Pixel-Wise Grasp Detection via Twin Deconvolution and Multi-Dimensional Attention

Guangli Ren, Wenjie Geng, Peiyu Guan, Zhiqiang Cao, *Senior Member, IEEE*, and Junzhi Yu, *Fellow, IEEE*

***Abstract*— The grasp detection is crucial to high-quality robotic grasping. Typically, the mainstream encoder-decoder regression solution is attractive due to its high accuracy and efficiency, however, it is still challenging to solve the checkerboard artifacts from the uneven overlap of convolution results in decoder, and features from the encoder also need to be further refined. In this paper, a novel pixel-wise grasp detection network is proposed, which is composed of an encoder, a multi-dimensional attention bottleneck, and a decoder based on twin deconvolution. The proposed decoder introduces a twin branch upon the original transposed convolution branch. Through the overlap degree matrix provided by the twin branch, the original branch is re-weighted and then the checkerboard artifacts of the original branch are eliminated. Besides, to deeply explore the intrinsic relationship of features and strengthen feature discrimination, residual multi-head self-attention, cross-amplitude attention, and channel attention are integrated together. As a result, adaptive feature refinement is achieved. The effectiveness of the proposed method is verified by experiments.**

***Index Terms*— Grasp detection, multi-dimensional attention, twin deconvolution.**

I. Introduction

NOWADAYS, robotic manipulation has become an important domain in embodied artificial intelligence [1]. As an interdisciplinary technology of computer graphics, reinforcement learning, and robotics, the manipulation skill learning usually produces action policy by adequate training of the manipulator in the simulated interactive environment, such as full-physics SAPIEN simulator [2], SOFA physical simulation framework [3], and Gazebo simulator [4]. Please refer to the encyclopedic survey [5] for more details. During the simulated training executed by trial and error, the proper guidance is helpful for efficiency improvement. For the most typical object grasping task among robotic manipulation, grasp detection can serve as such a guidance by making the network more focus on the positions with highly graspable probability.

Existing grasp detections aim to find suitable grasps of the target object from the image or point cloud [6]. The early studies on grasp detection mainly find a proper grasp by feature matching of query object with a library of object models [7], [8] and its extensibility is weak as it depends on these models. With successful applications of deep learning [9], [10], fruitful outcomes with candidate-based [11], [12], [13] and regression-based [14], [15] types are proposed. The first type regards grasp detection as a two-stage task, which generates candidate grasps and then evaluates grasps with score ranking. In contrast, regression-based grasp detection is efficient as it predicts the grasp directly through a regression network. It is subdivided into encoder regression [14], [16], [17] and encoder-decoder regression [15], [18], [19], [20] according to the network structure. The former predicts grasp through an encoder procedure. Although it achieves real-time detection, the prediction tends to be the average of the ground truth in some cases, which is possibly invalid. With the advantage of the decoder that up-samples the feature map for fine-grained pixel-wise prediction, the encoder-decoder regression solution attains good accuracy. Since the pioneering method GG-CNN (generative grasp convolutional neural network) is proposed [15], a variety of grasp detection networks are designed: enhanced GG-CNN with semi-supervised feature extraction [18], generative residual convolutional neural network (GR-ConvNet) [19], and two-stream grasping network (TsGNet) [20].

Notice that encoder-decoder regression networks usually utilize cascaded deconvolution (i.e. transposed convolution) to up-sample the feature map for realizing grasp prediction of each pixel. However, such up-sampling process can cause the feature map to display checkerboard-like pattern called checkerboard artifacts [21], which appears commonly in image generative networks [22]. This phenomenon makes the feature map unsmooth and thus results in performance degeneration. Actually, the checkerboard artifacts originate from the uneven overlap of convolution results at different positions [21]. If the relative overlap differences among positions are computable, they can be used to constrain uneven overlap and thus the checkerboard artifacts are expected to be eliminated. Besides, to better focus on regions of interest and suppress

Manuscript received 7 July 2022; revised 4 October 2022 and 28 December 2022; accepted 12 January 2023. Date of publication 18 January 2023; date of current version 4 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62073322, in part by the CIE-Tencent Robotics X Rhino-Bird Focused Research Program under Grant 2022-07, and in part by the Beijing Natural Science Foundation under Grant 2022MQ05. This article was recommended by Associate Editor S. Gao. *(Guangli Ren and Peiyu Guan contributed equally to this work.) (Corresponding author: Zhiqiang Cao.)*

Guangli Ren, Wenjie Geng, Peiyu Guan, and Zhiqiang Cao are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhiqiang.cao@ia.ac.cn).

Junzhi Yu is with the Department of Advanced Manufacturing and Robotics, BIC-ESAT, College of Engineering, Peking University, Beijing 100871, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3237866.

Digital Object Identifier 10.1109/TCSVT.2023.3237866

redundant features, the attention mechanism has shown its advantage [23], [24]. A multimodal local-global attention network is proposed in affective video content analysis [23]. By extending the self-attention mechanism to multilevel fusion for key parts selection both from multimodal local streams and global sequences, a high performance is attained. Gao et al. [24] designed a co-attention module to highlight the co-salient objects while suppressing background. Generally speaking, the attention solutions concern the correlation about different pixel positions, channels, or multimodal information. On one hand, the detail representation of features may not be insufficient, which shall affect the performance of grasp detection. On the other hand, it does not deeply explore the intrinsic relationship of features such as the correlation among the row and column features.

The aforementioned analyses motivate us to build advanced encoder-decoder network. For this paper, the main contributions are as follows. A pixel-wise grasp detection network with twin deconvolution and multi-dimensional attention is proposed, which achieves good accuracy with real-time performance. Different from [25] and [26] that deal with checkerboard artifacts by adding extra operation during the original transposed convolution, we proposed a novel twin deconvolution up-sampling scheme, where a twin branch is introduced in parallel with the original transposed convolution branch. In the twin branch, the degree of overlap at each position is computed through new deconvolution based on a kernel with the same weights. Thus, the relative overlap differences among positions can be represented by the overlap degree matrix, which is used to re-weight the output of the original branch. As a result, the overlapped convolution results are smoothed and the checkerboard artifacts are eliminated. Besides, multi-dimensional attention is designed to achieve adaptive feature refinement. Particularly, cross-amplitude attention is provided to mine the intrinsic relationship of features, where cross attention promotes the stability of features by capturing the correlation of features between the column and row corresponding to each spatial position, and amplitude attention improves detail representation of features by paying more attention to the regions with strong gradient intensity. With the combination of information from the residual multi-head self-attention, feature discrimination is further enhanced.

II. Related Work

This section discusses the grasp detection methods based on deep learning from three aspects: candidate-based, encoder regression, and encoder-decoder regression solutions.

A. Candidate-Based Solution

It infers grasps in the form of candidate-evaluation. Lenz et al. [11] proposed a cascade coarse-to-fine network based on sparse auto-encoder (SAE) to evaluate candidate grasps generated by sliding window searching, and then optimal grasp is obtained. A problem is that the searching based on sliding window is time-consuming. To generate candidate grasps rapidly, many excellent schemes are proposed. Mahler et al. [12] obtained discretely antipodal candidate grasps from depth images by uniform sampling based on surface normal. In [13], an object detector for fast grasp generation is achieved by integrating region proposal network with sub-networks for grasp orientation and bounding box prediction. Other effective strategies include particle swarm optimizer [27], object skeleton [28], and domain-independent unsupervised clustering [29].

B. Encoder Regression

This solution utilizes an encoder network to regress proper grasps. Redmon et al. [14] leveraged AlexNet as backbone to perform single-stage regression of graspable bounding boxes, which gets rid of standard sliding window or region proposal techniques. By extracting and fusing features based on VGG-16, a robust grasp detection is achieved with the combination of classification and regression [16]. In [17], a multi-modal grasp detection network is presented to predict the grasp configuration, where two ResNet-50 are executed in parallel for integrating the RGB and depth features. Although real-time detection is achieved, the prediction is sometimes invalid.

C. Encoder-Decoder Regression

It generates pixel-wise grasp prediction through an encoder-decoder network. A generative grasp convolutional neural network is proposed [15] to predict grasps at every pixel, and this one-to-one mapping from a depth image avoids discrete sampling of grasp candidates as well as long computation time. Mahajan et al. [18] used vector quantized variational autoencoder to enhance the generalization of GG-CNN through semi-supervised feature extraction with limited labelled training data. Kumra et al. introduced residual block and designed the generative residual convolutional neural network (GR-ConvNet) [19], which can generate robust grasps from n-channel input in real time. Yu et al. proposed a two-stream grasping network (TsGNet) by replacing standard convolution with depthwise separable convolution [20]. Besides, a global deconvolution is designed to reduce the number of parameters with better feature expression. After the bounding box and the segmentation mask of the object are obtained by a simultaneous detection and segmentation network, the target object is separated from the background, which is beneficial to avoid the background interference during grasp detection. Nevertheless, it is still challenging that the checkerboard artifacts make the feature maps in decoder distort and thus affect the performance of grasp prediction. In the field of image super-resolution, some researches are conducted to process checkerboard artifacts. Kinoshita and Kiya [25] designed a fixed convolutional layer with an order of smoothness in their convolutional neural network to constrain these artifacts. In [26], Sugawara et al. added the kernel of zero-order hold to compensate the output of up-sampling layers. Reference [21] directly replaces deconvolution using resize-convolution at the cost of extra computational complexity, which is implemented through nearest-neighbor resize followed by convolution. Different from the aforementioned processing, this paper designs a decoder based on twin deconvolution, where a twin deconvolution branch is leveraged to calculate the overlap degree matrix corresponding to the output

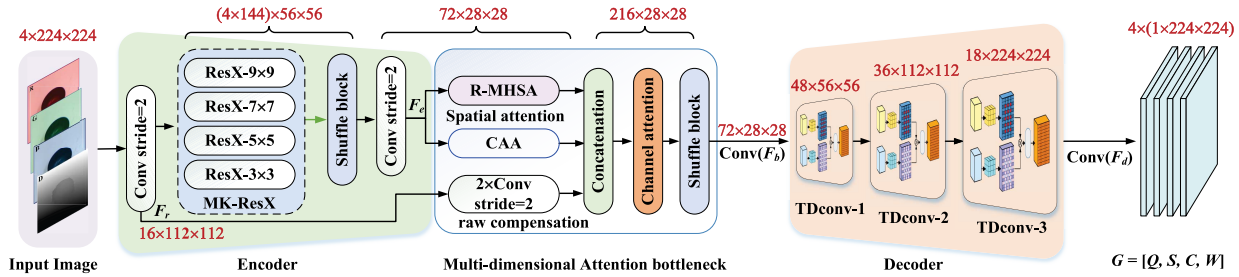

Fig. 1. The framework of the proposed network. Given an RGB-D input image whose size is 4 × 224 × 224, feature is extracted by the encoder. The output feature map $F_e \in \mathbb{R}^{72\times28\times28}$ of the encoder is further refined through a multi-dimensional attention bottleneck, where the outputs from the residual multi-head self-attention (R-MHSA), cross-amplitude attention (CAA), and raw compensation are concatenated in channel, which is then adjusted by the channel attention and a shuffle block for better feature representation $F_b \in \mathbb{R}^{216\times28\times28}$. Followed by a convolution operation, the feature map with the size of 72 × 28 × 28 is fed into the decoder, which adopts three cascaded twin deconvolutions TDconv-1, TDconv-2, and TDconv-3 for grasp prediction G.

of the original transposed convolution branch. By an element-wise division operation between the outputs of original and twin branches, the checkerboard artifacts are eliminated.

III. Method

This section details the proposed pixel-wise grasp detection network with twin deconvolution and multi-dimensional attention, which is termed as TDMAG-Net. Fig. 1 illustrates the framework of the proposed network, which is mainly divided into three modules: encoder, multi-dimensional attention bottleneck, and decoder. The encoder is dedicated to feature extraction of the input image and the feature map F_e is outputted. Then, F_e is refined in the bottleneck module by combining attentions in spatial and channel dimensions for better feature representation F_b. The result after the convolution of F_b is sent to the decoder based on twin deconvolution. Finally, the decoder output is processed by four parallel standard convolutions and the grasp prediction $G = [Q, S, C, W]$ is obtained, where Q, S, C, W correspond to feature maps of grasp quality, sine and cosine related to the grasp angle, and grasp width, respectively.

A. Encoder

It includes a MK-ResX block, a shuffle block [30], and two convolution layers (stride = 2). Inspired by the stronger feature description of ResNeXt block [31] with a fixed kernel size, a multi-kernel ResNeXt block (MK-ResX) is adopted, in which four ResNeXt with kernel sizes of 3 × 3, 5 × 5, 7 × 7, and 9 × 9 are processed in a parallel way. Through the encoder, the input RGB-D data is transformed into the feature map $F_e \in \mathbb{R}^{c\times h\times w}$, where c, h, and w refer to the channel number, height, and width, respectively.

B. Multi-Dimensional Attention Bottleneck

This module is used for adaptive feature refinement based on attention in both spatial and channel dimensions. It is mainly composed of residual multi-head self-attention (R-MHSA), cross-amplitude attention (CAA), raw compensation, a channel attention, and a shuffle block, as illustrated in Fig. 1. The former two blocks constitute spatial attention, where CAA is used to reinforce features by cross attention and amplitude attention, and R-MHSA is presented to enhance feature discrimination. Take the feature map F_e with high-level semantic information as input, these two blocks are complementary to extract more discriminative features. During the forward propagation of features, semantic features are enhanced, but in the meantime the detailed information is gradually inhibited, which impairs the feature completeness. To resolve this drawback, a raw compensation block parallel to R-MHSA and CAA is introduced. By processing the low-level feature map F_r, more detailed information is involved, where F_r is the output of the first convolution (stride = 2) on the original RGB-D image. After that, the channel attention [32] is adopted to capture the relation among different channels for better representation. Similar to encoder, another shuffle block [30] is added after the channel attention to output the feature map F_b.

1) R-MHSA Block: MHSA [33] is popular as it can acquire the correlation between each feature and other features at different spatial positions. This makes the network more focus on the object region by the learned correlation. Attracted by this advantage, MHSA is adopted. Meanwhile, considering that MHSA is sensitive to redundant features, a residual encoder-decoder is additionally introduced prior to MHSA for preprocessing the input feature map F_e. This residual architecture is composed of a standard convolution and a twin deconvolution (please see Section III.C) whose kernel sizes are both set to 3 × 3.

2) CAA Block: It mainly contains a cross attention submodule and an amplitude attention submodule, where the former focuses on spatial relationship from the perspective of column and row, while the latter aims at enhancing detail representation of features. As illustrated in Fig. 2, CAA block takes F_e as its input to infer two 2D attention maps termed as cross attention $A_{cross} \in \mathbb{R}^{1\times h\times w}$ and amplitude attention $A_{amp} \in \mathbb{R}^{1\times h\times w}$. Then A_{cross} and A_{amp} are concatenated in channel and further fused by convolution layers with kernel size 3 × 3. The attention map after fusion is utilized to reweight F_e through element-wise multiplication, and a refined feature map $F_f \in \mathbb{R}^{c\times h\times w}$ is outputted.

The cross attention submodule is designed to capture the correlation of features between column and row corresponding to each spatial position. Both average-pooling and max-pooling are imposed on F_e along height and width dimensions. Unlike regular two-dimensional pooling, our pooling makes an

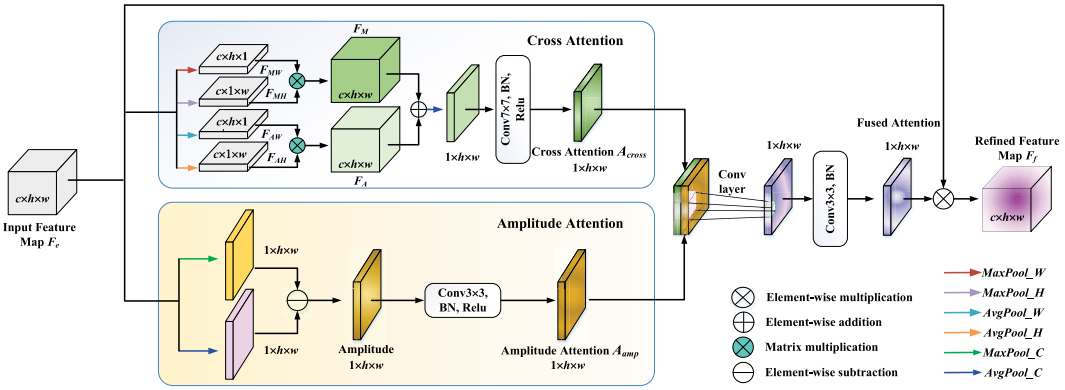

Fig. 2. Structure of cross-amplitude attention (CAA). The input feature F_e is respectively processed by the cross attention and amplitude attention submodules to generate two 2D attention maps A_{cross} and A_{amp}. Then these two attention maps are concatenated in channel and fused by convolution operation. The fused attention map is utilized to reweight the original feature map F_e for a refined feature map $F_f \in \mathbb{R}^{c\times h\times w}$ by an element-wise multiplication.

average/maximum operation among all the feature values in a row or a column, which achieves the mapping from $\mathbb{R}^{c\times h\times w}$ to $\mathbb{R}^{c\times h\times 1}$ or $\mathbb{R}^{c\times 1\times w}$. We label the outputs after average-pooling and max-pooling in height dimension as $F_{AH} \in \mathbb{R}^{c\times 1\times w}$ and $F_{MH} \in \mathbb{R}^{c\times 1\times w}$, respectively. Correspondingly, the outputs in width dimension are denoted with $F_{AW} \in \mathbb{R}^{c\times h\times 1}$ and $F_{MW} \in \mathbb{R}^{c\times h\times 1}$. F_{MW} and F_{MH} are fused by matrix multiplication to produce a feature map $F_M \in \mathbb{R}^{c\times h\times w}$, and similarly, $F_A \in \mathbb{R}^{c\times h\times w}$ is obtained using F_{AW} and F_{AH}. After that, these two feature maps are added and compressed in channel dimension through average-pooling, which is followed by a convolutional layer with batch normalization and ReLU to modulate features for the cross attention A_{cross}.

The amplitude attention submodule focuses on the regions with strong gradient intensity, which tends to contain more detailed information. Herein, the feature amplitude is introduced to express the variation of gradient, which is referred to the feature difference deviating from the average at each spatial position. By max-pooling and average-pooling of F_e, two feature maps containing the maximum value and average value of each position are obtained, respectively. After the calculation of the maximum value minus average value of each position, the feature amplitude is acquired, which is equivalent to element-wise subtraction of corresponding feature maps. Followed by a convolution operation, we get the amplitude attention $A_{amp} \in \mathbb{R}^{1\times h\times w}$.

It is worth mentioning that the cross attention promotes the stability of features, while amplitude attention improves detail representation of features, they are combined in pursuit of better features.

C. Decoder

The decoder attains better up-sampling in consistent with the spatial size of the input image, based on three cascaded twin deconvolutions (TDconv). They have the same structure with different kernel sizes of 3 × 3, 5 × 5, and 7 × 7, and the numbers of their output channels are 48, 36, and 18, respectively.

As mentioned above, the commonly used transposed convolution is thought to result in the checkerboard artifacts due to uneven overlap of convolution results. This problem affects the performance, especially for pixel-wise grasp prediction. Considering that the overlaps at the same position are basically distributed symmetrically on the convolution kernel, for a given position, its each convolution has a similar influence on the final result. By introducing a new transposed convolution based on a kernel with the same weights, the degree of overlap at each position is computed. The resultant overlap degree matrix is then utilized to re-weight the feature map of original transposed convolution through a division operation, which achieves smoothing of the overlapped convolution results. In this way, the checkerboard artifacts are eliminated. Fig. 3 shows the structure of a twin deconvolution.

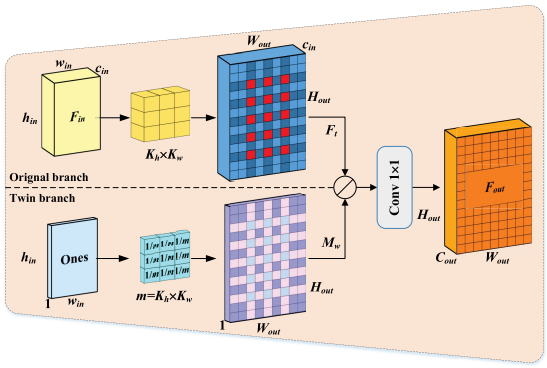

Fig. 3. Structure of a twin deconvolution. $F_{in} \in \mathbb{R}^{c_{in}\times h_{in}\times w_{in}}$ and $F_{out} \in \mathbb{R}^{C_{out}\times H_{out}\times W_{out}}$ denote the input feature map and output feature map, respectively, where c_{in}, h_{in}, w_{in} and $C_{out}, H_{out}, W_{out}$ are the channel number, height, and width corresponding to F_{in} and F_{out}.

Concretely, there are two branches in a twin deconvolution: original branch and twin branch, where the former is a standard transposed convolution and the latter is used to calculate the overlap degree corresponding to the original branch for removing checkerboard artifacts. The input of twin branch is a matrix $Ones \in \mathbb{R}^{1\times h_{in}\times w_{in}}$ with all the entries 1, whose spatial size is the same as that of the input feature map F_{in} of the original branch. Moreover, the kernel of the twin branch has the same spatial size as that of the original branch ($K_h \times K_w$) and its all entries are set to $1/m, m = K_h \times K_w$. With the

transposed convolution in the twin branch, the overlap degree matrix $M_w \in \mathbb{R}^{1\times H_{out}\times W_{out}}$ is computed corresponding to all spatial positions of the output $F_t \in \mathbb{R}^{c_{in}\times H_{out}\times W_{out}}$ from the original branch. Then, an element-wise division operation is performed between each channel of F_t and M_w. Followed by a pointwise convolution Conv1 $\times$ 1, the final output F_{out} of twin deconvolution is obtained.

After the sequential processing of three twin deconvolutions, we acquire the feature map F_d. The result after the convolution of F_d is sent to four parallel convolutions to get grasp prediction $G = [Q, S, C, W]$. Then, the best grasp is computed. The feature maps S and C are employed to avoid the singularity of the grasp angle θ [15], and each element in S and C corresponds to $sin(2\theta)$ and $cos(2\theta)$, respectively. By calculating $arctan(S(p)/C(p))/2$ for the corresponding elements in S and C, we obtain the grasp angle θ at the pixel coordinate p and then a new feature map A of grasp angle is obtained. Q is processed using a filter with Gaussian kernel [15] to ensure stable grasp. On this basis, the best grasp $[p^*, A(p^*), W(p^*)]$ is generated, where $p^* = \underset{p}{\text{argmax}}\, Q(p)$ corresponds to the center coordinate of the best grasp rectangle, $Q(p)$ is the grasp quality at the pixel coordinate p, $A(p^*)$ and $W(p^*)$ refer to the angle and width of the best grasp rectangle, respectively.

IV. Experiments

The proposed TDMAG-Net is verified on the Cornell grasping dataset [11], Jacquard grasping dataset [34], and an actual multi-object scene. The Cornell dataset contains 885 images with 640 $\times$ 480 from 240 real world objects and 8019 manually annotated rectangle-based grasps, where 5110 grasps are positive and others are negative. During the network training, data augmentation [19] with random crops, zooms, and rotations is performed to generate 4425 RGB-D images with 3982 images as training set. The Jacquard dataset contains 54k images of 11k objects and over 1 million grasp labels, where 95% of the data are used as training set.

According to [35], a grasp prediction is considered proper if the difference of orientation angle between the predicted grasp and the ground truth is less than 30°, and the pixel-wise intersection over union (IOU) of the predicted grasp rectangle and its ground truth is over 25%. Then the grasp accuracy is obtained by computing the proportion of proper grasps in all predictions. The evaluation metrics include image-wise accuracy (IW acc.) and object-wise accuracy (OW acc.) [15], where the former evaluates how well the model can generalize to new positions for objects that have been seen previously, and the latter implies the generalization of network for the new unseen objects. Our method runs on NVIDIA GTX1080 GPU and Intel Core i7-7770HQ CPU.

A. Ablation Studies

1) Ablation of TDMAG-Net: To testify the performance of our proposed TDMAG-Net, its eight variants are considered according to whether BasicConv, PResNeXt, MK-ResX, R-MHSA, CAA, raw compensation (RawCom), TransConv, and TDconv are adopted. The BasicConv is an encoder with three cascaded standard convolutions. PResNeXt is constructed by four parallel ResNeXt blocks with the same kernel size 3 $\times$ 3, which is seen as a substitution of MK-ResX. Also, TransConv is a standard decoder with three cascaded transposed convolutions. Table I presents the comparison results of different variants on the Cornell grasping dataset in terms of IW acc. and OW acc. From the results of TDMAG-Net-I and TDMAG-Net-III, one can see that our encoder and decoder achieve improvement in both IW acc. and OW acc. than the standard encoder-decoder scheme, benefiting from better feature representation of MK-ResX with the up-sampling advantage of TDconv. Based on TDMAG-Net-III, TDMAG-Net adds an attention bottleneck involving R-MHSA, CAA, and RawCom. The detection accuracy is further improved because of feature refinement of the bottleneck. For TDMAG-Net-II, it adopts PResNeXt to replace MK-ResX in the encoder. The results indicate that our MK-ResX makes a performance improvement by fusing the information with different scales. TDMAG-Net-IV, TDMAG-Net-V, TDMAG-Net-VI, and TDMAG-Net-VII are designed to validate the detailed blocks of our attention bottleneck module. The performance is increased in sequence by respectively considering RawCom, CAA, and R-MHSA blocks. Besides, the results of TDMAG-Net-VIII with TransConv are lower than those of the proposed TDMAG-Net, which shows that TDconv promotes the detection performance.

Fig. 4 visualizes six feature maps with the size of 28 $\times$ 28 from the outputs of R-MHSA, RawCom, and CAA. It is observed that these three attention blocks focus on different

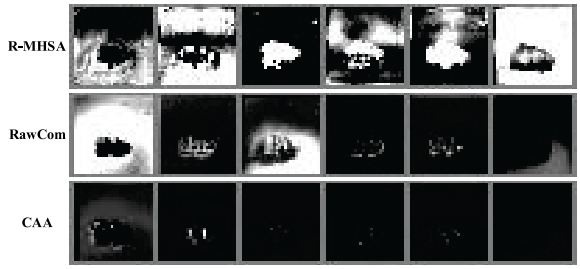

Fig. 4. Visualization of six feature maps (28 $\times$ 28) from the outputs of R-MHSA, RawCom, and CAA.

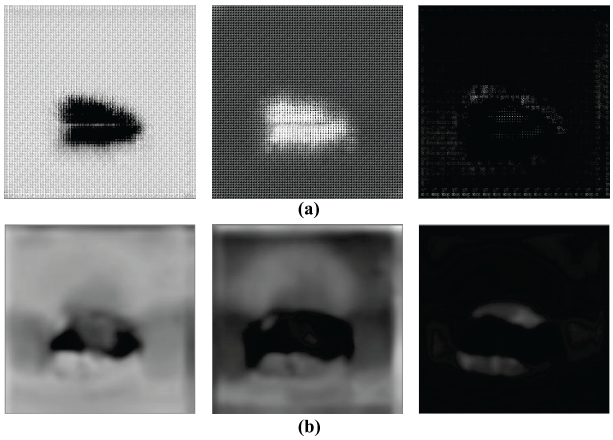

Fig. 5. Visualization of three feature maps from the decoder. (a) TDMAG-Net-I. (b) TDMAG-Net.

TABLE I
COMPARISON OF DIFFERENT VARIANTS OF TDMAG-NET ON THE CORNELL DATASET

Method	BasicConv	PResNeXt	MK-ResX	R-MHSA	CAA	RawCom	TransConv	TDconv	IW acc. (%)	OW acc. (%)
TDMAG-Net-I	√	×	×	×	×	×	√	×	87.64	85.39
TDMAG-Net-II	×	√	×	√	√	√	×	√	97.75	96.63
TDMAG-Net-III	×	×	√	×	×	×	×	√	94.38	92.13
TDMAG-Net-IV	×	×	√	×	×	√	×	√	95.51	93.26
TDMAG-Net-V	×	×	√	√	×	×	×	√	96.63	95.51
TDMAG-Net-VI	×	×	√	×	√	×	×	√	95.51	95.51
TDMAG-Net-VII	×	×	√	√	√	×	×	√	97.75	96.63
TDMAG-Net-VIII	×	×	√	√	√	√	√	×	97.75	95.51
TDMAG-Net	×	×	√	√	√	√	×	√	**98.87**	**97.75**

TABLE II
THE ABLATION OF R-MHSA BLOCK ON THE CORNELL DATASET

Method	Relative position encodings	Absolute position encodings	Residual encoder-decoder	IW acc. (%)	OW acc. (%)
MHSA-I	√	×	×	96.63	95.51
MHSA-II	×	√	×	95.51	94.38
R-MHSA	√	×	√	**98.87**	**97.75**

TABLE III
THE ABLATION OF CAA BLOCK ON THE CORNELL DATASET

Method	Cross attention	Amplitude attention		IW acc. (%)
		Max-pooling	Average-pooling	
CAA-I	√	×	×	96.63
CAA-II	×	√	×	95.51
	×	×	√	94.38
	×	√	√	96.63
CAA	√	√	√	98.87

TABLE IV
COMPARISON OF DIFFERENT METHODS ON THE CORNELL DATASET

Method	Parameters (Approx.)	Input		Accuracy (%)	
		RGB	Depth	IW	OW
SAE-Net [11]	>1050500	√	√	73.9	75.6
AlexNet-based [14]	>7300000	√	—	88.0	87.1
ResNet50-based [17]	>20000000	√	√	89.21	88.96
GR-ConvNet [19]	1900900	√	√	97.7	96.6
ROI-GD [36]	>30000000	√	—	93.6	93.5
Det_Seg w/o Seg [13]	>23000000	√	—	98.2	—
GG-CNN [15]	62420	—	√	73.0	69.0
TsGNet [20]	66754	√	√	93.13	92.99
TDMAG-Net w/RGB	376548	√	—	97.75	95.51
TDMAG-Net	377732	√	√	98.87	97.75

regions of interest and they work together to achieve comprehensive attention. To further demonstrate the twin deconvolutions in the decoder of our TDMAG-Net, the outputted feature map $F_d \in \mathbb{R}^{18\times224\times224}$ by TDconv-3 is also visualized. And the results of TDMAG-Net-I are provided for comparison. 3 out of 18 feature maps from TDMAG-Net-I and TDMAG-Net are presented in Fig. 5. It is observed from Fig. 5(b) that the checkerboard artifacts disappear, which indicates that the introduction of twin branch eliminates the checkerboard artifacts.

2) Ablation of R-MHSA Block: We consider MHSA-I, MHSA-II, and our R-MHSA according to whether relative position encodings, absolute position encodings, and residual encoder-decoder are involved. Relative position encodings reflect the relative relationship among pixels at different positions, while absolute position encodings assign a unique label to each pixel. The results of different settings are shown in Table II. It is seen that MHSA-I is better than MHSA-II, which shows that the relative position encodings are more suitable for the grasp detection task than absolute position encodings. Compared to MHSA-I, R-MHSA performs better, which proves that the residual encoder-decoder enhances the performance through the suppression of redundant features.

3) Ablation of CAA Block: Three settings CAA-I, CAA-II, and CAA are involved in this ablation. CAA-I only considers cross attention, and CAA-II mainly concerns amplitude attention. Besides, CAA-II is subdivided into three cases according to whether max-pooling and average-pooling in the element-wise subtraction operation are involved. Table III describes the results of different settings. For CAA-II, the integration of max-pooling and average-pooling performs better (96.63%) than only using one of them, which proves the validity of the element-wise subtraction. Also, CAA-I attains the accuracy of 96.63%. This implies that the cross attention is comparable to the amplitude attention. With the combination of cross attention and amplitude attention, CAA achieves the best result.

B. Comparison With Existing Methods

In this section, the proposed TDMAG-Net is compared with existing methods including SAE-Net [11], Det_Seg [13], AlexNet-based [14], ResNet50-based [17], ROI-GD [36], GR-ConvNet [19], GG-CNN [15], and TsGNet [20] on the Cornell grasping dataset. The first two methods are categorized into the candidate-based solution, the third to fifth methods belong to the encoder regression, and the latter three ones correspond to pixel-wise regression. Also, TDMAG-Net with only RGB input (TDMAG-Net w/RGB) is considered. Their results are presented in Table IV and the accuracy of the proposed method is good. Comparing the results of TDMAG-Net w/RGB and TDMAG-Net, the introduction of depth map improves the detection accuracy. In addition, the running time including computation of the best grasp is 15 ms, which shows the efficiency of the proposed method.

To further verify the proposed TDMAG-Net, the methods including GR-ConvNet [19], GG-CNN [15], GG-CNN2 [37], ROI-GD [36], Det_Seg [13] are used for comparison on the Jacquard grasping dataset. The former two methods are implemented with RGB-D input based on their open-source codes [38], [39], and the result of ROI-GD is from [13].

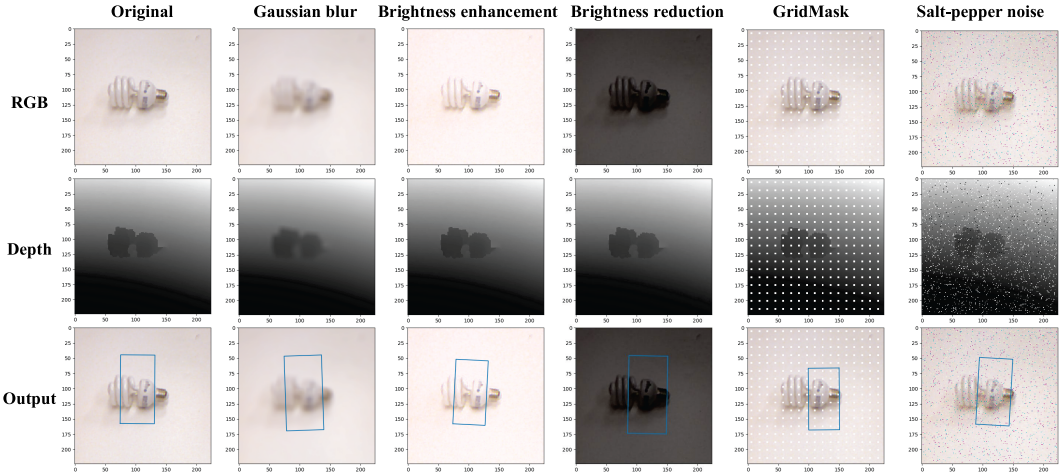

Fig. 6. The grasp detection results of TDMAG-Net with different interferences.

TABLE V
COMPARISON OF DIFFERENT METHODS ON THE JACQUARD DATASET

Method	Parameters (Approx.)	Input		IW Acc. (%)
		RGB	Depth	
GR-ConvNet [19]	1900900	√	√	89.8
GG-CNN [15]	62420	√	√	87.1
GG-CNN2 [37]	66000	—	√	84
ROI-GD [36]	>30000000	√	—	90.4
Det_Seg [13]	>23000000	√	—	92.95
TDMAG-Net	377732	√	√	92.55

The comparison results of different methods are shown in Table V. It is seen that our TDMAG-Net performs well with the help of geometry information of the object from depth map. Combining the results of Tables IV and V, the proposed TDMAG-Net is considered as effective.

C. Robustness Verification

To further verify the proposed method, different interferences are imposed on both original RGB and depth map. The first column of Fig. 6 presents the detection result of original input. The second to sixth columns correspond to the results after Gaussian blur with kernel size 10×10, brightness enhancement (15%), brightness reduction (60%), GridMask with size 3×3, and salt-pepper noise with intensity of 5% are exerted. Despite these interferences, the proposed method still achieves grasp detection.

D. Multi-Object Grasp Detection in an Actual Scene

We apply the proposed method in a realistic multi-object scene. With the Mask R-CNN [40] for object detection and segmentation, the target objects are separated from the background. Also, the background suppression in [20] is borrowed to prevent the disturbance of background by filling specific pixel value into non-target regions. Then the grasp detection is executed based on the RGB and depth images corresponding to each concerned object. Fig. 7(a) gives the experiment scene, where the objects with four categories (bottle, banana, apple, and orange) are concerned. In particular, a banana is placed into a cup and an orange is put on the book. The detection results of concerned objects are shown in Fig. 7(b) and Fig. 7(c) presents the corresponding grasp detection results.

(a) (b) (c)

Fig. 7. The grasp detection on an actual scene. (a) RGB image. (b) Detection results of concerned objects. (c) Grasp detection results.

E. Generality to Saliency Detection

In this section, the proposed grasp detection network is extended to the pixel-wise saliency detection task [41], [42], [43], [44], which also requires up-sampling operation in the decoder with the output of saliency map. Considering that the network needs to reconstruct the contours of the objects instead of learning grasp pose, more feature fusion is added to provide comprehensive features in the decoder. Specifically, the feature maps from the input, intermediate, and output of raw compensation are respectively concatenated with the inputs of twin deconvolutions TDconv-3, TDconv-2, TDconv-1. After the output of TDconv-3 is operated through a convolution followed by a Sigmoid activation function, saliency map of objects can be obtained. In addition, the size of input image is adjusted to 256×256. The network is trained on the MSRA10K dataset [45] with the cross-entropy loss.

Fig. 8 presents the results of different methods including UCF [41], DS [46], ELD [47], and RFCN [48] on the typical images from the ECSSD dataset [49], where the results of the latter three methods are from [41]. One can see from Fig. 8(c) that our method achieves saliency detection. UCF [41] performs better with more details. This is because UCF adopts an effective hybrid up-sampling based on the combination of deconvolution with restricted filter sizes and linear interpolation, which enforces the network to learn accurate

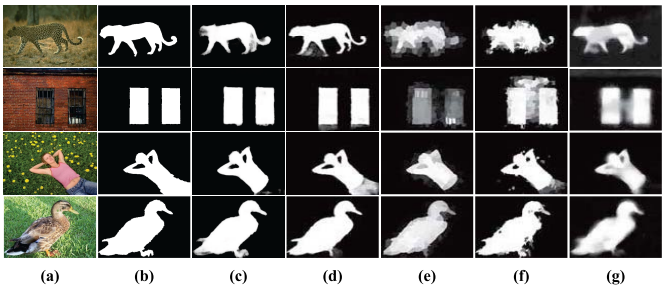

Fig. 8. The results of different saliency detection methods on the typical images from the ECSSD dataset. (a) Original images; (b) Ground truth; (c) Our method; (d) UCF [41]; (e) DS [46]; (f) ELD [47]; (g) RFCN [48].

boundary for saliency detection. Inspired by the UCF model, the idea of boundary recovery shall be considered in our future work, such as introducing pyramid feature fusion or adjusting the weighting proportion of saliency foreground and general background in the training loss.

F. Discussion

The proposed grasp detection method based on twin deconvolution and multi-dimensional attention provides an enhanced up-sampling scheme with adaptable feature refinement. In the transposed convolution, the checkerboard artifacts bring in noise to up-sampling results. To solve this problem, a parallel twin branch is introduced to produce the overlap degree matrix for smoothing the output of the original branch. This solves the checkerboard artifacts. Moreover, a multi-dimensional attention module is designed to adjust the feature map in global, local, and channel dimensions, which provides more discriminative features for the subsequent grasp prediction.

The grasp detection methods including our TDMAG-Net are trained in public grasping datasets. In actual environments, the performance of the grasp detection may be affected. A possible improvement is to take advantage of segmentation to deal with complex cases of multiple objects. Det_Seg [13] has made a beneficial attempt, where semantic segmentation is adopted and combined with grasp candidates to fine-tune grasp detection results. The results show good potential. Inspired by this, instance segmentation is used to generate valuable information about object profile and position, which is fed into twin deconvolutions of our decoder. This provides a promising direction to further improve the performance.

V. Conclusion

In this paper, we proposed a novel pixel-wise grasp detection network with twin deconvolution and multi-dimensional attention. By adding a twin branch upon original transposed convolution branch, twin deconvolution is proposed with better up-sampling performance, and the problem of checkerboard artifacts is solved. Moreover, multi-dimensional attention is presented to fine-tune the features from the dimensions of spatial and channel. With cross-amplitude attention and residual multi-head self-attention, the intrinsic relationship of features is mined and discrimination of features is enhanced. Experimental results show the effectiveness of the proposed method. In the near future, the proposed method shall be combined with instance segmentation in actual robot system for better grasp.

References

[1] K. Ehsani et al., "ManipulaTHOR: A framework for visual object manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4497–4506.
[2] T. Mu et al., "ManiSkill: Generalizable manipulation skill benchmark with large-scale demonstrations," 2021, *arXiv:2107.14483*.
[3] P. M. Scheikl et al., "Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery," *IEEE Robot. Autom. Lett.*, vol. 8, no. 2, pp. 560–567, Feb. 2023.
[4] Z. Tu, C. Yang, X. Wu, Y. Zhu, W. Wu, and N. Jia, "Moving object flexible grasping based on deep reinforcement learning," in *Proc. Int. Conf. Control, Autom. Robot.*, 2022, pp. 34–39.
[5] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied AI: From simulators to research tasks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 2, pp. 230–244, Apr. 2022.
[6] G. Ren et al., "A fast search algorithm based on image pyramid for robotic grasping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 6520–6525.
[7] A. Herzog, P. Pastor, M. Kalakrishnan, L. Righetti, T. Asfour, and S. Schaal, "Template-based learning of grasp selection," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 2379–2384.
[8] F. T. Pokorny, Y. Bekiroglu, and D. Kragic, "Grasp moduli spaces and spherical harmonics," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 389–396.
[9] X. Chen, H. Li, Q. Wu, K. N. Ngan, and L. Xu, "High-quality R-CNN object detection using multi-path detection calibration network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 715–727, Feb. 2021.
[10] R. Xia, G. Li, Z. Huang, H. Meng, and Y. Pang, "CBASH: Combined backbone and advanced selection heads with object semantic proposals for weakly supervised object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6502–6514, Oct. 2022.
[11] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, nos. 4–5, pp. 705–724, 2015.
[12] J. Mahler et al., "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017, *arXiv:1703.09312*.
[13] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13452–13458.
[14] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1316–1322.
[15] D. Morrison et al., "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," in *Proc. Robot., Sci. Syst.*, 2018.
[16] Q. Zhang et al., "Robust robot grasp detection in multimodal fusion," in *Proc. MATEC Web Conf.*, vol. 139, 2017.
[17] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 769–776.
[18] M. Mahajan, T. Bhattacharjee, A. Krishnan, P. Shukla, and G. C. Nandi, "Robotic grasp detection by learning representation in a vector quantized manifold," in *Proc. Int. Conf. Signal Process. Commun. (SPCOM)*, Jul. 2020, pp. 1–5.
[19] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 9626–9633.
[20] Y. Yu, Z. Cao, Z. Liu, W. Geng, J. Yu, and W. Zhang, "A two-stream CNN with simultaneous detection and segmentation for robotic grasping," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 2, pp. 1167–1181, Feb. 2022.
[21] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, Oct. 2016.
[22] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2222–2234.
[23] Y. Ou, Z. Chen, and F. Wu, "Multimodal local-global attention network for affective video content analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1901–1914, May 2021.

[24] G. Gao, W. Zhao, Q. Liu, and Y. Wang, "Co-saliency detection with co-attention fully convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 877–889, Mar. 2021.

[25] Y. Kinoshita and H. Kiya, "Fixed smooth convolutional layer for avoiding checkerboard artifacts in CNNs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3712–3716.

[26] Y. Sugawara, S. Shiota, and H. Kiya, "Super-resolution using convolutional neural networks without any checkerboard artifacts," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 66–70.

[27] Z. Chu, M. Hu, and X. Chen, "Robotic grasp detection using a novel two-stage approach," 2020, *arXiv:2011.14123*.

[28] M. Vohra, R. Prakash, and L. Behera, "Real-time grasp pose estimation for novel objects in densely cluttered environment," in *Proc. 28th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Oct. 2019, pp. 1–6.

[29] S. V. Pharswan, M. Vohra, A. Kumar, and L. Behera, "Domain-independent unsupervised detection of grasp regions to grasp novel objects," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 640–645.

[30] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[31] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[32] S. Woo et al., "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[33] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16514–16524.

[34] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 3511–3516.

[35] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3304–3311.

[36] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "ROI-based robotic grasp detection for object overlapping scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4768–4775.

[37] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robot. Res.*, vol. 39, nos. 2–3, pp. 183–201, Mar. 2020.

[38] (2021). *Robotic-Grasping*. [Online]. Available: https://github.com/skumra/robotic-grasping

[39] (2020). *GGCNN*. [Online]. Available: https://github.com/dougsm/ggcnn

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[41] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 212–221.

[42] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[43] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[44] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1755–1769, Jul. 2020.

[45] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[46] X. Li et al., "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[47] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 660–668.

[48] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[49] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

Guangli Ren received the B.S. degree in intelligent science and technology from Dalian Maritime University, Dalian, China, in 2015, and the M.S. degree in technology of computer application from the Capital Normal University, Beijing, China, in 2018. He is currently pursuing the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include visual SLAM and robotic manipulation.

Wengjie Geng received the B.S. and M.S. degrees from Harbin Engineering University, Harbin, China, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include visual perception and robotic grasping.

Peiyu Guan received the B.E. degree in electronic information science and technology from Jilin University, Changchun, China, in 2017, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2022. She is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. Her research interests include service robot.

Zhiqiang Cao (Senior Member, IEEE) received the B.E. degree in industrial automation and M.E. degree in control theory and control engineering from the Shandong University of Technology, Jinan, China, in 1996 and 1999, respectively, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2002. He is currently a Professor with the Institute of Automation, Chinese Academy of Sciences. His research interests include service robots and intelligent robot.

Junzhi Yu (Fellow, IEEE) received the B.E. degree in safety engineering and the M.E. degree in precision instruments and mechanology from the North University of China, Taiyuan, China, in 1998 and 2001, respectively, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003. From 2004 to 2006, he was a Post-Doctoral Research Fellow with the Center for Systems and Control, Peking University. He was an Associate Professor with the Institute of Automation, Chinese Academy of Sciences, in 2006, where he was a Full Professor in 2012. In 2018, he joined the College of Engineering, Peking University, as a Tenured Full Professor. His current research interests include intelligent robots, motion control, and intelligent mechatronic systems.