dot{\boldsymbol{\vartheta}}_{\rm r}\end{bmatrix} + \begin{bmatrix}\tilde{z}\dot{\boldsymbol{p}}_{\rm er}\\ \boldsymbol{0}\end{bmatrix} - \begin{bmatrix}\lambda_{\rm v} z\boldsymbol{e}_{\rm v}\\ \lambda_\omega\boldsymbol{e}_\omega\end{bmatrix}. \tag{31}$$

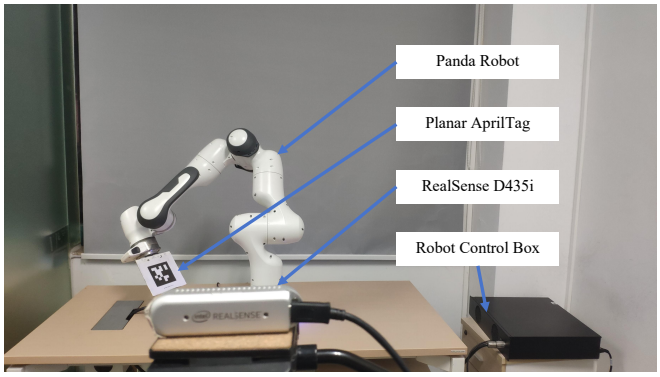

Fig. 2. An experimental environment for monocular ETH robots, where the end-effector is identified by an attached planar AprilTag.

C. Composite Learning Adaptation

We define IE and PE to facilitate analysis as follows [10].

Definition 1: A bounded signal $\Phi(t) \in \mathbb{R}^{11\times3}$ is of IE if $\exists$ $T_{\rm e}, \zeta_{\rm d}, \sigma \in \mathbb{R}^+$ such that $\int_{T_{\rm e}-\zeta_{\rm d}}^{T_{\rm e}} \Phi(\zeta)\Phi^T(\zeta){\rm d}\zeta \geq \sigma I$.

Definition 2: A bounded signal $\Phi(t) \in \mathbb{R}^{11\times3}$ is of PE if $\exists$ $\zeta_{\rm d}, \sigma \in \mathbb{R}^+$ such that $\int_{t-\zeta_{\rm d}}^{t} \Phi(\zeta)\Phi^T(\zeta){\rm d}\zeta \geq \sigma I, \forall t \geq 0$.

Define a generalized prediction error

$$\boldsymbol{\xi}(t) := \Psi_{\rm e}(t)\hat{\boldsymbol{\theta}} - \Psi_{\rm e}(t)\boldsymbol{\theta} \in \mathbb{R}^{11} \tag{32}$$

in which $\Psi_{\rm e} \in \mathbb{R}^{11\times11}$ is given by

$$\Psi_{\rm e}(t) := \begin{cases}\int_{t-\zeta_{\rm d}}^{t}\Phi(\zeta)\Phi^T(\zeta){\rm d}\zeta, & t < T_{\rm e}\\ \int_{T_e-\zeta_{\rm d}}^{T_e}\Phi(\zeta)\Phi^T(\zeta){\rm d}\zeta, & t \geq T_{\rm e}\end{cases}.$$

Note that one has $\Psi_{\rm e}\boldsymbol{\theta} = \int_{t-\zeta_{\rm d}}^{t}\Phi(\zeta)\boldsymbol{y}(\zeta){\rm d}\zeta$. Using eigendecomposition, one obtains $\Psi_{\rm e} = U\Sigma U^T$ with $\Sigma := {\rm diag}(\lambda_1, \lambda_2, \ldots, \lambda_{11})$, where $\lambda_i \in \mathbb{R}$ ($i = 1$ to 11) is the ith eigenvalue of Ψ, and $U \in \mathbb{R}^{11\times11}$ is orthogonal, i.e., $UU^T = I$. Note that $\Psi_{\rm e}$ is symmetric and positive-semidefinite, and each eigenvalue λ_i is non-negative. Define a regularized inversion of $\Psi_{\rm e}$:

$$\Psi_\varrho^{+} := U\Sigma_\varrho U^T \in \mathbb{R}^{11\times11} \tag{33}$$

with $\Sigma_\varrho := {\rm diag}(\lambda_{\varrho1}, \lambda_{\varrho2}, \ldots, \lambda_{\varrho11}) \in \mathbb{R}^{11\times11}$, where $\lambda_{\varrho i} := 1/\max\{\lambda_i, \varrho\} \in \mathbb{R}^+$ ($i = 1$ to 11) is the ith eigenvalue of Ψ_ϱ^{+}, and $\varrho \in \mathbb{R}^+$ is a small threshold. Then, one concludes that: 1) If $\Psi_{\rm e} \in \mathbb{R}^{11\times11}$ is positive-semidefinite, Ψ_ϱ^{+} in (33) is positive-definite; 2) $\boldsymbol{0} \leq \Psi_\varrho^{+}\Psi_{\rm e} \leq I$, $\Psi_\varrho^{+}\Psi_{\rm e} = I$ iff $\Psi_{\rm e} \geq \varrho I$, and $\Psi_\varrho^{+}\Psi_{\rm e} = \boldsymbol{0}$ iff $\Psi_{\rm e} = \boldsymbol{0}$ [14]. Now, design a regularized composite learning law as follows:

$$\dot{\hat{\boldsymbol{\theta}}} = \Gamma\left(Y_1(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\dot{\boldsymbol{q}} - Y_2(\boldsymbol{q},\boldsymbol{\eta})K_{\rm p}\boldsymbol{e}_{\rm v} - \kappa\Psi_\varrho^{+}\boldsymbol{\xi}\right) \tag{34}$$

with $\boldsymbol{\eta} = \dot{\boldsymbol{p}}_{\rm er}$, where $\Gamma \in \mathbb{R}^{11\times11}$ is a positive-definite matrix of learning rates, and $\kappa \in \mathbb{R}^+$ is a weighting factor. We give the following reasonable assumption from [5].

Assumption 1: The low-level joint controller of the robotic system guarantees $\dot{\boldsymbol{e}} \in L_2 \cap L_\infty$ so that there exists a constant $l_{\rm m} \in \mathbb{R}^+$ to satisfy $\int_0^t \dot{\boldsymbol{e}}^T(\tau)\dot{\boldsymbol{e}}(\tau){\rm d}\tau \leq l_{\rm m}, \forall t \geq 0$.

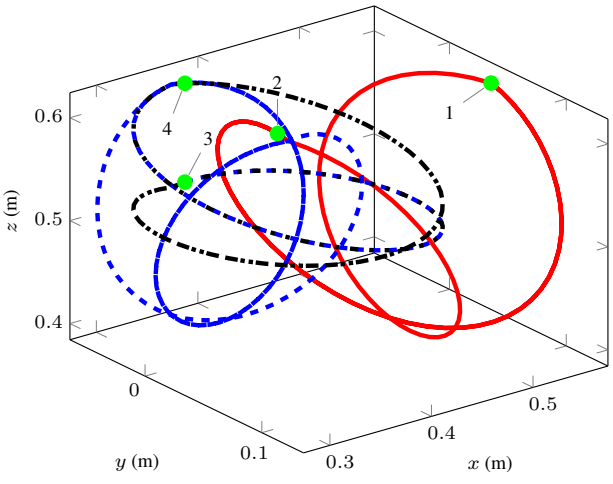

Fig. 3. The desired trajectory of the feature point $\mathcal{Q}_{\rm e}$ in the Euclidean space, where the notation "1" to "4" (marked in green dot) denote the waypoints at instants $t = 0$, 30, 60 and 100 s, respectively.

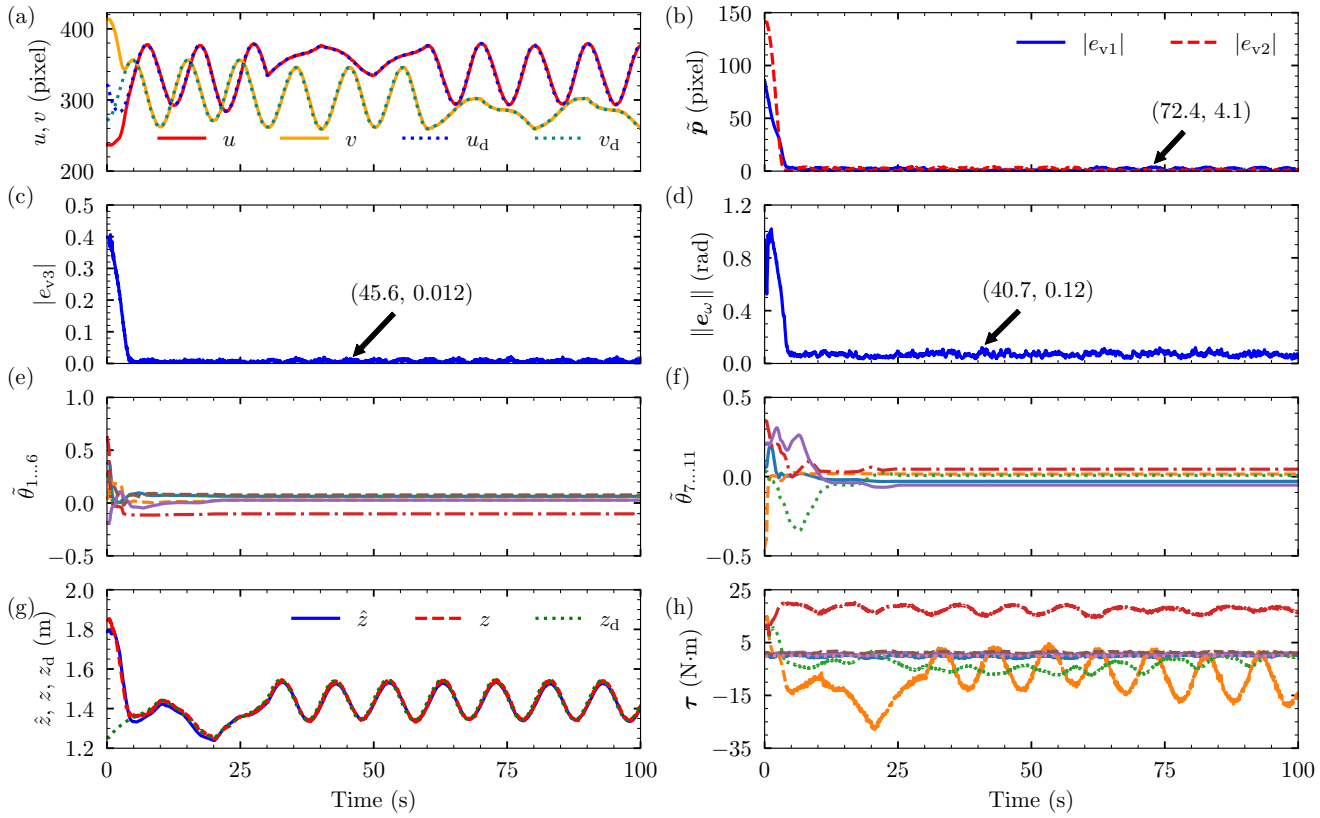

Fig. 4. Control results of the proposed CL-HBVS in experiments. (a) Pixel tracking trajectories in the u- and v-axes of the image plane. (b) Pixel errors $e_{\mathrm{v}1}$ and $e_{\mathrm{v}2}$ in the u- and v-axes of the image plane, respectively. (c) Absolute value of the depth ratio error $e_{\mathrm{v}3}$. (d) Norm of the rotation error $\boldsymbol{e}_\omega$. (e) Parameter errors $\tilde{\theta}_i$ (i =1 to 6). (f) Parameter errors $\tilde{\theta}_i$ (i = 7 to 11). (g) Estimated, current, and desired depths $\hat{z}$, z, and z_{d}. (h) Control torque $\boldsymbol{\tau} \in \mathbb{R}^7$.

Consider an articulated robot system with an ETH monocular camera, as depicted in Fig. 1, where a plane $\aleph$ is attached to the end-effector frame $\{E\}$, containing at least 4 feature points. Only the feature point $\mathcal{Q}_{\mathrm{e}}$ in the world frame $\{B\}$ is known, which is applied to calculate its Cartesian coordinate ${}^{\mathrm{b}}\boldsymbol{x}(\boldsymbol{q})$. If the kinematic system described by (22) is driven by the CL-HBVS control law (25) with (34), then the closed-loop system is stable in the following sense: 1) All signals involved are bounded and the tracking errors $\boldsymbol{e}_{\mathrm{v}}(t)$ and $\boldsymbol{e}_\omega(t)$ asymptotically converge to $\mathbf{0}$ on $t \geq 0$; 2) if IE exists, both the tracking errors $\boldsymbol{e}_{\mathrm{v}}(t)$, $\boldsymbol{e}_\omega(t)$ and the parameter estimation error $\tilde{\boldsymbol{\theta}}(t)$ asymptotically converge to $\mathbf{0}$ on $t \geq T_{\mathrm{e}}$, implying exact depth estimation as $\hat{z}(t) \to z(\boldsymbol{q}(t))$. The proof is similar to [5], so it is omitted here. Note that even if d_{34} is unknown a priori, (34) still works as discussed in [14].

IV. Experimental Studies

Experiments on the vision-based robotic system are carried out as illustrated in Fig. 2, where the environment includes a 7-DoF collaborative robot named Franka Emika Panda, and a fixed Intel RealSense camera D435i to capture images at a resolution of 640×480 pixels, which provides visual signals at a rate of 30 frames per second. An AprilTag [20] with four visual markers positioned at its corners is affixed to the end-effector. During experiments, pixel positions $\boldsymbol{p}_{\mathrm{r}i}$ ($i = 1$ to 4) of each reference feature points $\mathcal{Q}_{\mathrm{r}i}$ are (189, 241), (268, 241), (266, 163) and (185, 163), and the initial robot joint position is $\boldsymbol{q}(0) = [0, -0.5236, -1.8326, -2.4435, -0.2618, 2.0071, -1.5708]^T$ rad. The camera's intrinsic parameter matrix is

$$K = \begin{bmatrix} 608 & 0 & 327 \\ 0 & 608 & 232 \\ 0 & 0 & 1 \end{bmatrix}.$$

Since the camera's extrinsic parameter D is unavailable in practice, the following calibrated one from ViSP [21]:

$$\hat{D}_{\mathrm{r}} = \begin{bmatrix} -0.1184 & 0.9902 & -0.0735 & 0.1382 \\ 0.0427 & -0.0689 & -0.9967 & 0.6554 \\ -0.9920 & -0.1212 & -0.0341 & 1.8341 \end{bmatrix}$$

is chosen as a reference. Also, as d_{34} is unknown, a reference one $\hat{d}_{34\mathrm{r}} = 1.8341$ obtained from $\hat{D}_{\mathrm{r}}$ is used for calculating $\boldsymbol{y}(\boldsymbol{p}_i, {}^{\mathrm{b}}\boldsymbol{x}_i)$ in (11) and $\boldsymbol{y}_2(\boldsymbol{\eta})$ in (28).

An initial estimate of D is set as $\hat{\boldsymbol{\theta}}(0) = [0.2263, 0.9567, 0.1830, 1.0000, -0.1830, 0.2263, -0.9567, 0.1000, -0.9567, 0.1830, 0.2263]^T$, which ensures a suitable initial discrepancy with D, better showing the evolution of the parameter estimate $\hat{\boldsymbol{\theta}}$. We design a desired trajectory with 100 s shown in Fig. 3, which consists of three sub-tasks: 1) Tracking the red trajectory at $t \in [0, 30)$ s; 2) tracking the blue trajectory at $t \in$ [30,60) s; 3) tracking the black trajectory at $t \in$ [60,100] s.

The proposed CL-HBVS (25) with (34) is implemented with $K_{\mathrm{p}} = \mathrm{diag}(4 \times 10^{-5}, 4 \times 10^{-5}, 0.2)$, $K_\omega = 0.2I$, $\Gamma =$

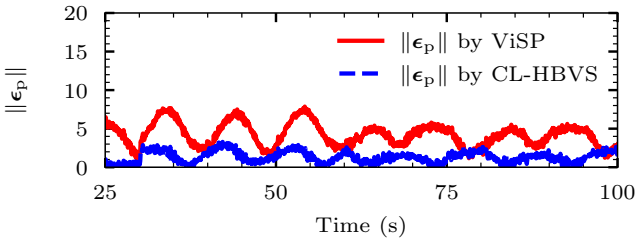

Fig. 5. Prediction errors ϵ_p by ViSP and the proposed CL-HBVS.

TABLE I
ESTIMATED, CURRENT AND DESIRED DEPTH VALUES $\hat{z}$, z, AND z_d DURING EXPERIMENTS

Time (s)	31	42	53	64	75	86	95	100
$\hat{z}$ (m)	1.48	1.51	1.53	1.51	1.47	1.40	1.46	1.40
z (m)	1.49	1.52	1.54	1.52	1.48	1.41	1.47	1.42
z_d (m)	1.50	1.53	1.54	1.51	1.46	1.40	1.45	1.43
$\frac{\|\tilde{\theta}\|}{\|\theta\|}$ (%)	4.88	4.87	4.88	4.88	4.88	4.88	4.88	4.88

$0.4I$, $\lambda_v = \lambda_\omega = 2$, $\kappa = 15$, $\varrho = 3$, and $T_e = \tau_d = 25$. Experiments are run on a host PC under Linux OS Ubuntu 20.04 with ROS Noetic Ninjemys. The ViSP and the Franka Control Interface provided via libfranka are integrated into the ROS. To evaluate the effectiveness of depth tracking and estimation by the proposed method, we use the pose estimation algorithm embedded in ViSP to get exact information, where z is calculated by the known geometric information of the AprilTag, and $\hat{z}$ is obtained by (7). Finally, define a model prediction error $\epsilon_p(t) := \boldsymbol{y}(\boldsymbol{p}, {}^b\boldsymbol{x}) - \Phi^T(\boldsymbol{p}, {}^b\boldsymbol{x})\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is obtained by (34) ($\kappa = 15$) and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_r$ for the proposed CL-HBVS and ViSP, respectively, and $\hat{\boldsymbol{\theta}}_r \in \mathbb{R}^{11}$ is the first 11 elements of the calibrated matrix $\hat{D}_r$.

Experiment results are demonstrated in Fig. 4. One observes that the current pixel $\boldsymbol{p} = [u, v, 1]^T$ of the feature point $\mathcal{Q}_e$ accurately tracks its desired position $\boldsymbol{p}_d = [u_d, v_d, 1]^T$ with a steady-state error of about 0.4 pixel [see Figs. 4(a) and (b)], the current depth z converges to the desired depth z_d with an error of about 0.02 m [see Fig. 4 (c)], and the norm of the rotation error $\|\boldsymbol{e}_\omega\|$ is about 0.1 rad on average [see Fig. 4(d)]. Besides, the convergence of $\tilde{\boldsymbol{\theta}}$ with an error of 0.1 is achieved at about 25 s [see Fig. 4 (e)-(f)], which indicates an accurate depth estimation [see Fig. 4 (g)]. Table I gives the depth values $\hat{z}$, z and z_d at some instants, where both the depth estimation error $\tilde{z}$ and the depth tracking error $e_z := z - z_d \in \mathbb{R}$ are less than 2 cm. In addition, the control torque $\boldsymbol{\tau} \in \mathbb{R}^7$ [see Fig. 4 (h)] is shown to verify that the control input of the low-level joint controller is reasonable. What's more, the prediction error ϵ_p obtained by the proposed CL-HBVS is smaller than that by ViSP [see Fig. 5], indicating that $\hat{\boldsymbol{\theta}}$ by the CL-HBVS is closer to the true value $\boldsymbol{\theta}$ compared to the calibrated one $\hat{\boldsymbol{\theta}}_r$. To assess the performance of $\tilde{\boldsymbol{\theta}}$, the percentage estimation error $\|\tilde{\boldsymbol{\theta}}\|/\|\boldsymbol{\theta}_r\|$ (%) with less than 5 % is also shown in Table I.

V. CONCLUSIONS

This paper has presented a passivity-based kinematic control method named CL-HBVS for robots with ETH monocular cameras to achieve 6-DoF pose tracking in the 3D space under unknown camera extrinsic parameters. Parameter convergence is achieved under the weakened IE condition, which leads to the exact estimation of the time-varying depth and prevents singularity in the estimated rotation matrix. Experiments based on a 7-DoF robot manipulator have validated the effectiveness of the proposed method in estimating the extrinsic parameters and time-varying depth and controlling 6-DoF robot pose.

REFERENCES

[1] F. Chaumette and S. Hutchinson, "Visual servo control. I. basic approaches," *IEEE Robot. Autom. Mag.*, vol. 13, no. 4, pp. 82–90, Dec. 2006.

[2] J. Chen, D. M. Dawson, W. E. Dixon, and A. Behal, "Adaptive homography-based visual servo tracking for a fixed camera configuration with a camera-in-hand extension," *IEEE Trans. Control Syst. Technol.*, vol. 13, no. 5, pp. 814–825, Sep. 2005.

[3] C.-C. Cheah, C. Liu, and J.-J. E. Slotine, "Adaptive tracking control for robots with unknown kinematic and dynamic properties," *Int. J. Robot. Res.*, vol. 25, no. 3, pp. 283–296, Mar. 2006.

[4] H. Wang, C. C. Cheah, W. Ren, and Y. Xie, "Passive separation approach to adaptive visual tracking for robotic systems," *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 6, pp. 2232–2241, Apr. 2018.

[5] Y. Li, H. Wang, Y. Xie, C. C. Cheah, and W. Ren, "Adaptive image-space regulation for robotic systems," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 2, pp. 850–857, Dec. 2021.

[6] Y. Zhang and C. Hua, "A new adaptive visual tracking scheme for robotic system without image-space velocity information," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 52, no. 8, pp. 5249–5258, Aug. 2022.

[7] G. Hu, S. Gupta, N. Fitz-Coy, and W. E. Dixon, "Lyapunov-based visual servo tracking control via a quatemion formulation," in *Proc. IEEE Conf. Decis. Control*, San Diego, CA, USA,, 2006, pp. 3861–3866.

[8] J. Chen, V. K. Chitrakaran, and D. M. Dawson, "Range identification of features on an object using a single camera," *Automatica*, vol. 47, no. 1, pp. 201–206, Jan. 2011.

[9] S. Mehta, V. Jayaraman, T. Burks, and W. Dixon, "Teach by zooming: A unified approach to visual servo control," *Mechatronics*, vol. 22, no. 4, pp. 436–443, Jun. 2012.

[10] Y. Pan and H. Yu, "Composite learning robot control with guaranteed parameter convergence," *Automatica*, vol. 89, pp. 398–406, Mar. 2018.

[11] B. Lai, Z. Li, W. Li, C. Yang, and Y. Pan, "Homography-based visual servoing of eye-in-hand robots with exact depth estimation," *IEEE Trans. Ind. Electron.*, vol. 71, no. 4, pp. 3832–3841, Apr. 2024.

[12] Z. Li, B. Lai, and Y. Pan, "Image-based composite learning robot visual servoing with an uncalibrated eye-to-hand camera," *IEEE/ASME Trans. Mechatronics*, vol. 29, no. 4, pp. 2499–2509, Aug. 2024.

[13] Z. Li, W. Li, and Y. Pan, "Composite learning image-based visual servoing of redundant robots with nullspace compliance," *IEEE Control Syst. Lett.*, vol. 8, pp. 315–320, Jan. 2024.

[14] Z. Li, B. Lai, H. Wang, and Y. Pan, "Homography-based visual servoing of robot pose under an uncalibrated eye-to-hand camera," *IEEE/ASME Trans. Mechatronics*, vol. 29, no. 3, pp. 1891–1902, Jun. 2024.

[15] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NJ, USA: Cambridge Univ. Press, 2004.

[16] J. J. Craig, *Introduction to Robotics: Mechanics and Control*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2005.

[17] X. Liang, H. Wang, Y.-H. Liu, W. Chen, and J. Zhao, "A unified design method for adaptive visual tracking control of robots with eye-in-hand /fixed camera configuration," *Automatica*, vol. 59, pp. 97–105, Sep. 2015.

[18] Y.-H. Liu, H. Wang, C. Wang, and K. K. Lam, "Uncalibrated visual servoing of robots using a depth-independent interaction matrix," *IEEE Trans. Robot.*, vol. 22, no. 4, pp. 804–817, Aug. 2006.

[19] H. Wang, "Passivity-based adaptive control for visually servoed robotic systems," in *Proc. Austral. Control Conf.*, Canberra, Australia, 2014, pp. 152–157.

[20] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Daejeon, South Korea, 2016, pp. 4193–4198.

[21] E. Marchand, F. Spindler, and F. Chaumette, "ViSP for visual servoing: A generic software platform with a wide class of robot control skills," *IEEE Robot. Autom. Mag.*, vol. 12, no. 4, pp. 40–52, Dec. 2005.