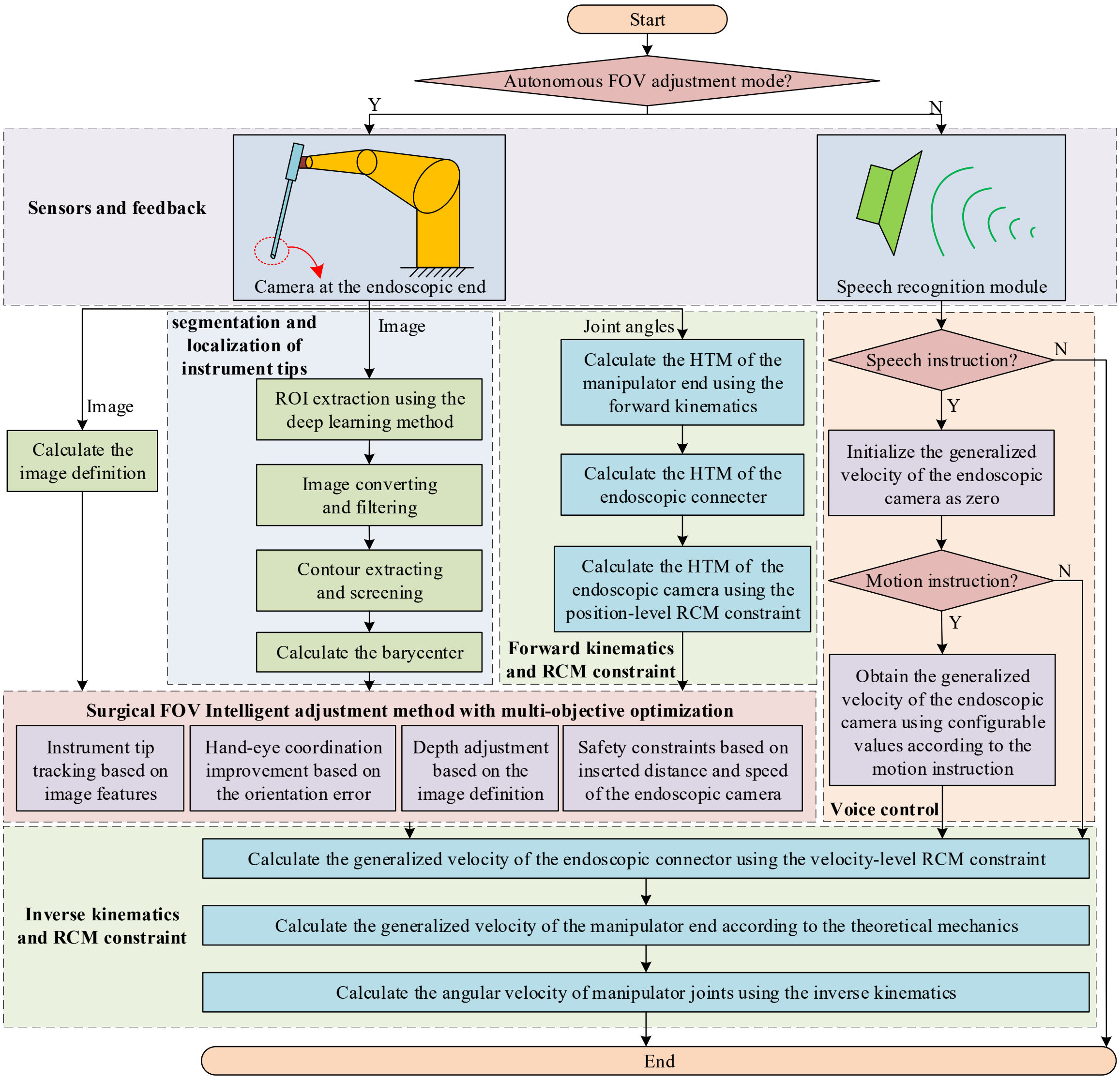
APPENDIX A

In order to better reflect the comprehensibility of the AEHR system and the control process, the calculation process of the entire algorithm is supplemented, as shown in **APPENDIX A.1**.


Start
Autonomous FOV adjustment mode?
Y
N
Sensors and feedback
Camera at the endoscopic end
Speech recognition module
Image
segmentation and localization of instrument tips
Image
ROI extraction using the deep learning method
Image converting and filtering
Contour extracting and screening
Calculate the barycenter
Calculate the image definition
Joint angles
Calculate the HTM of the manipulator end using the forward kinematics
Calculate the HTM of the endoscopic connecter
Calculate the HTM of the endoscopic camera using the position-level RCM constraint
Forward kinematics and RCM constraint
Speech instruction?
N
Y
Initialize the generalized velocity of the endoscopic camera as zero
Motion instruction?
N
Y
Obtain the generalized velocity of the endoscopic camera using configurable values according to the motion instruction
Voice control
Surgical FOV Intelligent adjustment method with multi-objective optimization
Instrument tip tracking based on image features
Hand-eye coordination improvement based on the orientation error
Depth adjustment based on the image definition
Safety constraints based on inserted distance and speed of the endoscopic camera
Inverse kinematics and RCM constraint
Calculate the generalized velocity of the endoscopic connector using the velocity-level RCM constraint
Calculate the generalized velocity of the manipulator end according to the theoretical mechanics
Calculate the angular velocity of manipulator joints using the inverse kinematics
End

APPENDIX A.1 The algorithm flowchart of the AEHR system

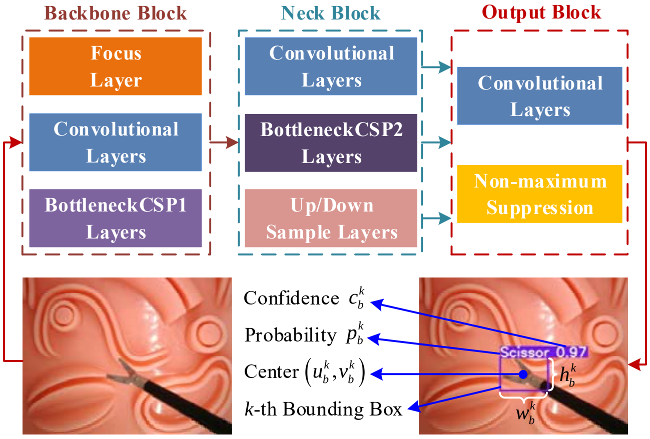
APPENDIX B

The localization of surgical instrument tips can provide effective and important feedback information to intelligent FOV adjustment of AEHRs. In order to accurately extract the system tracking point, this paper adopts the DLM for segmentation and localization of instrument tips, which can reduce the localization error of instrument tips. Firstly, the region of interest (ROI) (i.e., the instrument tip region) is extracted in visual feedback using the YOLOv5 model. Then, the instrument tips are segmented, and the barycenter of instrument tips is extracted. Finally, the weighted average barycenter of all instrument tips is calculated as the system tracking point.

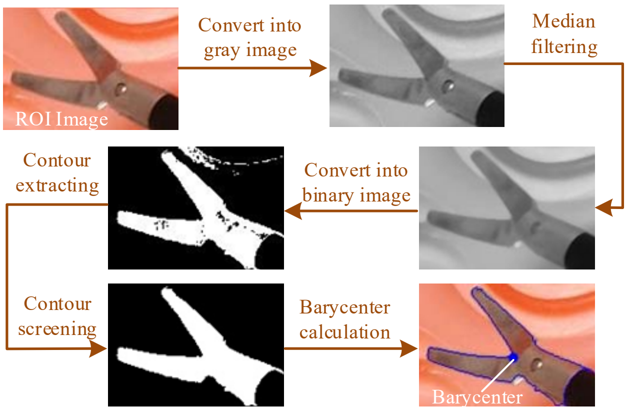
ROI extraction: The YOLOv5 model has the ability of fast and accurate target region regression prediction, which can meet the real-time requirement of ROI extraction. It is assumed that I_S is the $w_s \times h_s$ surgical image captured by the endoscopic camera, which contains the projection of K instruments. The schematic diagram of ROI extraction is shown in **APPENDIX B.1**, and the prediction information (i.e., center, size, confidence, and class) of the kth instrument tip region is expressed as $\left(u_b^k, v_b^k, w_b^k, h_b^k, c_b^k, p_b^k\right)$.

Segmentation and localization of instrument tips: The segmentation and localization process of instrument tips is shown in **APPENDIX B.2**. It mainly includes the following steps: (1) crop out the kth ROI image I_{TS}^k in I_S according to $\left(u_b^k, v_b^k, w_b^k, h_b^k\right)$; (2) convert I_{TS}^k into the gray image I_{TG}^k; (3) process I_{TG}^k using median filter according to the filter size k_s

and obtain the filtered image I_{TGF}^{k}; (4) convert I_{TGF}^{k} into the binary image I_{TB}^{k} according to the threshold $\left[tr_{\min}, tr_{\max}\right]$; (5) extract Q contours in I_{TB}^{k}; (6) calculate the area a^q of q-th (q = 1,2,…,Q) contour $\boldsymbol{c}^q$ and the area ratio $\eta_{\text{area}} = a^q / \left(w_b^k \times h_b^k\right)$ of $\boldsymbol{c}^q$ in I_{TS}^{k}; (7) calculate the barycenter $\boldsymbol{s}_{\text{tip}}^{k}$ of kth instrument tip according to $\boldsymbol{c}^q$.



APPENDIX B.1 The schematic diagram of ROI extraction



APPENDIX B.2 The segmentation and localization process of instrument tips

The whole algorithm of the surgical instrument tips segmentation and localization is shown in **Algorithm B**.

Algorithm B Segmentation and localization of surgical instrument tips

```
Input: $I_S$ - source image captured by the endoscopic
Output: $\boldsymbol{s}_{\text{tips}}$ - system tracking point
    $\left\{\left(u_b^k, v_b^k, w_b^k, h_b^k, c_b^k, p_b^k\right) \mid k = 1,2,\ldots,K\right\} \leftarrow$ input $I_S$ into the YOLOv5
model to predict ROIs
    for $k$ =1 to $K$ do
        $I_{\mathrm{TS}}^{k} \leftarrow$ crop $I_S$ using $\left(u_b^k, v_b^k, w_b^k, h_b^k\right)$
        $I_{\mathrm{TG}}^{k} \leftarrow$ convert $I_{\mathrm{TS}}^{k}$ into a gray image
        $I_{\mathrm{TGF}}^{k} \leftarrow I_{\mathrm{TG}}^{k}$ using median filter, of which the size is $k_{\text{size}}$
        $I_{\mathrm{TB}}^{k} \leftarrow$ convert $I_{\mathrm{TGF}}^{k}$ into a binary image using $\left(tr_{\min}, tr_{\max}\right)$
        $\left\{\boldsymbol{c}^q \mid q = 1,2,\ldots,Q\right\} \leftarrow$ extract contours of $I_{\mathrm{TB}}^{k}$
        for $q$ = 1 to $Q$ do
            $a^q \leftarrow$ calculate the area of $\boldsymbol{c}^q$
            calculate the area ratio of $\boldsymbol{c}^q$ in $I_{\mathrm{TS}}^{k}$ : $\eta_{\text{area}} = a^q / \left(w_b^k \times h_b^k\right)$
            if $\eta_{\text{area}} > ra_{\min}$ and $\eta_{\text{area}} < ra_{\max}$
                $I_{\mathrm{TBS}}^{k} \leftarrow$ choose the region of $\boldsymbol{c}^q$ from $I_{\mathrm{TB}}^{k}$
            end if
        end for
        $\boldsymbol{s}_{\text{tip}}^{k} \leftarrow$ (12) ~ (13) with $I_{\mathrm{TBS}}^{k}$
    end for
    $\boldsymbol{s}_{\text{tips}} \leftarrow$ (14) with $\left\{\boldsymbol{s}_{\text{tip}}^{k} \mid k = 1,2,\ldots,K\right\}$
return $\boldsymbol{s}_{\text{tips}}$
```

APPENDIX C

Proof of Theorem

On the one hand, the proposed method decouples and controls the position and attitude of the endoscopic camera at the same time. On the other hand, it adjusts the endoscopic depth to optimize the image definition under the premise of meeting the position-level safety constraint.

Theorem C1:

According to the principle of pinhole imaging, it can be obtained:

$$\begin{bmatrix} u_{\text{tips}}^{i} \\ v_{\text{tips}}^{i} \\ 1 \end{bmatrix} = \frac{1}{{}^{\text{lap}}z_i} \begin{bmatrix} f_u & 0 & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} {}^{\text{lap}}x_i \\ {}^{\text{lap}}y_i \\ {}^{\text{lap}}z_i \end{bmatrix} \tag{C.1}$$

As shown in Fig.5, the direction vector of the endoscopic camera velocity is determined by:

$$\boldsymbol{v}_r^i = \frac{\boldsymbol{z}_{\text{lap}} \times \hat{\boldsymbol{p}}_{\text{tips}}^{i} \times \hat{\boldsymbol{p}}_{\text{tips}}^{i}}{\left\| \boldsymbol{z}_{\text{lap}} \times \hat{\boldsymbol{p}}_{\text{tips},i}^{i} \times \hat{\boldsymbol{p}}_{\text{tips}}^{i} \right\|} \tag{C.2}$$

where, $\boldsymbol{z}_{\text{lap}} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^{\mathrm{T}}$, $\hat{\boldsymbol{p}}_{\text{tips}}^{i} = \begin{bmatrix} x_{\text{tips}}^{i} / z_{\text{tips}}^{i} & y_{\text{tips},i}^{i} / z_{\text{tips}}^{i} & 1 \end{bmatrix}^{\mathrm{T}}$, $\boldsymbol{p}_{\text{tips}}^{i} = [\, x_{\text{tips}}^{i}, y_{\text{tips}}^{i}, z_{\text{tips}}^{i} \,]^{\mathrm{T}}$ is the 3D coordinate of the ith device tracking point.

Therefore, the linear velocity of the endoscopic camera can be expressed as:

$$\boldsymbol{v}_c^i = \begin{bmatrix} v_{cx}^{i} \\ v_{cy}^{i} \\ v_{cz}^{i} \end{bmatrix} = \upsilon_s^i \boldsymbol{v}_r^i = \upsilon_s^i \begin{bmatrix} x_{\text{tips}}^{i} z_{\text{tips}}^{i} \big/ l \\ y_{\text{tips},i}^{i} z_{\text{tips}}^{i} \big/ l \\ \left(x_{\text{tips}}^{i,2} + y_{\text{tips}}^{i,2} \right) \big/ l \end{bmatrix} \tag{C.3}$$

where, $l = \sqrt{\left(x_{\text{tips}}^{i\ 2} + y_{\text{tips}}^{i\ 2} \right)} \left\| \boldsymbol{p}_{\text{tips}}^{i} \right\|$, $\upsilon^i \geq 0$ is the magnitude of the camera velocity, $\boldsymbol{v}_r^i = \left[v_{rx}^{i}, v_{ry}^{i}, v_{rz}^{i} \right]^{\mathrm{T}}$.

When the endoscopic camera moves and the surgical instrument tip is stationary, the velocity of the surgical instrument tip relative to the endoscopic camera is:

$$\boldsymbol{v}_f^i = \begin{bmatrix} -v_{rx}^{i} - z_{\text{tips}}^{i} \omega_{ry}^{i} \\ -v_{ry}^{i} + z_{\text{tips}}^{i} \omega_{rx}^{i} \\ -v_{rz}^{i} - y_{\text{tips}}^{i} \omega_{rx}^{i} + x_{\text{tips}}^{i} \omega_{ry}^{i} \end{bmatrix} \tag{C.4}$$

where, $\omega_{rx}^{i} = -v_{ry}^{i} / d_{in}$, $\omega_{ry}^{i} = v_{rx}^{i} / d_{in}$.

When both the endoscopic camera and the surgical instrument tips are moving, the linear velocity of the surgical instrument tips relative to the endoscopic camera is:

$$\hat{\boldsymbol{v}}_f^i = \begin{bmatrix} \dot{x}_{\text{tips}}^i - v_{rx}^i - \dfrac{v_{rx}^i z_{\text{tips}}^i}{d_{in}} \\ \dot{y}_{\text{tips}}^i - v_{ry}^i - \dfrac{v_{ry}^i z_{\text{tips}}^i}{d_{in}} \\ \dot{z}_{\text{tips}}^i - v_{rz}^i + \dfrac{v_{ry}^i y_{\text{tips}}^i}{d_{in}} + \dfrac{v_{rx}^i x_{\text{tips}}^i}{d_{in}} \end{bmatrix} = \begin{bmatrix} \dot{x}_{\text{tips}}^i - \dfrac{\upsilon^i}{l} x_{\text{tips},i}^i z_{\text{tips},i}^i - \dfrac{\upsilon^i}{d_{in} l} x_{\text{tips},i}^i z_{\text{tips}}^{i2} \\ \dot{y}_{\text{tips}}^i - \dfrac{\upsilon^i}{l} y_{\text{tips}}^i z_{\text{tips}}^i - \dfrac{\upsilon^i}{d_{in} l} y_{\text{tips}}^i z_{\text{tips}}^{i,2} \\ \dot{z}_{\text{tips}}^i + \dfrac{\upsilon^i}{l}\left(x_{\text{tips}}^{i,2} + y_{\text{tips}}^{i,2}\right) + \dfrac{\upsilon^i}{d_{in} l}\left(x_{\text{tips}}^{i,2} + y_{\text{tips}}^{i,2}\right) z_{\text{tips}}^i \end{bmatrix} \tag{C.5}$$

Differentiate (C.1) and substitute (C.5) into the result of the derivation to get:

$$\begin{aligned} \dot{u}_{\text{tips}}^i &= \frac{f_u}{z_{\text{tips}}^i} v_{fx}^i - \frac{f_u\, x_{\text{tips}}^i}{z_{\text{tips}}^{i2}} v_{fz}^i \\ &= \frac{f_u}{z_i^{i2}} \Big[\dot{x}_{\text{tips}}^i z_{\text{tips}}^i - x_{\text{tips}}^i \dot{z}_{\text{tips}}^i \\ &\qquad - \frac{\upsilon^i}{l} x_{\text{tips}}^i \left(x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2} + z_{\text{tips}}^{i2} \right) \\ &\qquad - \frac{\upsilon^i}{d_{in} l} x_{\text{tips}}^i z_{\text{tips}}^i \left(x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2} + z_{\text{tips}}^{i2} \right) \Big] \end{aligned} \tag{C.6}$$

$$\begin{aligned} \dot{v}_{\text{tips}}^i &= \frac{f_v}{z_{\text{tips}}^i} v_{fy}^i - \frac{f_v\, y_{\text{tips}}^i}{z_i^2} v_{fz}^i \\ &= \frac{f_v}{z_{\text{tips}}^{i2}} \Big[\dot{y}_{\text{tips}}^i z_{\text{tips}}^i - y_{\text{tips}}^i \dot{z}_{\text{tips}}^i \\ &\qquad - \frac{\upsilon^i}{l} y_{\text{tips}}^i \left(x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2} + z_{\text{tips}}^{i2} \right) \\ &\qquad - \frac{\upsilon^i}{d_{in} l} y_{\text{tips}}^i z_{\text{tips}}^i \left(x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2} + z_{\text{tips}}^{i2} \right) \Big] \end{aligned} \tag{C.7}$$

The established Lyapunov function is the distance between the pixel coordinates of the actual instrument tips and the center of the camera imaging image, that is,

$$L = \left(u_{\text{tips}}^i - u_0\right)^2 + \left(v_{\text{tips}}^i - v_0\right)^2 \tag{C.8}$$

Taking the derivative of (C.8), we can get:

$$\dot{L} = 2\left(u_{\text{tips}}^i - u_0\right)\dot{u}_{\text{tips}}^i + 2\left(u_{\text{tips}}^i - v_0\right)\dot{v}_{\text{tips}}^i \tag{C.9}$$

Substituting (C.1), (C.6), and (C.7) into (C.9), yield:

$$\begin{aligned} \dot{L} = {} & \frac{f_u^2 x_{\text{tips}}^{i2}}{z_{\text{tips}}^{i3}} \left[\frac{z_{\text{tips}}^i}{x_{\text{tips}}^i} \dot{x}_{\text{tips}}^i - \dot{z}_{\text{tips}}^i - \left(1 + \frac{z_{\text{tips}}^i}{d_{in}}\right) \upsilon^i \frac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \right] \\ & + \frac{f_v^2 y_{\text{tips}}^{i2}}{z_{\text{tips}}^{i3}} \left[\frac{z_{\text{tips}}^i}{y_{\text{tips}}^i} \dot{y}_{\text{tips}}^i - \dot{z}_{\text{tips}}^i - \left(1 + \frac{z_{\text{tips}}^i}{d_{in}}\right) \upsilon^i \frac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \right] \end{aligned} \tag{C.10}$$

When $d_{in} \to +\infty$, (10) can be further simplified as:

$$\begin{aligned} \dot{L} = {} & \frac{f_u^2 x_{\text{tips}}^{i2}}{z_{\text{tips}}^{i3}} \left[\frac{z_{\text{tips}}^i}{x_{\text{tips}}^i} \dot{x}_{\text{tips}}^i - \dot{z}_{\text{tips}}^i - \frac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \upsilon^i \right] + \\ & \frac{f_v^2 y_{\text{tips}}^{i2}}{z_{\text{tips}}^{i3}} \left[\frac{z_{\text{tips}}^i}{y_{\text{tips}}^i} \dot{y}_{\text{tips}}^i - \dot{z}_{\text{tips}}^i - \frac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \upsilon^i \right] \\ \le {} & \frac{f_u^2 x_{\text{tips}}^{i2}}{z_{\text{tips}}^{i3}} A + \frac{f_v^2 y_{\text{tips}}^{i2}}{z_{\text{tips}}^{i3}} B \end{aligned} \tag{C.11}$$

with

$$\begin{aligned} A &= \frac{z_{\text{tips}}^i}{x_{\text{tips}}^i} \dot{x}_{\text{tips}}^i - \dot{z}_{\text{tips}}^i - \frac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \upsilon^i \\ B &= \frac{z_{\text{tips}}^i}{y_{\text{tips}}^i} \dot{y}_{\text{tips}}^i - \dot{z}_{\text{tips}}^i - \frac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \upsilon^i \end{aligned} \tag{C.12}$$

Since $z_{\text{tips}}^i > 0$, (12) has the following characteristics:

$$\begin{aligned} A &\le \max\left\{ \frac{z_{\text{tips}}^i}{\left\| x_{\text{tips}}^i \right\|} \upsilon^i{}_s, \upsilon^i \right\} - \frac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \upsilon^i \\ B &\le \max\left\{ \frac{z_{\text{tips}}^i}{\left\| y_{\text{tips}}^i \right\|} \upsilon^i, \upsilon^i \right\} - \frac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \upsilon^i \end{aligned} \tag{C.13}$$

Further,

$$\begin{aligned} & A: \begin{cases} A \le \upsilon^i - \dfrac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \upsilon^i \le 0, & \text{if } \dfrac{z_{\text{tips}}^i}{\left\| x_{\text{tips}}^i \right\|} \le 1 \\ A \le \dfrac{z_{\text{tips}}^i}{\left\| x_{\text{tips}}^i \right\|} \upsilon^i - \sqrt{\dfrac{x_{\text{tips}}^{i2} + z_{\text{tips}}^{i2}}{x_{\text{tips}}^{i2}}} \upsilon^i \le 0, & \text{else} \end{cases} \\ & B: \begin{cases} B \le \upsilon^i - \dfrac{\left\| \boldsymbol{p}_{\text{tips}}^i \right\|}{\sqrt{x_{\text{tips}}^{i2} + y_{\text{tips}}^{i2}}} \upsilon^i \le 0, & \text{if } \dfrac{z_{\text{tips}}^i}{\left\| y_{\text{tips}}^i \right\|} \le 1 \\ B \le \dfrac{z_{\text{tips}}^i}{\left\| y_{\text{tips}}^i \right\|} \upsilon^i - \sqrt{\dfrac{y_{\text{tips}}^{i2} + z_{\text{tips}}^{i2}}{y_{\text{tips}}^{i2}}} \upsilon^i \le 0, & \text{else} \end{cases} \end{aligned} \tag{C.14}$$

Based on (C.14) and (C.11), $\dot{L} \le 0$ can be obtained, so the distance error between the pixel coordinates of the actual instrument end and the center of the camera image is always convergent.

Theorem C2:

According to the hand-eye coordination model, $\exists \beta_1 \ge 0$, when $\forall \beta \ge \beta_1$, $0 \le {}^{\text{con}}\omega_{\text{coor}} \le \omega_{\max}$.

$$\dot{\beta} = -\eta_{\text{coor}}\, {}^{\text{con}}\omega_{\text{coor}} \le 0 \tag{C.15}$$

Empathy, $\exists \beta_2 \le 0$, when $\forall \beta \le \beta_2$, $-\omega_{\max} \le {}^{\text{con}}\omega_{\text{coor}}^t \le 0$.

$$\dot{\beta} = -\eta_{\text{coor}}\, {}^{\text{con}}\omega_{\text{coor}} \ge 0 \tag{C.16}$$

Theorem C3:

Let $\eta_{\mathrm{def}} \ge \eta_{\mathrm{track}}$. According to the optimization model considering image definition, $\exists\left|\eta_q\right| \ge \eta_0$, when $\forall d_{obj} < d_{foc}$, $\lambda_{\mathrm{def}} = v_{\max}$.

$$\begin{aligned}\dot{d}_{\mathrm{obj}} &= \eta_{\mathrm{track}}\,{}^{lap}\boldsymbol{v}_{\mathrm{track}}\cdot\left[0\quad 0\quad 1\right]^{\mathrm{T}} + \eta_{\mathrm{def}}\lambda_{\mathrm{def}}\\ &\ge\left(-\eta_{\mathrm{track}}+\eta_{\mathrm{def}}\right)v_{\max}\\ &\ge 0\end{aligned} \tag{C.17}$$

Empathy, $\exists\left|\eta_q\right| \ge \eta_0$, when $\forall d_{\mathrm{obj}} > d_{\mathrm{foc}}$, $\lambda_{\mathrm{def}} = -v_{\max}$.

$$\begin{aligned}\dot{d}_{\mathrm{obj}} &= \eta_{\mathrm{track}}\,{}^{\mathrm{lap}}\boldsymbol{v}_{\mathrm{track}}\cdot\left[0\quad 0\quad 1\right]^{\mathrm{T}} + \eta_{\mathrm{def}}\lambda_{\mathrm{def}}\\ &\le\left(\eta_{\mathrm{track}}-\eta_{\mathrm{def}}\right)v_{\max}\\ &\le 0\end{aligned} \tag{C.18}$$

Theorem C4:

Let $\eta_{\mathrm{safe}} \ge \eta_{\mathrm{track}}$. According to the position-level safety constraint, $\exists d_{in1} \in \left[d_{\mathrm{alart,min}}, d_{\mathrm{safe,min}}\right]$, when $\forall d_{in} \in \left[d_{\mathrm{alart,min}}, d_{in1}\right]$, $\lambda_{safe} = v_{\max}$.

$$\begin{aligned}\dot{d}_{in} &= \eta_{\mathrm{track}}\,{}^{\mathrm{con}}\boldsymbol{v}_{\mathrm{track}}\cdot\left[0\quad 0\quad 1\right]^{\mathrm{T}} + \eta_{\mathrm{safe}}\lambda_{\mathrm{safe}}\\ &\ge\left(-\eta_{\mathrm{track}}+\eta_{\mathrm{safe}}\right)v_{\max}\\ &\ge 0\end{aligned} \tag{C.19}$$

Empathy, $\exists d_{in2} \in \left[d_{\mathrm{safe,max}}, d_{\mathrm{alart,max}}\right]$, when $\forall d_{in} \in \left[d_{in2}, d_{\mathrm{alart,max}}\right]$, $\lambda_{safe} = -v_{\max}$.

$$\begin{aligned}\dot{d}_{in} &= \eta_{\mathrm{track}}\,{}^{\mathrm{con}}\boldsymbol{v}_{\mathrm{track}}\cdot\left[0\quad 0\quad 1\right]^{\mathrm{T}} + \eta_{\mathrm{safe}}\lambda_{\mathrm{safe}}\\ &\le\left(\eta_{\mathrm{track}}-\eta_{\mathrm{safe}}\right)v_{\max}\\ &\le 0\end{aligned} \tag{C.20}$$

APPENDIX D

In the experimental system, the D-H parameters of the 6-DOF robot are shown in **APPENDIX D.1**. During instrument tips identification, the architecture of surgical instrument detection with YOLOv5 model is shown in **APPENDIX D.2.**

APPENDIX D.1 The D-H parameter of the robot

Link i	α_i (rad)	a_i (mm)	d_i (mm)	offset(rad)	θ_i (rad)
1	$\pi/2$	0	199.34	$-\pi$	θ_1
2	0	-250.00	0	0	θ_2
3	0	-250.00	0	0	θ_3
4	$\pi/2$	0	109.10	0	θ_4
5	$-\pi/2$	0	108.00	0	θ_5
6	0	0	75.86	0	θ_6

APPENDIX D.2 Architecture of surgical instrument detection with YOLOv5 model

Structure name	Filter size	Input shape	Output shape
Adaptive resize	-	640×480×3	608×608×3
Focus	32	608×608×3	304×304×32
BatchNormalization+ Convolution	64	304×304×32	152×152×64
BottleneckCSP	64	152×152×64	152×152×64
BatchNormalization+ Convolution	128	152×152×64	76×76×128
BottleneckCSP×3	128	76×76×128	76×76×128
BatchNormalization+ Convolution	256	76×76×128	38×38×256
BottleneckCSP×3	256	38×38×256	38×38×256
BatchNormalization+ Convolution	512	38×38×256	19×19×512
Neck	-	76×76×128 38×38×256 19×19×512	76×76×255 38×38×255 19×19×255