Xiaofei QIN, Wenkai HU, Chen XIAO, Changxiang HE, Songwen PEI, Xuedian ZHANG, 2023. Attention-based efficient robot grasp detection network. *Frontiers of Information Technology & Electronic Engineering*, 24(10):1430-1444. https://doi.org/10.1631/FITEE.2200502

Attention-based efficient robot grasp detection network

Key words: Robot grasp detection; Attention mechanism; Encoder–decoder; Neural network

Corresponding author: Xuedian ZHANG
E-mail: obmmd_zxd@163.com
ORCID: https://orcid.org/0000-0002-7636-7517

Motivation

1. Although many existing methods have achieved good performance, the balance between the real-time requirement and accuracy still needs improvement.
2. Although the predicted rectangles of the top row seem to meet the intersection over union (IoU) threshold, if the grasp is performed,a fingertip on one side of the gripper will collide with the object.
3. The IoU item in the loss function is improved by using an hourglass-shaped predicted grasp box instead of a rectangular one when IoU is calculated, which will reduce the influence of the central area in the predicted grasp boxes and create good correspondence between high IoU values and grasping success rates.

Problem formulation

First, the grasp detection algorithm is used to detect the pixel-level grasp configuration, which is noted as $\boldsymbol{G}$. In this study, $\boldsymbol{G}$ is defined as follows:

$$\boldsymbol{G} = (\boldsymbol{W}, \boldsymbol{\Theta}, \boldsymbol{Q}) \in \mathbb{R}^{3\times H\times W}, \tag{1}$$

where $\boldsymbol{W}$, $\boldsymbol{\Theta}$, and $\boldsymbol{Q}$ represent three output images with the same size as the input image. Each pixel in these images can be regarded as the width, rotation angle, and grasp quality score of a grasp rectangle candidate.

Second, an argmax operation is performed on $\boldsymbol{Q}$ of $\boldsymbol{G}$ to obtain the pixel coordinates of the highest grasp quality score, and then the coordinates are used as the index to obtain the width, rotation angle, and grasp quality score of the grasp rectangle in the pixel coordinate system $\boldsymbol{G_p}$. In this study, $\boldsymbol{G_p}$ is defined as follows:

$$\boldsymbol{G_p} = (\boldsymbol{p}, w_\mathrm{p}, \theta_\mathrm{r}, q)\,, \tag{2}$$

where $\boldsymbol{p}$ denotes the pixel coordinates (u, v) of the grasp center point, w_p is the width of the pixel-level grasp rectangle, θ_r is the deflection angle relative to the horizontal line of the image in the range of $[\frac{-\pi}{2}, \frac{\pi}{2}]$, and q is the pixel-level grasp quality score in the range of $[0, 1]$.

$$\boldsymbol{G}_\mathrm{c} = (\boldsymbol{P}, w_\mathrm{c}, \theta_\mathrm{r}, q)\,, \tag{3}$$

where $\boldsymbol{P}$ denotes the coordinates (x, y, z) of the grasp center point in the camera coordinate system, w_c is the width of the grasp in the camera coordinate system (i.e., the difference of the opening and closing degrees of the gripper), θ_r is the same as that in Eq. (2) because in this study the eye-in-hand robot grasping system is used, and q represents the grasp quality score.

Finally, $\boldsymbol{G}_\mathrm{c}$ is converted into the robot coordinate system through external parameters of the robot and camera, which is noted as $\boldsymbol{G}_\mathrm{r}$. The definition of $\boldsymbol{G}_\mathrm{r}$ is similar to that of $\boldsymbol{G}_\mathrm{c}$, and will not be given here. The convertion process from $\boldsymbol{G_p}$ to $\boldsymbol{G}_\mathrm{r}$ can be represented as follows:

$$\boldsymbol{G}_\mathrm{r} = \mathrm{Trans}_\mathrm{rc}\left(\boldsymbol{G}_\mathrm{c}\right) = \mathrm{Trans}_\mathrm{rc}\left(\mathrm{Trans}_\mathrm{cp}\left(\boldsymbol{G_p}\right)\right), \tag{4}$$

$$\mathrm{Trans}_\mathrm{rc}(\boldsymbol{G}_\mathrm{c}) = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ \boldsymbol{0} & \boldsymbol{1} \end{bmatrix} \boldsymbol{G}_\mathrm{c} = \boldsymbol{G}_\mathrm{r}, \tag{5}$$

where $\boldsymbol{R}$ is the rotation matrix from the camera coordinate system to the robot coordinate system, and $\boldsymbol{T}$ is the translation matrix. Similarly, $\boldsymbol{R}$ in $\mathrm{Trans}_\mathrm{cp}$ is the rotation matrix from the pixel coordinate system to the camera coordinate system, and $\boldsymbol{T}$ in $\mathrm{Trans}_\mathrm{cp}$ is the translation matrix.

Method

An end-to-end grasp detection convolutional neural network (CNN) named AE-GDN

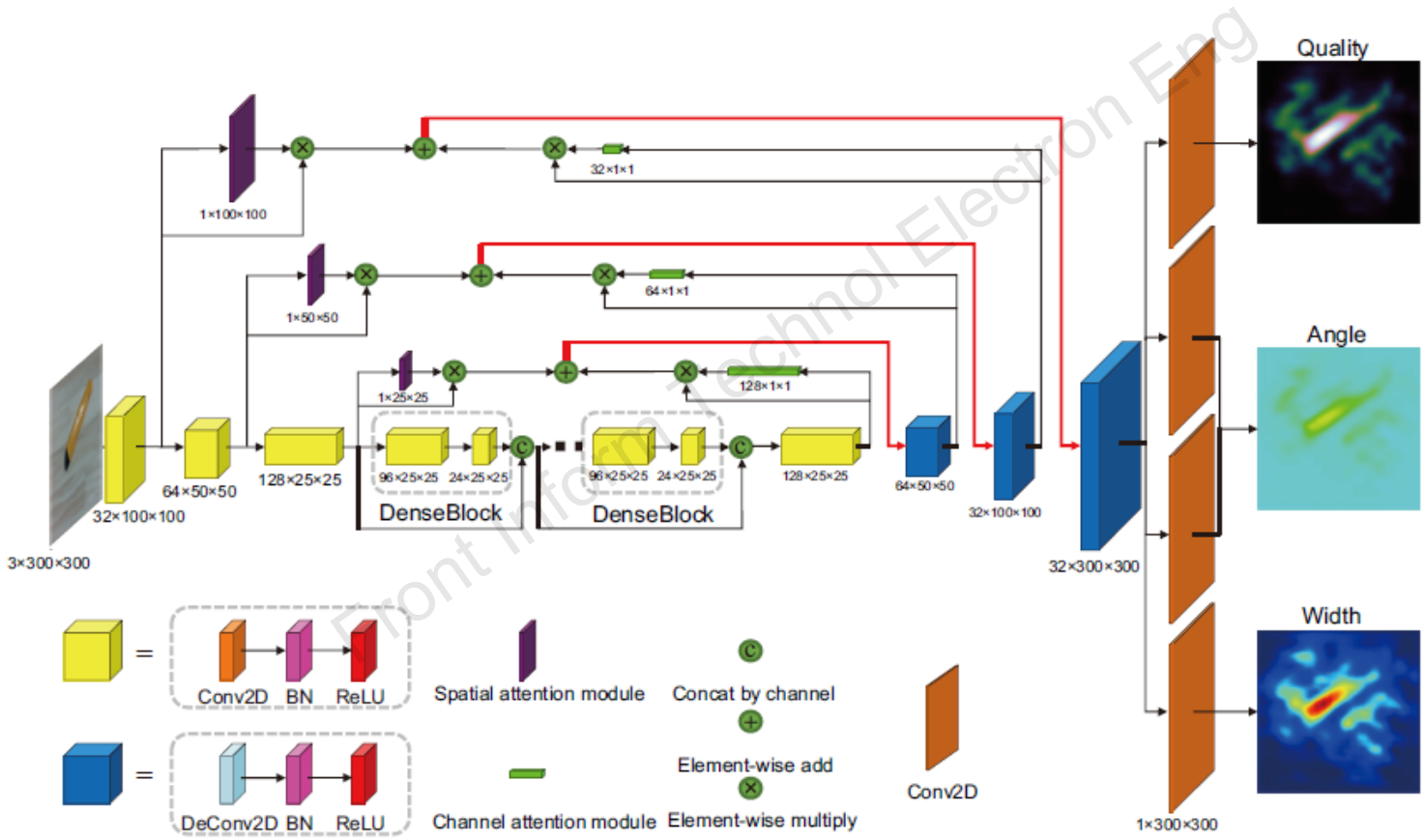

Fig. 3 Model architecture

BN: batch normalization; ReLU: rectified linear unit. In the figure, the red lines represent the fused feature inputted into the decoder stages. References to color refer to the online version of this figure

Method

Model architecture and loss function

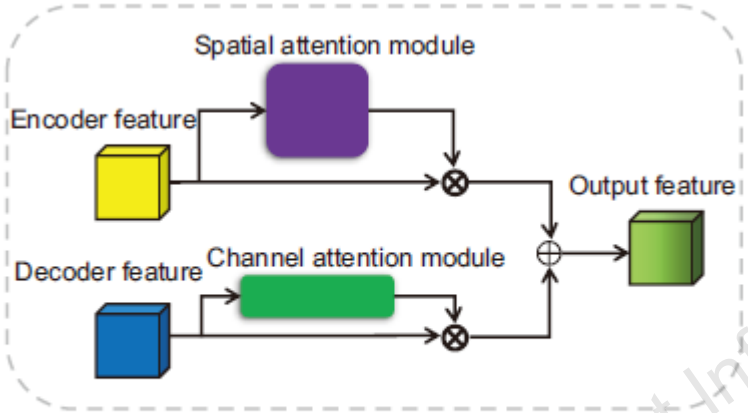

Fig. 4 The improved attention module for feature fusion

$$\mathcal{L} = \mathcal{L}_{\text{quality}} + \mathcal{L}_{\text{angle}}^{\cos} + \mathcal{L}_{\text{angle}}^{\sin} + \mathcal{L}_{\text{width}} + \mathcal{L}_{\text{GIoU}}, \quad (6)$$

$$\mathcal{L}_{\text{quality}} = \text{MSE}\left(q_{\text{g}}, \hat{q}_{\text{g}}\right), \quad (7)$$

$$\text{GIoU} = \text{IoU} - \frac{\mathcal{E}\left(P,G\right) - \mathcal{U}\left(P,G\right)}{\mathcal{E}\left(P,G\right)}, \quad (8)$$

$$\text{GIoU} = -\frac{E}{G+P+E} = -\frac{1}{\frac{G+P}{E}+1}. \quad (9)$$

$$\begin{aligned} \text{GIoU} &= \frac{I}{G+P-I} - \frac{E}{G+P+E} \\ &= \frac{1}{\frac{G+P}{I}-1} - \frac{1}{\frac{G+P}{E}+1}. \end{aligned} \quad (10)$$

$$\mathcal{L}_{\text{GIoU}} = 0.5 - \text{GIoU}. \quad (11)$$

Method

The effect of the generalized intersection over union (GIoU)

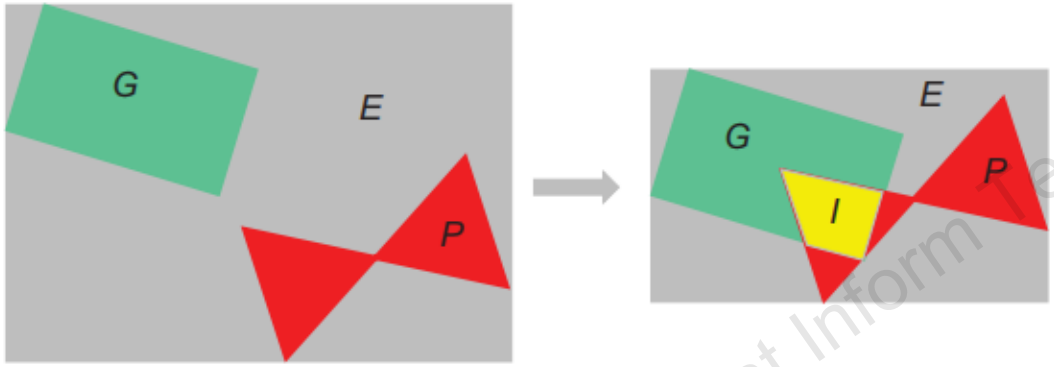

Fig. 5 Examples of the ground truth (GT) rectangle and the predicted grasp box for calculating the generalized intersection over union (GIoU)

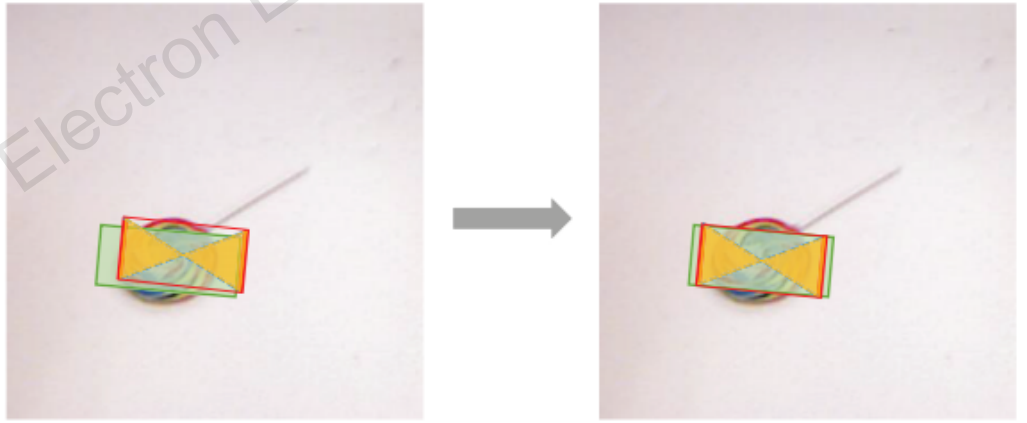

Fig. 6 The effect of the hourglass box, which drives the grasp detection result to change from the left (collision between the object and gripper despite a high IoU value) to the right during training. References to color refer to the online version of this figure

Conclusions

1. This paper proposes an efficient pixel-level grasp detection network AE-GDN based on an attention mechanism and an hourglass box matching mechanism.
2. The hourglass box matching mechanism creates good correspondence between high IoUs and high-quality grasp rectangles.
3. The inference speed of the proposed AE-GDN can meet the real-time requirement of robot grasping tasks.

Xiaofei QIN received the Ph.D. degree in control theory and control engineering from Institute of Automation, Chinese Academy of Sciences, Beijing, in 2010. Since 2015, he has been an associate professor with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. Dr. Qin was a member of Shanghai Association for Automation since 2015 and a member of Chinese Association for Artificial Intelligence since 2017. His research interests include artificial intelligence algorisms and applications, robotic technology and applications, power electronic technology, and motion control methods.

Wenkai HU received his B.E. degree in the University of Shanghai for Science and Technology in 2020. Currently, he is working toward his MS degree at the University of Shanghai for Science and Technology. His current research interests include deep learning, neural networks, and robot grasp detection.

Xuedian ZHANG Professor, PhD supervisor, is currently the Executive Vice Dean of the School of Opto-electronic Information and Computer Engineering, University of Shanghai for Science and Technology. His research interests include photoelectric detection technology and devices, biomedical imaging, artificial intelligence, embedded systems, optical fiber sensing technology, etc.