



Research article

Soft objects grasping evaluation using a novel VCFN-YOLOv8 framework

Guoshun Cui^{a,b}, Shiwei Su^{a,b}, Hanyu Gao^{a,b}, Kai Zhuo^a, Kun Yang^{a,c,*}, Hang Wu^{d,e,**}^a Shanxi Key Laboratory of Artificial Intelligence & Micro Nano Sensors, College of Integrated Circuits, Taiyuan University of Technology, Taiyuan 030024, China^b Shanxi Institute of 6D Artificial Intelligence Biomedical Science, Taiyuan 030031, China^c Jinan Shengquan Group Share-Holding Co., Ltd., Jinan 250000, China^d Medical Support Technology Research Department, Academy of Military Sciences, Tianjin 300000, China^e School of Artificial Intelligence, Nankai University, Tianjin 300350, China

ARTICLE INFO

Article history:

Received 7 December 2024

Revised 24 February 2025

Accepted 19 March 2025

Available online 11 April 2025

Keywords:

Grasping evaluation

Multimodal fusion

Intelligent perception

YOLOv8

ABSTRACT

Humans can quickly perform adaptive grasping of soft objects by using visual perception and judgment of the grasping angle, which helps prevent the objects from sliding or deforming excessively. However, this easy task remains a challenge for robots. The grasping states of soft objects can be categorized into four types: sliding, appropriate, excessive and extreme. Effective recognition of different states is crucial for achieving adaptive grasping of soft objects. To address this problem, a novel visual-curvature fusion network based on YOLOv8 (VCFN-YOLOv8) is proposed to evaluate the grasping state of various soft objects. In this framework, the robotic arm equipped with the wrist camera and the curvature sensor is established to perform generalization grasping and lifting experiments on 11 different objects. Meanwhile, the dataset is built for training and testing the proposed method. The results show a classification accuracy of 99.51% on four different grasping states. A series of grasping evaluation experiments is conducted based on the proposed framework, along with tests for the model's generality. The experiment results demonstrate that VCFN-YOLOv8 is accurate and efficient in evaluating the grasping state of soft objects and shows a certain degree of generalization for non-soft objects. It can be widely applied in fields such as automatic control, adaptive grasping and surgical robot.

© 2025 The Author(s). Published by Elsevier B.V. on behalf of Shandong University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Due to rapid technological advancements, the demand for robotic control and dexterous hand operations in service and production industries is continually increasing. The grasping ability of robotic dexterous hands has attracted widespread attention [1]. Efficient and accurate evaluation of the grasping state is a crucial part of improving the overall grasping quality of robotic dexterous hands.

Visual sensing is the main way humans perceive the surrounding world, with over 70% of the information humans receive from the external environment coming from visual data. Computer vision mimics human visual perception to obtain information and makes decisions to produce corresponding results [2]. However,

relying solely on visual perception has certain limitations. For example, it may be difficult to capture all information about an object due to the camera angle or the lighting conditions. Therefore, multimodal information fusion becomes a key approach to solving these issues. Traditional evaluations of robotic arm dexterous hand grasping quality mainly focus on object slipping and deformation caused by excessive grasping, which lead to visual and force feedback. Many researchers have studied grasp stability evaluation, slip detection and deformation detection [3,4]. However, the framework they used often involves decision-making based on visual and tactile fusion. This approach works well for objects with noticeable elastic force changes after deformation. However, for soft objects, due to their material properties, the elastic force feedback changes during deformation are not obvious [5], making the grasping state assessment using visual-tactile fusion less effective. In some cases, tactile feedback may not fully capture the shape of soft objects. Due to the nonlinear and unstructured characteristics of soft objects, they do not show a stable shape during contact. This instability makes it harder to capture clear and consistent information using conventional

* Corresponding author at: Shanxi Key Laboratory of Artificial Intelligence & Micro Nano Sensors, College of Integrated Circuits, Taiyuan University of Technology, Taiyuan 030024, China.

** Corresponding author.

E-mail addresses: yangkun01@tyut.edu.cn (K. Yang), wuhang1991@nankai.edu.cn (H. Wu).

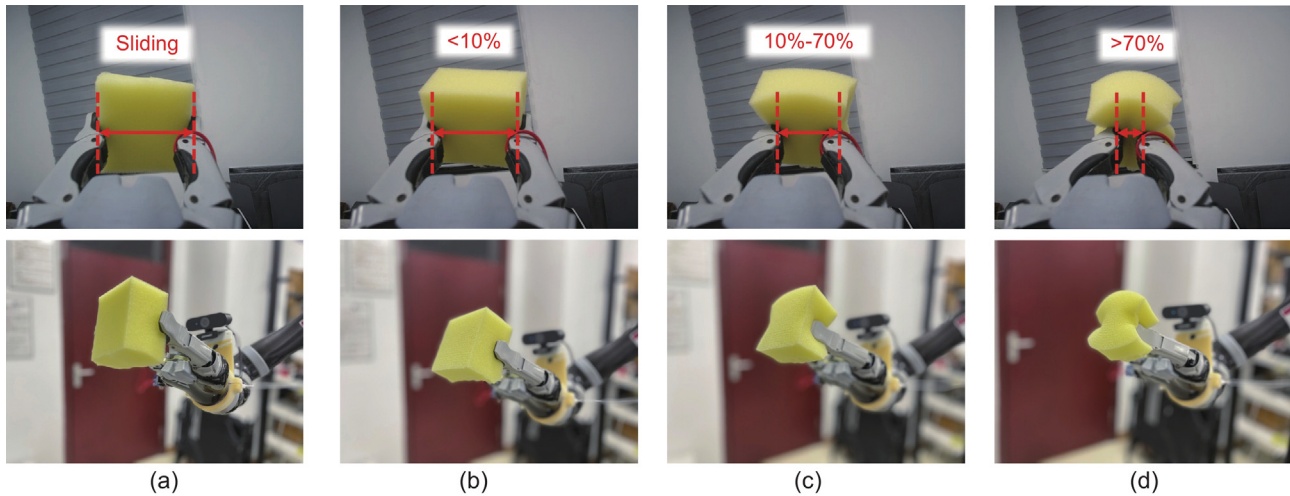


Fig. 1. Different grasping states. (a): Sliding. (b): Appropriate. (c): Excessive. (d): Extreme. The images in the top row are captured by a camera mounted on the wrist of the robotic arm, while the bottom images are captured by a side camera. The red part indicates the compression ratio.

tactile feedback methods. Therefore, a more reliable method is needed to assess the grasping state of soft objects.

In this paper, the task of evaluating the grasping state of soft objects is defined as a four-class problem illustrated in Fig. 1. These four types of problem include sliding (there is relative sliding between the object and the gripper during the grasping), appropriate (the compression rate of the object does not exceed 10%), excessive (the compression rate of the object does not exceed 70%) and extreme (the compression rate of the object exceeds 70%).

In the task of evaluating the grasping state of soft objects, humans can intuitively carry out the evaluation process, thanks to feedback from both visual input and finger grasping angles [6]. Similarly, robots can use this method to complete the evaluation task. The focus of this study is to enable robots to use visual-curvature information fusion perception to evaluate the grasping state of soft objects. The primary challenge in solving this dynamic classification problem under the dual-modal information framework is to establish a mechanism for cross-modal feature extraction and attribute association between visual and curvature data [7]. Inspired by a bionic approach, a novel Visual-Curvature Fusion Network based on YOLOv8 (VCFN-YOLOv8) is proposed to assess the grasping state of various soft objects. Additionally, a camera is used to capture extensive images from different angles to gather rich visual information and a wide range of grasping and lifting experiments are conducted to train and test the proposed model. The visual and curvature information is captured separately from a rotatable camera located on the robotic arm's wrist and a curvature sensor fixed at the back of the robot gripper. The experimental setup is shown in Fig. 2. Finally, A series of comparative experiments based on different sequence lengths, structures and modes are conducted on the proposed VCFN-YOLOv8 framework. The proposed method is also carried out a series of grasping evaluation experiments and generalization tests, further verifying its effectiveness and generality.

The structure of this paper is as follows: Section 2 describes the related work on detection and state evaluation of grasping tasks. Section 3 describes the preprocessing of the sensors' information and the detailed architecture of VCFN-YOLOv8. Section 4 introduces the design of the experiments and analysis of the experimental results. In the conclusion, the work of this paper is summarized and the future application of the proposed framework is prospected. The main contributions of this paper are as follows:

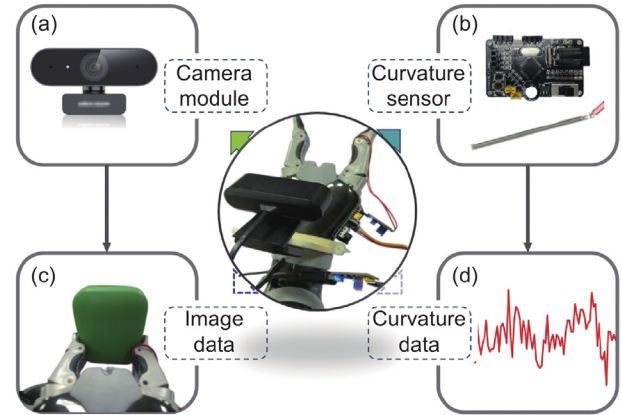


Fig. 2. The experiment setup of the paper. (a): The HIKVISION DS-E12 camera mounted on the wrist. (b): The dexterous hand gripper is equipped with a curvature sensor module. (c): Images taken by the wrist-mounted camera. (d): Curvature data from the curvature sensor. (e): The robotic arm, model JACO2-j2n6s200, produced by KINOVA.

- A grasping state evaluation system for soft objects is built. The system consists of the data acquisition module, the JACO2-j2n6s200 robotic arm and the algorithm fusion component.
- A novel multimodal fusion network (VCFN-YOLOv8) is proposed to address the problem of assessing the grasping state of soft objects. The use of the vision and the curvature can overcome the shortcomings of traditional frameworks and improves the accuracy of grasping evaluations. The results show a classification accuracy of 99.51% on four different grasping states.
- Accurate and effective grasping state evaluation experiments are conducted on unfamiliar objects, with an accuracy rate of 92%. Extended validation is conducted on the adaptive grasping of deformable objects, and the experimental setup completes the adaptive grasping within 6 s. The designed experiment verifies the generality of the architecture.

2. Related work

Different modalities such as vision, tactus and curvature are applied in the recognition of grasping. Vision is an important modality for acquiring information [8]. Popović et al. [9] proposed a grasping strategy that does not use any object-specific priori knowledge. The strategy used vision to extract second-order relations of 3D features that represent the edge structure. The grasping method was applied to humanoid robots. Liang et al. [10] introduced a 6-DoF position estimation method based on 3D vision reconstruction for the precise localization of objects during robotic grasping. The method achieved high-degree-of-freedom robot grasping in unknown environments. The method could still maintain a certain degree of robustness in complex environments. Ge et al. [11] designed a visual strategy for grasping unobstructed objects in complex scenes using the Mask-RCNN visual image segmentation network and color information. The accuracy of this framework on the real scene dataset was 92%. In actual grasping task, the method achieved a grasp success rate of 91%. Luo et al. [12] proposed a novel visual servo controller based on natural features. This controller extracted natural features from the target image and processed them to provide servo feature points. The controller was equipped with the airborne manipulator system to realize aero-grasping.

In addition to visual methods, tactile sensing has been widely used due to its advantages in contact perception. Kwiatkowski et al. [3] proposed a new method for the grasping stability assessment using CNN. The method improved upon the results of the unsupervised feature learning approach by using tactile feedback. Cockburn et al. [4] introduced a new framework based on tactile sensors for improved robotic grasping. An unsupervised feature learning method was used in the framework to predict grasp performance. Hyttinen et al. [13] proposed a framework to evaluate the grasping stability of objects. The framework evaluated grasping stability by learning tactile features. The grasping results were correctly predicted with the accuracy of 89% in the test cases.

With the increasing demand for grasp of the soft objects, curvature perception is gradually finding its way into various fields. Taghipour et al. [14] conducted a comparative study of grating-based curvature sensors between Long Period Fiber Gratings (LPFG) and Fiber Bragg Gratings (FBG), and proposed a comprehensive simulation model. Through sensor integration technology, closed-loop control of modular robotic architectures could be achieved. Giada Gerboni from Stanford University [15] developed feedback control of a soft robot curvature module based on flexible fluid actuator (FFA) using commercial curvature sensors. Zhong et al. [16] designed a curvature sensor. Based on the proposed curvature sensor, two gloves were designed and manufactured.

Different tasks might require different perceptual modalities to obtain more diverse information [17,18]. In some complex scenes, the single modality might be greatly affected by environmental interference and cannot provide enough information. Scholars began to focus on multi-modal research to obtain more comprehensive environmental and operational information. After decades of research, with the emergence of various sensors and innovations in fusion algorithms, the range of modalities became increasingly diverse [19–21]. Therefore, selecting the appropriate modality fusion for different problems in various environments became a critical issue [22,23]. Cui et al. [24, 25] proposed a 3D convolution-based (C3D-VTFN) and a self-attention mechanism-based (VTFSa) visual-tactile fusion deep neural network to evaluate the grasping state of various deformable objects. The experiments showed that the classification accuracy of C3D-VTFN model could reach 99.97% and the VTFSa

model outperformed traditional methods by a margin of 7%. Han et al. [26] proposed a transformer-based grasping framework for rigid grippers, in which visual and tactile information were used to predict grasp outcomes with a multilayer perceptron. Depierre et al. [27] extended a neural network with a scoring module to evaluate the grasping ability of a given position and introduced a novel loss function that associates grasping parameter regression with the grasping ability score. Although the grasping of rigid objects has achieved good results, the grasping effect of soft objects is not satisfactory. Therefore, this work focuses on the multimodal grasp perception of soft objects, aiming to develop a method that achieves accurate and robust grasping.

3. Methodology

3.1. Method statement

According to the descriptions in Fig. 2, the ultimate goal of the proposed framework is to determine the grasping state by fusing the visual and curvature information. Given a visual image X_v and a sequence of curvature data $(X_{c1}, X_{c2}, X_{c3}, \dots, X_{cn})$, the YOLOv8 network Y_v is first used to extract the visual feature information $F_v = Y_v(X_v)$. Then, the curvature information feature extraction module Y_c is applied to obtain the curvature feature information $F_c = Y_c(X_{c1}, X_{c2}, X_{c3}, \dots, X_{cn})$. These two sets of features are fed into the feature fusion module $F_{v,c}$ to construct fused features with attribute associations. Finally, these fused features are input into the classification function \mathcal{F}_c to predict the current grasp state g , which is formalized as:

$$F_{v,c} = Y_v(X_v) \otimes Y_c(X_{c1}, X_{c2}, X_{c3}, \dots, X_{cn}). \quad (1)$$

$$g = \mathcal{F}_c(F_{v,c}) \quad g \in 0, 1, 2, 3. \quad (2)$$

Where 0, 1, 2 and 3 represent four grasp states: sliding, appropriate, excessive and extreme, respectively. Thus, the grasp state evaluation problem is defined as a four-class classification problem. The meaning assigned to “ \otimes ” in Eq. (1) is the fusion of visual information and curvature information. In this process, the weighted probability index of the four evaluated state(\mathcal{P}) is calculated and formulated as:

$$\mathcal{P}_{g=0,1,2,3} = \frac{\omega_{visual} \cdot p_{visual} + \omega_{curvature} \cdot p_{curvature}}{\omega_{visual} + \omega_{curvature}}. \quad (3)$$

In Eq. (3), ω_{visual} and $\omega_{curvature}$ are the weights of visual modality and curvature modality. p_{visual} and $p_{curvature}$ are the predicted probabilities of visual modality and curvature modality.

3.2. Preprocessing

3.2.1. Curvature data preprocessing

Due to the presence of Gaussian noise, the stability of the raw curvature signal is poor, which will introduce significant interference to the subsequent feature fusion work. Therefore, based on the characteristics of the collected signals, in order to reduce the influence of noise and improve signal quality, the original curvature signal is subjected to the following simple and efficient filtering operations. The first filtering (Ffilter) is carried out in the front-end acquisition device processor (STM32F103C8T6) using a mean filtering method to initially smooth the raw signal. It can be formalized as:

$$y[n] = \frac{1}{5} \sum_{i=0}^4 x[n+i]. \quad (4)$$

In this equation, $x[n+i]$ is the input sequence, $y[n]$ is the output sequence of the Filter and the window size is 5.

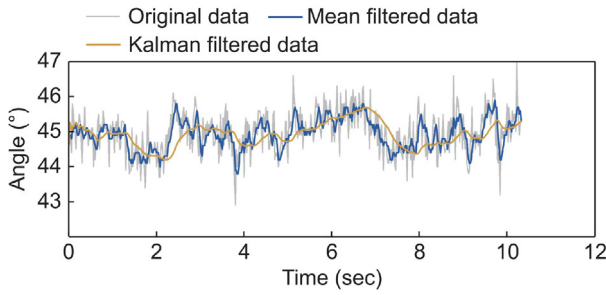


Fig. 3. The curvature data preprocessing process and results.

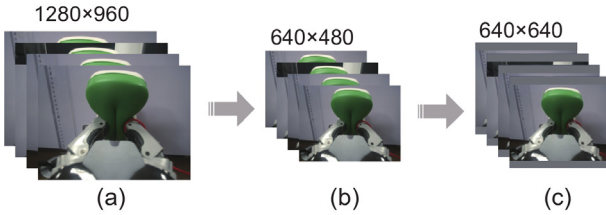


Fig. 4. The preprocessing process and results of the input image. (a) Original; (b) Downsampling; (c) Supplementation.

The preliminarily processed signal is then sent to the host computer by serial bus, where a Kalman filter is applied for the second filtering (Sfilter) to obtain the desired ideal data. The input value of Sfilter is the observation data sequence $x_n = z_n + v_n$, where x_n , z_n and v_n represent the observed values, true positions and the noise. Specifically, the observed noise follows a normal distribution $v_n \sim \mathcal{N}(0, R)$, with $R = 5$. During the prediction phase, Sfilter predicts the current state based on the previous state:

$$z_n^- = F \times z_{n-1}. \quad (5)$$

$$P_n^- = F P_{n-1} F^T + Q. \quad (6)$$

In Eq. (5), z_{n-1} and z_n^- represents the states of the previous time and the current time respectively. Eq. (6) represents the prediction of the covariance matrix, where $F = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $Q = 10^{-5}$. In the update phase, The Sfilter updates the state based on the current observation value x_n and the predicted value. At this stage, the Kalman gain K_n is calculated using the following formula:

$$K_n = P_n^- \begin{bmatrix} 1 \\ 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} P_n^- \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 5 \right)^{-1} \quad (7)$$

The relationship between input and output can be derived and is formulated as:

$$\hat{z}_n = \hat{z}_n^- + K_n(x_n - \hat{z}_n^-). \quad (8)$$

In this formula, \hat{z}_n is the Kalman filtered data, and x_n is the input observation value. In addition, the initial value of the filter is set. The initial state is $z_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and the initial covariance is

$$P_0 = \begin{bmatrix} 1000 & 0 \\ 0 & 1000 \end{bmatrix}. \text{ By applying the formulas in this section,}$$

the curvature data has been processed. The comparison of the processing results of curvature data is shown in Fig. 3. From the data curve in the figure, it can be seen that the filtered data greatly improves the noise situation without losing the main information.

3.2.2. Image data preprocessing

The image data is captured by the HIKVISION DS-E12 camera installed on the robotic arm's wrist. Each frame of the image has a resolution of 1280×960 with the size of about 170KB. The image data input size required by the model is 640×640 . To reduce the image size, the image is downsampled and compressed. The compressed image resolution is 640×480 , with a size of about 50KB, which can reduce the data volume by about 70%. In practical operation, the scaling factor s is calculated as follows:

$$s = \min\left(\frac{W_{target}}{W_{orig}}, \frac{H_{target}}{H_{orig}}\right). \quad (9)$$

In Eq. (9), W_{target} and H_{target} are the target width and height of the image. W_{orig} and H_{orig} are the original width and height of the image. In this study, s is always equal to 1. The scaled image size is $W_{scaled} \times H_{scaled}$ and it can be formulated as:

$$W_{scaled} = W_{orig} \times s. \quad (10)$$

$$H_{scaled} = H_{orig} \times s. \quad (11)$$

In the above Eqs. (10) and (11), W_{scaled} and H_{scaled} represent the scaled width and height respectively. Only the pixels in the height direction are filled and this process is expressed as:

$$P_W = \frac{W_{target} - W_{scaled}}{2}. \quad (12)$$

$$P_H = \frac{H_{target} - H_{scaled}}{2}. \quad (13)$$

Where P_W and P_H represent the filling in the width and height directions respectively. The image input size for the YOLOv8 model is set, and the images undergo the processes including: maintaining the aspect ratio, scaling the image and padding the scaled image to meet the required input size. The final image can be visually represented as:

$$FinalImage = Pad(ScaledImage, P_W, P_H). \quad (14)$$

This processing can reduce the data volume while preserving the information carried. The specific image input size and padding values can be adjusted according to specific needs. The process is illustrated in Fig. 4.

3.3. Fusion model

The overall structure of the VCFN-YOLOv8 framework is shown in Fig. 5. The input of the model consists of the visual image and the curvature signal sequence, while the output is the current grasping state of the gripper. The whole framework can be separated into preprocessing, feature extraction, feature fusion and classification. The YOLOv8 is fast and can detect objects in real time, which is important for robotic vision tasks. Compared to previous versions, YOLOv8 uses advanced detection methods and adapts to different environments, such as lighting changes and cluttered backgrounds. Specifically, the backgrounds of the evaluated subjects in this study are not singular. Therefore, the YOLOv8-based model has good speed, accuracy and adaptability, which is suitable for this study. The detail descriptions of the network are illustrated as follows:

(1) In the preprocessing module, the curvature data processing part is composed of the Ffilter and Sfilter. The image part is composed of the image preprocessing module.

(2) The feature extraction module consists of a curvature feature extraction module (CFE) and an image feature extraction module based on YOLOv8 (IFE). The CFE consists of front and back networks. The front network executes a simplified random forest consisting of three decision trees for curvature feature extraction tasks, while the back network executes linear regression tasks.

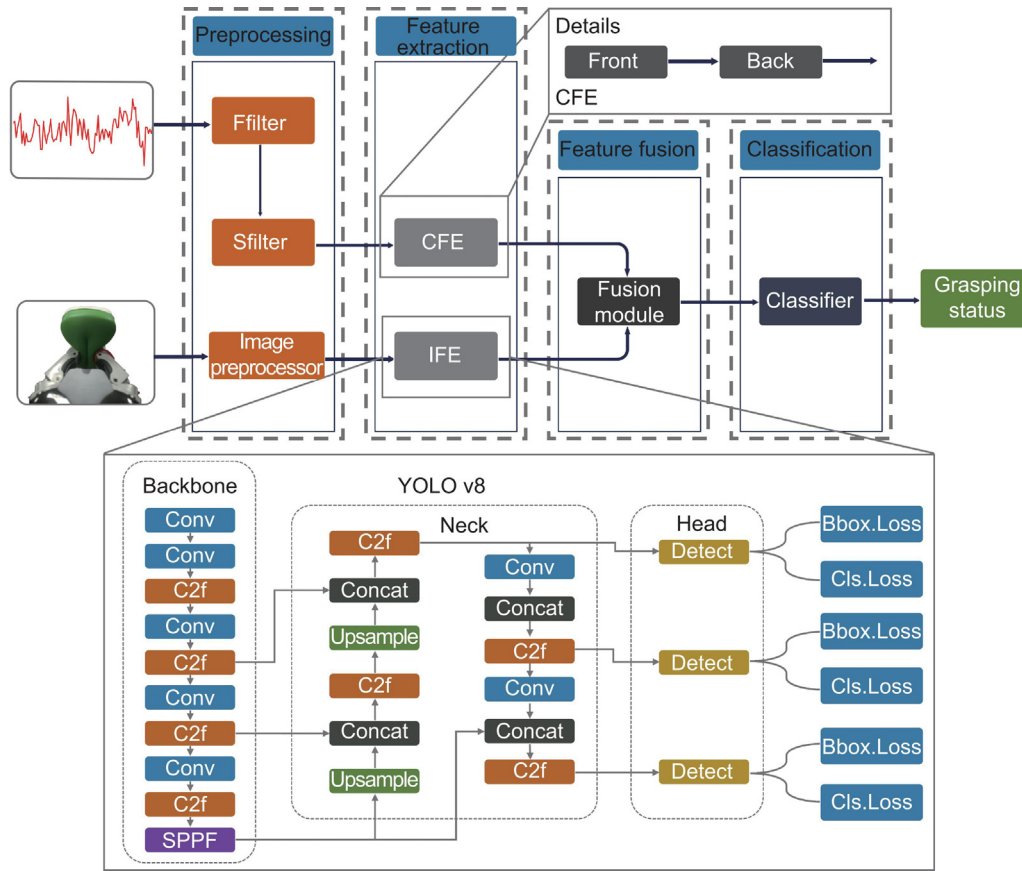


Fig. 5. Diagram of the VCFN-YOLOv8 framework.

The IFE consists of three main parts: backbone, neck, and head. The backbone serves as the foundation and is responsible for extracting features from the input image. The head produces the final detection result. The neck is located between the backbone and the head, used for fusing and enhancing features.

(3) The feature fusion module is used to combine the features of different sensors. In this process, the two originally independent modalities are fused based on Eq. (1).

(4) The classification module has four output channels, each providing one evaluation result, as shown in Fig. 1. The classification module is formulated as Eq. (2).

4. Experiments and results

In this section, extensive experiments are described to test the performance of the proposed evaluation method. The experiments are conducted on a computer running Windows 10, equipped with a multi-core 3.2 GHz Intel Core i9-12900K CPU, 64 GB of system memory (RAM), and two NVIDIA GeForce RTX 3090 graphics cards. All the experiments are carried out using a KINOVA JACO2-j2n6s200 robotic arm with a two-fingered gripper as the end effector. Each finger of the gripper is equipped with an actuator and the maximum grasping force is 25N. The maximum travel time of the gripper is 1.2 s. Specifically, the curvature sensors cover the opening and closing parts of the two fingers of the gripper, and the HIKVISION DS-E12 camera is mounted on the wrist of the robotic arm. The hardware implementation of acquisition system is illustrated in Fig. 6. The data from the curvature sensor are collected by the ADC module of the STM32F103C8T6 microcontroller and are sent to the computer via a serial bus to process subsequent data.

In the following experiment, the values of parameters are set as follow: in Eq. (3), $\omega_{visual} = 0.8$ and $\omega_{curvature} = 0.2$. In Section 3.2.2, $W_{orig} = 640$, $H_{orig} = 480$ and $W_{target} = H_{target} = 640$. In Eq. (14), $P_W = 0$ and $P_H = 80$.

4.1. Dataset

The visual curvature dataset (VCDS) is constructed through extensive grasping and lifting experiments on 9 soft deformable objects with different sizes, shapes, textures, and weights, which are shown in Fig. 7. In the grasping experiments, the objects are grasped at four preset widths corresponding to four different camera angles, and then slowly lifted by 100.0 mm (with a lifting speed set at 10.0 mm/s). Among them, four camera angles are defined as facing directly above the grasping objects, and then rotating clockwise by 90° , 180° and 270° respectively. During the grasping and lifting process, the visual and curvature data are collected with the frequencies of 30 Hz and 60 Hz respectively. For each object, 16 grasping experiments are conducted, and after filtering, approximately 270 to 280 frames of visual images and 540 to 560 frames of curvature signal sequences are collected for each trial. After removing redundant data, the final VCDS dataset consists of 8,000 single-frame visual images and 160 sets of curvature signal sequence samples.

4.2. Performance testing

The accuracy and F1-score are used to evaluate the model's classification performance comprehensively and accurately, with higher accuracy and F1-scores indicating better classification performance. It is generally believed that factors such as different input sequence lengths and model structures can impact the performance of the model.

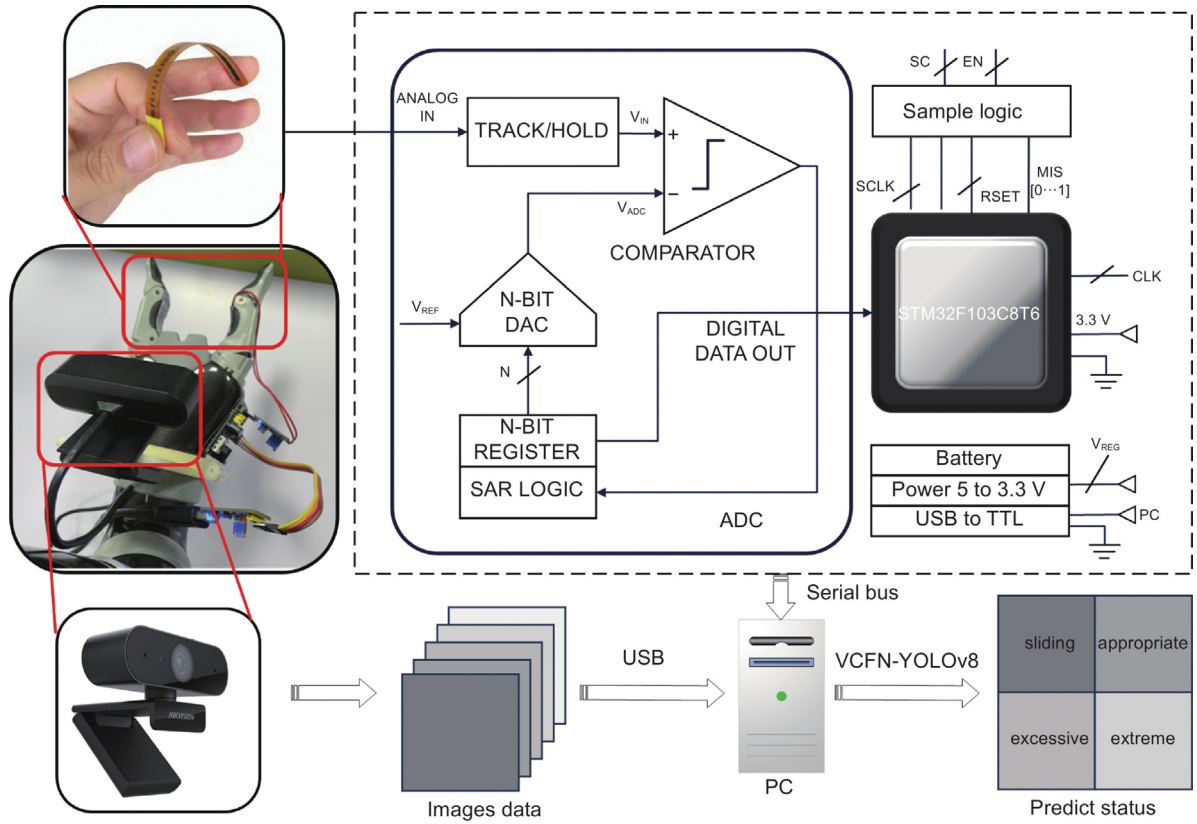


Fig. 6. The implementation process from the hardware perspective. The left side shows the experimental equipment, with the curvature sensor at the top, the modified gripper in the middle and the camera at the bottom. On the right, the top section depicts the hardware structure of curvature modality, while the bottom section shows the hardware structure of the visual modality and the predicted results of the entire structure.

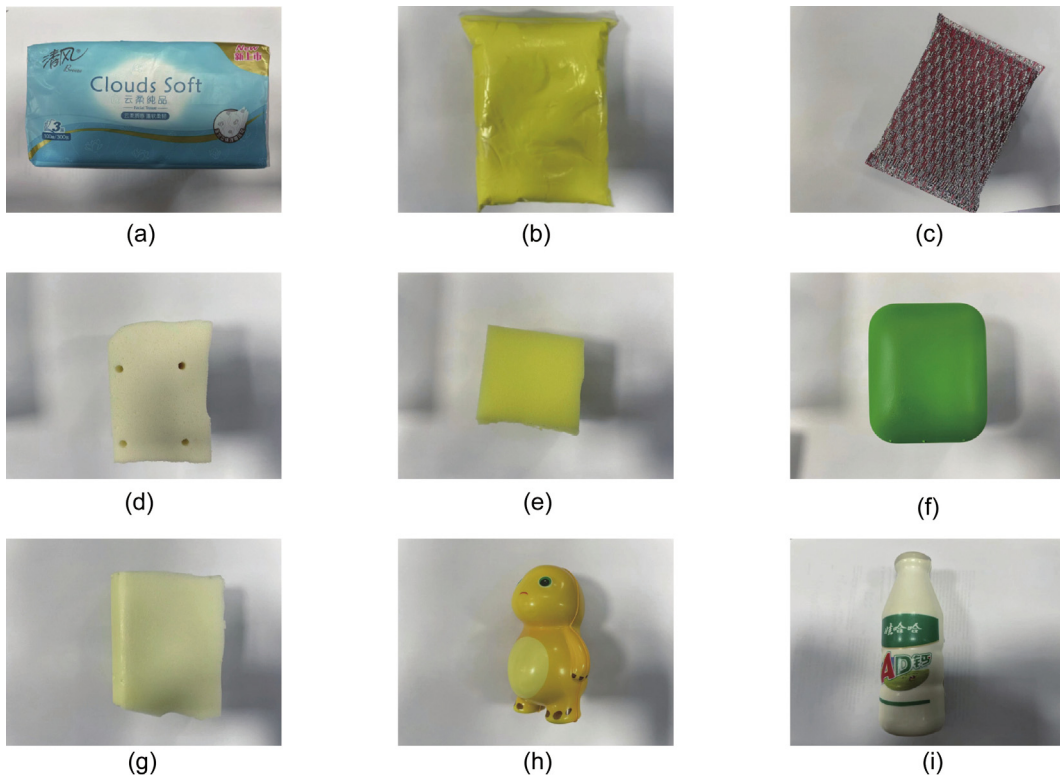


Fig. 7. The soft objects used in the grasp experiments.

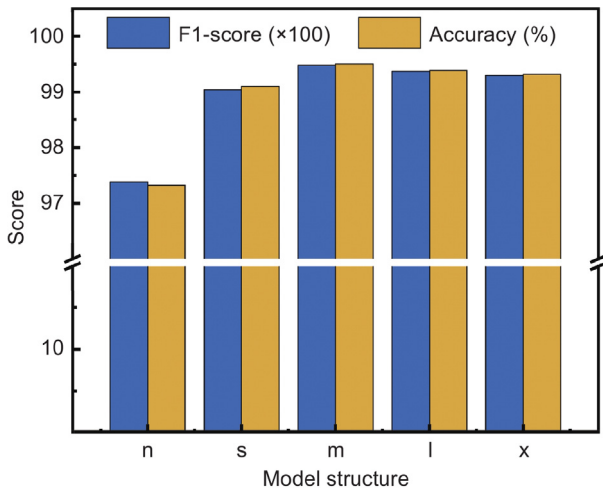


Fig. 8. Experimental results of models with different structures.

4.2.1. Structure comparison experiments

Due to the varying parameter counts in different model structures, more complex models require greater computational resources and higher-end hardware support, theoretically resulting in higher accuracy but slower computation speeds. The models with different structures (including YOLOv8n-clc, YOLOv8s-clc, YOLOv8m-clc, YOLOv8l-clc and YOLOv8x-clc) are compared to evaluate the grasping state of soft deformable objects. The experimental results are shown in Fig. 8.

The results indicate that the YOLOv8n-clc model had the lowest accuracy and F1 score. However, more complex models do not necessarily offer better classification performance. As shown in the data, the most complex YOLOv8x-clc model performed worse than the simpler YOLOv8m-clc model, with its accuracy and F1 score lower by 0.19% and 0.18%, respectively. This is because, increasing the number of parameters to achieve better performance can make the model more complex and focus on more data details, which can reduce classification accuracy. The results suggest that the YOLOv8m-clc model provides the best classification performance.

4.2.2. Sequence lengths comparison experiments

Different sequence lengths imply variations in information volume, computational load and noise. Models with visual sequence lengths of 2, 4, 8, 16, 32 and 64 are compared, and experiments are conducted using the corresponding curvature sequence lengths. The experimental results are shown in Fig. 9.

The results show that the model's classification has the best performance when the input sequence length is set as 16, with a 0.51% improvement in accuracy compared to a length of 2. When the input sequence is too long, classification accuracy slightly decreases due to increased computational load and noise interference. Therefore, there is no need to increase sequence length to improve model performance.

4.2.3. Modal comparison experiments

In this section, three comparative experiments are conducted to test various modality combinations, aiming to validate the performance advantages of visual-curvature fusion perception. First, the data are preprocessed accordingly. For the Visual-only mode, The visual feature extraction module and classification module in the VCFN-YOLOv8 framework are selected (as shown in Fig. 5). Similarly, the curvature-only mode is tested by combining the curvature feature extraction module and the classification module. The comparison results are shown in Table 1.

Table 1

	Curvature-only	Visual-only	V-C fusion
Accuracy	86.04	95.33	99.51
Precision	86.11	95.29	99.46
Recall	86.14	95.49	99.51
F1-score	86.12	95.39	99.48

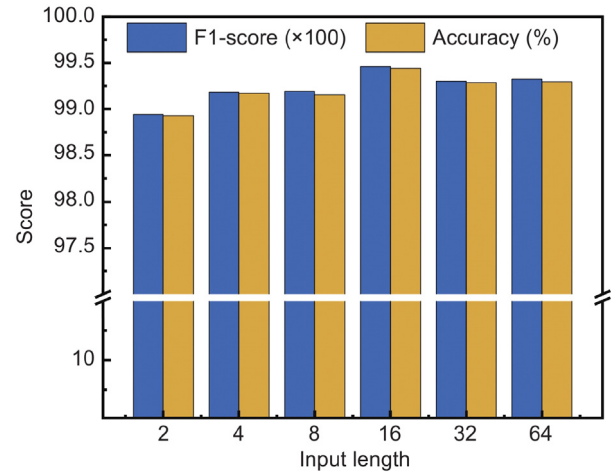


Fig. 9. Experimental results of models with different input length.

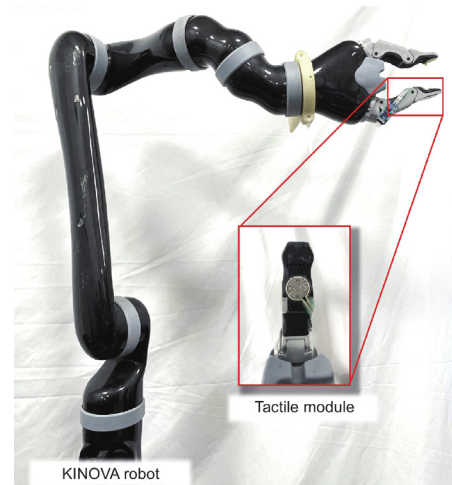


Fig. 10. Overview of the experimental scene.

According to the experimental results in Table 1, the visual-curvature fusion perception achieved better results than any single modality perception. In theory, visual images provide geometric information about the contact conditions, which helps better distinguish the extreme grasp state. The visual images contribute more information about the entire grasping process, which is why the Visual-only mode performs better, aligning with people's everyday understanding.

4.2.4. Method comparison experiments

Furthermore, tactile sensing experiments are conducted to verify the advantage of the VCFN-YOLOv8 framework. The tactile module RP-C1.8-LT is attached to the fingers of the KINOVA robot to obtain tactile data of grasping evaluation. The experimental scene is set up as shown in Fig. 10. From the set of experimental objects, light clay, latex sponge and ordinary sponge are selected for the grasping evaluation experiments of tactile methods. These

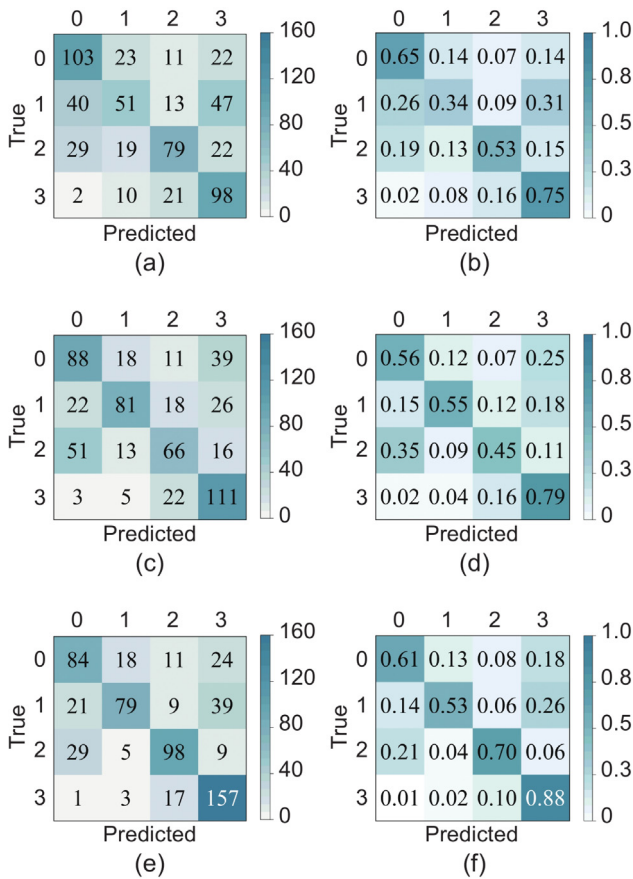


Fig. 11. Results of using the tactile method. (a) Confusion matrix of the grasping state evaluation results using tactile method for the light clay. (b) Normalized confusion matrix of the grasping state evaluation results using tactile method for the light clay. (c) Confusion matrix of the grasping state evaluation results using tactile method for the latex sponge. (d) Normalized confusion matrix of the grasping state evaluation results using tactile method for the latex sponge. (e) Confusion matrix of the grasping state evaluation results using tactile method for the ordinary sponge. (f) Normalized confusion matrix of the grasping state evaluation results using tactile method for the ordinary sponge.

three selected objects are subjected to a total of 1784 experiments, where light clay and latex sponge are tested 590 times and ordinary sponge is tested 604 times. During the process of grasping evaluation, these three objects are uniformly grasped and released.

The confusion matrix and normalized confusion matrix of the experimental results are shown in Fig. 11. The accuracy is selected as evaluation indicator for comparison between different methods, with the results shown in Fig. 12.

According to the experimental results, the tactile method is inferior to the curvature method and V-C fusion method in terms of the accuracy. This suggests that in the perception of soft objects, the single tactile feedback may not fully reflect the geometric characteristics of the object, especially when the object has a complex shape or undergoes a large range of deformation. In addition, the information from the confusion matrix of Fig. 11 shows that the evaluating accuracy of the “Extreme” state is higher. This is because in this state, a part of the experimental object will produce a larger supporting force feedback after being compressed to the minimum value.

4.2.5. Computational efficiency

In this study, the computational time of the VCFN-YOLOv8 framework is summarized from three parts that are the process

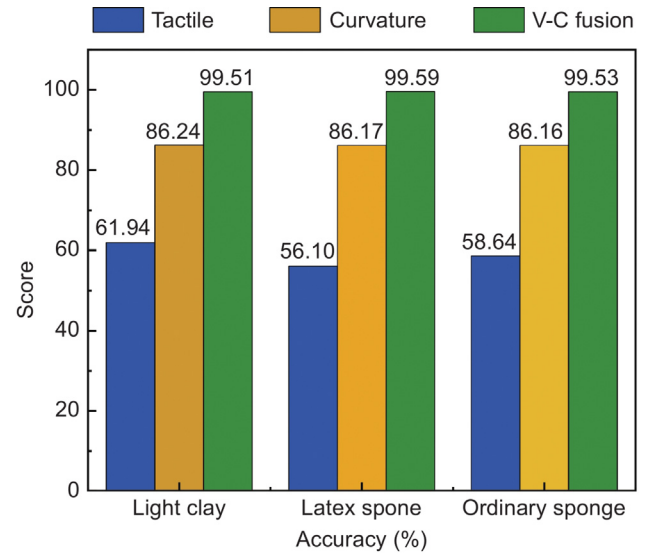


Fig. 12. Experimental results of different methods.

Table 2

The table of time value.

Variable	T_{cmp}	t_{pc}	t_{pi}	t_{ec}	t_{ei}	t_{fc}
value(ms)	38	19.5	6.9	16.7	1.3	1.8

of preprocessing, feature extraction, fusion and classification. This process can be formalized as:

$$T_{cmp} = \max \{t_{pc}, t_{pi}\} + \max \{t_{ec}, t_{ei}\} + t_{fc}. \quad (15)$$

In this equation, t_{pc} represents the curvature data preprocessing time, t_{pi} represents the image data preprocessing time, t_{ec} represents the curvature data feature extraction time and t_{ei} represents the image data feature extraction time, t_{fc} represents the time of fusion and classification and T_{cmp} represents the time of conducting an evaluation. The values of these time variables in the experiments are shown in Table 2.

The computational efficiency is defined as frame rate in this study. Obviously, it can be formalized as:

$$E_{cmp} = \frac{1}{T_{cmp}}. \quad (16)$$

In Eq. (16), E_{cmp} represents the computational efficiency and its value is 26.3 fps.

4.3. Generalization experiments

Four sets of experiments are designed to verify the model's versatility and effectiveness in practical applications. The experiments included grasping state evaluation and adaptive grasping for both soft and non-soft deformable objects.

4.3.1. Grasping state evaluation experiments

To test the model's versatility, the experimental setup described in the introduction is used to evaluate grasping states (sliding, appropriate, excessive and extreme) for soft hollow foam and disposable paper cups, performing 25 trials for each state. The confusion matrices for the grasping state evaluations of soft hollow foam and disposable paper cups are shown in Figs. 13.

The results indicate that the framework demonstrated good versatility in evaluating the grasping state of unfamiliar soft deformable objects and non-soft deformable objects. The classification accuracy of grasping state for soft hollow foam and disposable paper cups reached 96% and 88%, respectively. Due to the

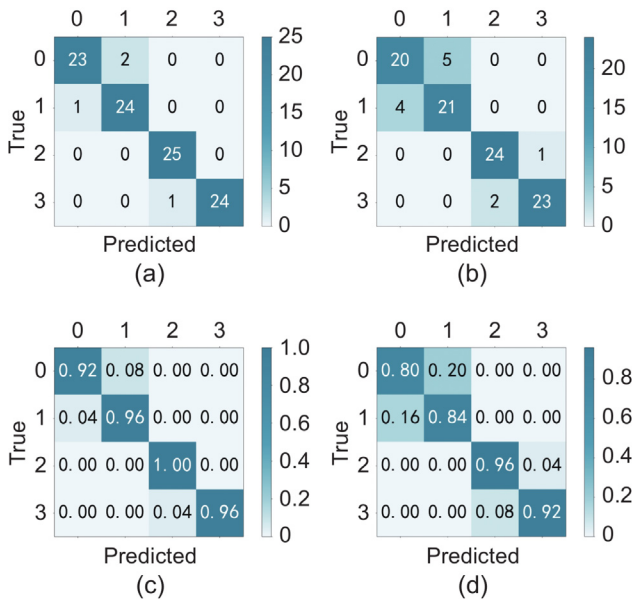


Fig. 13. Results of grasp state evaluation tests. (a) Confusion matrix for the soft hollow foam. (b) Confusion matrix for the disposable paper cup. (c) Normalized confusion matrix for the soft hollow foam. (d) Normalized confusion matrix for the disposable paper cup.

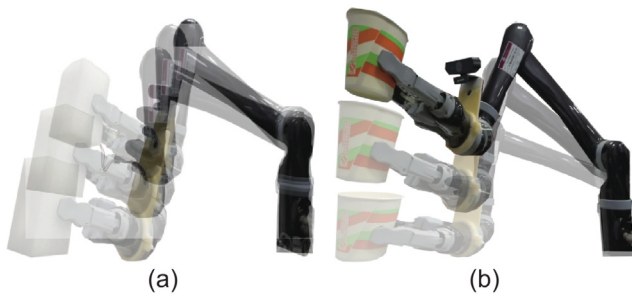


Fig. 14. Adaptive grasping process. (a) Grasping process of soft hollow foam; (b) Grasping process of disposable paper cups.

characteristics of disposable paper cups, the differences between sliding and appropriate grasping states are small, which led to reduced accuracy for these states.

4.3.2. Adaptive grasping experiments

To further validate the model's effectiveness in practical applications, detailed adaptive grasping experiments for soft deformable objects are conducted based on real-time feedback from the VCFN-YOLOv8 model. The detailed grasp regulation strategies are as follows:

$$W_{t+1} = W_t + (g(t) - g(0)) \quad (17)$$

Where W_t and $g(t)$ represent the grasping width of the gripper and the evaluated state at the current moment. The W_{t+1} represents the grasping width of the gripper at the next moment. The unit of W_{t+1} and W_t is in millimeter. The $g(0) = 1$ denotes the appropriate state. When the grasp is too tight, the width of the gripper tends to increase, and vice versa. Specifically, appropriate grasping experiments for soft hollow foam and disposable paper cups are performed, as shown in Figs. 14. The adjustment time, curvature feedback values and real-time grasping status are recorded, as shown in Figs. 15 and 16.

Figs. 15 and 16 indicate that, after a certain adjustment period, the appropriate grasping of soft hollow foam and disposable

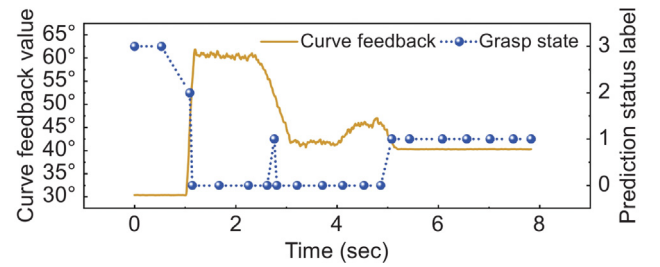


Fig. 15. Real-time variation curves of different values during the soft hollow foam grasping experiment.

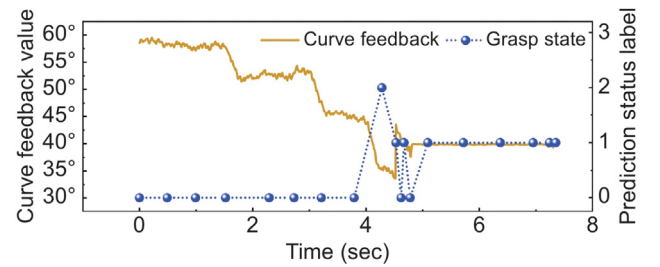


Fig. 16. Real-time variation curves of different values during the disposable paper cup grasping experiment.

paper cups is achieved successfully by the robotic arm's gripper. During this process, the grasping state evaluation is continuously updated with changes in the actual state, and the gripper adjusts the grasping width accordingly. For example, Fig. 15 shows that the gripper initially performed an extreme grasp and then adjusted to achieve appropriate grasping. In Fig. 16, during the appropriate grasping experiment for the disposable paper cup, the gripper achieved the target after three attempts and fine-tuning in about 4.6 s. Although the adjustment process varied across experiments, it effectively demonstrated the model's validity.

In conclusion, the model's versatility and practical effectiveness is validated by four sets of experiments effectively.

5. Conclusion

To effectively improve the grasping evaluation effect of soft objects, a novel VCFN-YOLOv8 fusion method is presented. The YOLOv8m-cls structure is used in the VCFN-YOLOv8 framework to extract visual features and achieve organic fusion of visual and curvature information through cross-modal feature extraction and attribute association mechanisms, providing a solution for grasping evaluating tasks in the dual-modal information fusion domain. Additionally, a VCDS dataset is established through extensive grasping and lifting experiments from different camera angles, which demonstrated the model's effectiveness and accuracy. Finally, the effectiveness and generalization of the framework are validated by experiments. The proposed architecture can be widely applied in fields such as workshop object sorting and quality inspection. Future work will explore information fusion models that are more in line with human perceptual characteristics, such as solutions for fusion perception of visual, olfactory and tactile senses.

CRediT authorship contribution statement

Guoshun Cui: Writing – original draft, Validation, Resources, Methodology, Formal analysis. **Shiwei Su:** Writing – review & editing, Visualization, Software, Formal analysis. **Hanyu Gao:** Writing – original draft, Visualization, Validation, Investigation,

Data curation. **Kai Zhuo**: Writing – review & editing, Validation. **Kun Yang**: Writing – review & editing, Supervision, Funding acquisition, Formal analysis. **Hang Wu**: Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Fundamental Research Project of Shanxi Province (202403021211229).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.birob.2025.100232>.

References

- [1] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou, Y. Mezouar, Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey, *Int. J. Robot. Res.* 37 (7) (2018) 688–716, <http://dx.doi.org/10.1177/0278364918779698>.
- [2] R. Reddy, S.R. Nagaraja, Integration of robotic arm with vision system, in: 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014, pp. 1–5, <http://dx.doi.org/10.1109/ICCIC.2014.7238302>.
- [3] J. Kwiatkowski, D. Cockburn, V. Duchaine, Grasp stability assessment through the fusion of proprioception and tactile signals using convolutional neural networks, in: 2017 IEEE/RSS International Conference on Intelligent Robots and Systems, IROS, 2017, pp. 286–292, <http://dx.doi.org/10.1109/IROS.2017.8202170>.
- [4] D. Cockburn, J.-P. Roberge, T.-H.-L. Le, A. Maslyczyk, V. Duchaine, Grasp stability assessment through unsupervised feature learning of tactile images, in: 2017 IEEE International Conference on Robotics and Automation, ICRA, 2017, pp. 2238–2244, <http://dx.doi.org/10.1109/ICRA.2017.7989257>.
- [5] C.B. Teeple, T.N. Koutros, M.A. Graule, R.J. Wood, Multi-segment soft robotic fingers enable robust precision grasping, *Int. J. Robot. Res.* 39 (14) (2020) 1647–1667.
- [6] W. Zheng, Y. Xie, B. Zhang, J. Zhou, J. Zhang, Dexterous robotic grasping of delicate fruits aided with a multi-sensory e-glove and manual grasping analysis for damage-free manipulation, *Comput. Electron. Agric.* 190 (2021) 106472, <http://dx.doi.org/10.1016/j.compag.2021.106472>.
- [7] H. Ning, X. Zheng, X. Lu, Y. Yuan, Disentangled representation learning for cross-modal biometric matching, *IEEE Trans. Multimed.* 24 (2022) 1763–1774, <http://dx.doi.org/10.1109/TMM.2021.3071243>.
- [8] Y. Gan, B. Zhang, J. Shao, Z. Han, A. Li, X. Dai, Embodied intelligence: Bionic robot controller integrating environment perception, autonomous planning, and motion control, *IEEE Robot. Autom. Lett.* 9 (5) (2024) 4559–4566, <http://dx.doi.org/10.1109/LRA.2024.3377559>.
- [9] M. Popović, D. Kraft, L. Bodenhausen, E. Başeski, N. Pugeault, D. Kragic, T. Asfour, N. Krüger, A strategy for grasping unknown objects based on co-planarity and colour information, *Robot. Auton. Syst.* 58 (5) (2010) 551–565, <http://dx.doi.org/10.1016/j.robot.2010.01.003>.
- [10] J. Liang, J. Zhang, B. Pan, S. Xu, G. Zhao, G. Yu, X. Zhang, Visual reconstruction and localization-based robust robotic 6-DoF grasping in the wild, *IEEE Access* 9 (2021) 72451–72464, <http://dx.doi.org/10.1109/ACCESS.2021.3079245>.
- [11] J. Ge, L. Mao, J. Shi, Y. Jiang, Fusion-mask-RCNN: Visual robotic grasping in cluttered scenes, *Multimedia Tools Appl.* 83 (7) (2024) 20953–20973.
- [12] B. Luo, H. Chen, F. Quan, S. Zhang, Y. Liu, Natural feature-based visual servoing for grasping target with an aerial manipulator, *J. Bionic Eng.* 17 (2020) 215–228.
- [13] E. Hyttinen, D. Kragic, R. Detry, Learning the tactile signatures of prototypical object parts for robust part-based grasping of novel objects, in: 2015 IEEE International Conference on Robotics and Automation, ICRA, 2015, pp. 4927–4932, <http://dx.doi.org/10.1109/ICRA.2015.7139883>.
- [14] A. Taghipour, A. Rostami, M. Bahrami, H. Baghban, M. Dolatyari, Comparative study between LPGA- and FBG-based bending sensors, *Opt. Commun.* 312 (2014) 99–105, <http://dx.doi.org/10.1016/j.optcom.2013.09.020>.
- [15] G. Gerboni, A. Diodato, G. Ciuti, M. Cianchetti, A. Menciassi, Feedback control of soft robot actuators via commercial flex bend sensors, *IEEE/ASME Trans. Mechatronics* 22 (4) (2017) 1881–1888, <http://dx.doi.org/10.1109/TMECH.2017.2699677>.
- [16] Z. Shen, J. Yi, X. Li, L.H.P. Mark, Y. Hu, Z. Wang, A soft stretchable bending sensor and data glove applications, in: 2016 IEEE International Conference on Real-Time Computing and Robotics, RCAR, 2016, pp. 88–93, <http://dx.doi.org/10.1109/RCAR.2016.7784006>.
- [17] R.P. Rocha, P.A. Lopes, A.T. de Almeida, M. Tavakoli, C. Majidi, Fabrication and characterization of bending and pressure sensors for a soft prosthetic hand, *J. Micromech. Microeng.* 28 (3) (2018) 034001, <http://dx.doi.org/10.1088/1361-6439/aaa1d8>.
- [18] K. Huebner, K. Welke, M. Przybylski, N. Vahrenkamp, T. Asfour, D. Kragic, R. Dillmann, Grasping known objects with humanoid robots: A box-based approach, in: 2009 International Conference on Advanced Robotics, 2009, pp. 1–6.
- [19] J. Shi, J. Zheng, X. Liu, W. Xiang, Q. Zhang, Novel short-time fractional Fourier transform: Theory, implementation, and applications, *IEEE Trans. Signal Process.* 68 (2020) 3280–3295, <http://dx.doi.org/10.1109/TSP.2020.2992865>.
- [20] M. He, L. Qin, X. Deng, K. Liu, MFI-YOLO: Multi-fault insulator detection based on an improved YOLOv8, *IEEE Trans. Power Deliv.* 39 (1) (2024) 168–179, <http://dx.doi.org/10.1109/TPWRD.2023.3328178>.
- [21] H.-W. Lee, The study of mechanical arm and intelligent robot, *IEEE Access* 8 (2020) 119624–119634, <http://dx.doi.org/10.1109/ACCESS.2020.3003807>.
- [22] S. Li, H. Yu, W. Ding, H. Liu, L. Ye, C. Xia, X. Wang, X.-P. Zhang, Visual-tactile fusion for transparent object grasping in complex backgrounds, *IEEE Trans. Robot.* 39 (5) (2023) 3838–3856, <http://dx.doi.org/10.1109/TRO.2023.3286071>.
- [23] J. Hackett, M. Shah, Multi-sensor fusion: a perspective, in: Proceedings, IEEE International Conference on Robotics and Automation, vol.2, 1990, pp. 1324–1330, <http://dx.doi.org/10.1109/ROBOT.1990.126184>.
- [24] S. Cui, R. Wang, J. Wei, F. Li, S. Wang, Grasp state assessment of deformable objects using visual-tactile fusion perception, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, 2020, pp. 538–544, <http://dx.doi.org/10.1109/ICRA40945.2020.9196787>.
- [25] S. Cui, R. Wang, J. Wei, J. Hu, S. Wang, Self-attention based visual-tactile fusion learning for predicting grasp outcomes, *IEEE Robot. Autom. Lett.* 5 (4) (2020) 5827–5834, <http://dx.doi.org/10.1109/LRA.2020.3010720>.
- [26] Y. Han, K. Yu, R. Batra, N. Boyd, C. Mehta, T. Zhao, Y. She, S. Hutchinson, Y. Zhao, Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer, *IEEE/ASME Trans. Mechatronics* (2024).
- [27] A. Depierre, E. Dellandréa, L. Chen, Scoring graspability based on grasp regression for better grasp prediction, in: 2021 IEEE International Conference on Robotics and Automation, ICRA, 2021, pp. 4370–4376, <http://dx.doi.org/10.1109/ICRA48506.2021.9561198>.