AAGDN: Attention-Augmented Grasp Detection Network Based on Coordinate Attention and Effective Feature Fusion Method

Zhenning Zhou, Xiaoxiao Zhu, and Qixin Cao

***Abstract*—High-precision robotic grasping is necessary for extensive grasping applications in the future. Most previous grasp detection methods fail to pay enough attention to learn grasp-related features and the detection accuracy is limited. In this letter, a novel attention-augmented grasp detection network (AAGDN) is presented to generate accurate grasp poses for unknown objects. The proposed AAGDN has three elaborate designs making it achieve higher accuracy than existing methods. First, we construct a coordinate attention residual module to extract positional information and improve the spatial sensitivity of features. Then, we propose an effective feature fusion module to bridge the resolution and semantic gaps of different-level features and obtain efficient feature representations. Lastly, a feature augmentation pyramid module is developed to enhance grasp-related features as needed and reduce the loss of information. Extensive experiments on three public datasets and various real-world scenes prove that the proposed AAGDN achieves better performance than current methods. Our model obtains the state-of-the-art 99.3% and 96.2% grasp detection accuracy on the Cornell and Jacquard dataset, respectively. Moreover, in physical grasping experiments, the AAGDN attains the 94.6% success rate for unseen objects in cluttered scenes, which further demonstrates the accuracy and robustness of our method to grasp novel objects.**

***Index Terms*—Grasping, deep learning in grasping and manipulation, grasp detection, attention mechanism.**

I. Introduction

With the advantages of high automation and intelligence, robotic grasping has been widely studied for years. However, high-precision robotic grasping is still difficult because of the uncertainty about the object's shape, size, and pose, and it remains an open problem. The main challenge is to learn and build accurate intrinsic relationships between the visual information of input image and the optimal grasp pose, which requires the model to extract the grasp-related features precisely and transfer the learned knowledge to novel objects. It is worth noting that the grasping position related to the spatial features of the object represents the most critical grasp configuration, determines the grasp quality and therefore needs more attention in grasping tasks.

Researchers have made robots more intelligent in grasping tasks through many studies [1]. Force-closure-based methods have been proven reliable for grasping objects [2], but they need necessary searches for candidate grasps. The robotic grasp detection method is proposed to avoid this problem [3]. The grasp detection network can generate the optimal grasp pose and suitable grasp width without any request for large-scale sampling [4]. The grasp representation method that adopts rectangles to represent the feasible grasp area is first presented in [5]. Based on this method, researchers develop many deep learning-based methods. Among them, the detection-based methods regard this grasping task as a segmentation problem about the optimal grasp area and show better performance. The detection-based methods first generate the grasp heatmaps from the n-channel input, then employ the pixel having the maximum grasp success rate from the quality heatmap to be the final grasping position. Many studies have achieved great results with various network structures. The researchers in [6] design a grasp model by referring to the ideas of Fast RCNN [7]. The Generative Grasping CNN structure is proposed in [8] to achieve one-to-one mapping between the grasp pose and the depth image. Kumra et al. [9] propose the generative convolutional network to predict antipodal grasp configurations. However, the accuracy of these methods for novel objects needs to be further improved.

Moreover, in the latest studies [4], [10], [11], the attention mechanism has shown the powerful effect on helping models focus on the grasp-related features. However, there remains some problems that limit the performance. Current methods only encode the grasp semantic information but fail to extract the rich spatial information fully, so the model is unable to focus on the grasping-position features and achieve high grasp quality. In addition, existing studies generally adopt direct concatenation or addition operations to combine the different-level features. Nevertheless, the low-level features contain abundant spatial details while high-level features are rich in semantic information, the simple aggregation may weaken the effectiveness of information propagation and decrease the accuracy of grasping. Furthermore, some features are gradually ignored or lost during encoding, which causes some grasp-related information can not be restored when upsampling and limits the grasp detection performance.

To overcome these problems, we propose the AAGDN. First, we construct the CoA-ResNet to extract positional information and enhance spatial sensitivity of features. Second, we propose an Effective Feature Fusion Module (EFFM) to weight the importance of different-level features at each pixel when fusing them. Third, we develop the Feature Augmentation Pyramidal Module (FAPM) to enhance significant grasp-related features

Manuscript received 20 December 2022; accepted 6 April 2023. Date of publication 20 April 2023; date of current version 28 April 2023. This letter was recommended for publication by Associate Editor A. Morales and Editor H. Liu upon evaluation of the reviewers' comments. This work was supported by the National Key R&D Program of China under Grant 2018AAA0102700. *(Corresponding author: Qixin Cao.)*

The authors are with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhenning_zhou@sjtu.edu.cn; ttl@sjtu.edu.cn; qxcao@sjtu.edu.cn).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2023.3268596, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3268596

that are perhaps ignored and lost with the deepening of the network. The proposed AAGDN outperforms the latest methods and behaves competitive to the state-of-the-art. We visualize the comparison results, investigate the specific effectiveness of proposed modules, and conduct ablation studies. Finally, extensive grasping experiments are performed for various objects, including household, 3D-printed and metal objects.

The contributions can be summarized in four folds:

1) We construct a novel coordinate attention residual module, which improves the grasping-position attention and spatial sensitivity of features, assisting the grasp model to more accurately locate the optimal grasp position.
2) An effective feature fusion module is proposed to conduct adaptive response of multi-level features when fusing them, which bridges the representation gaps of features and ensures efficient information propagation in the combination phase.
3) We develop a new feature augmentation pyramid module to enhance grasp-related features that may be neglected and lost during encoding and refine the grasp performance.
4) The grasp network AAGDN with our proposed modules achieves the state-of-the-art performance on public datasets and real-world robotic grasping experiments.

II. Related Works

A. Grasp Detection With Deep Learning

Deep learning has been widely used in grasp detection and the ongoing studies can be divided into regression-based, classification-based and detection-based methods.

Regression-based methods mean that the models predict grasps for the given objects directly. Redmon et al. [12] propose the single-stage regression method to recognize the object and find good grasp rectangles. Approaches like [6], [13] detect the location of grasps based on Region of Interest (ROI). The neural network is first employed by Lenz et al. [5] as a classifier to detect grasps. Researchers in [14] train a neural network for classifying the possible grasps. Chu et al. [6] predict multiple grasp candidates by defining this problem as classification with null hypothesis competition. Recently, an orientation anchor box mechanism [15] is proposed to generate grasp angles based on the predefined oriented prior rectangles. The detection-based methods consider this task as a segmentation problem about the optimal grasp area. Methods [6], [16] refer to the key ideas from the object detection field to develop detectors. Guo et al. [17] design a hybrid model to improve the grasp stability with tactile information. In [8], researchers propose a generative convolutional neural network to predict the grasp pose and quality score for each pixel. Kumra et al. [9] achieve robust antipodal robotic grasps for unknown objects. Ainetter et al. [18] demonstrate that the semantic segmentation can assign grasp candidates to object classes. In sum, the regression-based and classification-based methods are direct and effective, but the detection-based methods have a more explicit meaning and perform better in comparison.

B. Attention Mechanisms for Grasp Detection

The attention mechanism is first proposed in [19]. The latest studies in grasp detection prove that the attention block can assist the model to focus on the grasp-related features as needed. The powerful transformer with the self-attention is adopted in [10] to learn global feature relevance, but this may ignore the details of local features and require a high cost in computation time. The research [4], [11] introduce the squeeze-and-excitation attention into the grasp model to give different weights for the features and obtain high accuracy. However, these works just encode the grasp semantic information and reweight the importance of channels, but pay less attention to the rich spatial features in each channel. This way, the model may easily lose some spatial information and fail to focus on the grasping-position features, which may lead to poor results. We fill this gap with the developed CoA-ResNet module.

C. Multi-Level Feature Fusion Methods

Feature fusion is widely adopted to integrate the features of different levels and obtain multi-level feature representation. The most common fusion method is to upsample the high-level features and then concatenate or add the upsampled features and low-level features together. Besides, some approaches have been presented to improve the feature fusion effectiveness. Yu et al. [20] apply an extra branch to protect spatial information while also producing features with high resolutions. Zhao et al. [21] propose a Cascade Feature Fusion module to integrate multi-resolution features under proper label guidance. Li et al. [22] implement cross-level feature aggregation by reusing the high-dimensional information. However, these methods ignore the semantic and resolution gaps between different-level features. Authors in [23] propose to fuse multi-level features using gates selectively, but with high computational cost. By considering these, we develop the EFFM to integrate multi-level features by applying the attention fusion map, which can improve the effectiveness of grasp-related information in the combination phase.

III. Problem Statement

The problem of robotic grasping is described as generating the optimal grasp pose and performing it on the robot. Different from the 5-dimensional grasp rectangle proposed in [12], an improved grasp representation similar to the one developed in [24] is adopted in this letter, which can be denoted as:

$$\boldsymbol{G_r} = \{\boldsymbol{p_r}, \boldsymbol{\theta_r}, \boldsymbol{w_r}, \boldsymbol{q}\} \tag{1}$$

where $\boldsymbol{p_r} = (x, y, z)$ is the center position of the gripper's tip in Cartesian coordinates, $\boldsymbol{\theta_r}$ represents the gripper orientation angle around the Z axis, and $\boldsymbol{w_r}$ is the required opening width for gripper. The grasping quality score $\boldsymbol{q}$ is added to predict the probability of grasp success, and the best grasping configuration can be acquired by $\boldsymbol{G}^* = \arg\max_q \boldsymbol{G}_r$.

In the image space, a grasp is described as:

$$\boldsymbol{G_i} = \{\boldsymbol{p_i}, \boldsymbol{\theta_i}, \boldsymbol{w_i}, \boldsymbol{q}\} \tag{2}$$

where $\boldsymbol{p}_\mathbf{i} = (u_\mathbf{i}, v_\mathbf{i})$ is the grasp center, $\boldsymbol{w}_\mathbf{i}$ is the required grasp width with the range of [0, $\mathrm{W}_{\mathrm{max}}$]. $\boldsymbol{\theta}_\mathbf{i}$ indicates the grasp angle at position $\boldsymbol{i}$ and takes value in $[-\frac{\pi}{2}, \frac{\pi}{2}]$. $\boldsymbol{q}$ denotes the grasp quality at every pixel with the range of [0, 1].

The generated heatmaps in the image space are described with grasp angle, width and quality, as shown in (3). Each pixel in three images $\mathbf{\Theta}$, $\mathbf{W}$ and $\mathbf{Q}$ respectively denotes the corresponding value calculated using (2).

$$\mathbf{G} = (\mathbf{\Theta}, \mathbf{W}, \mathbf{Q}) \in \mathbb{R}^{3\times h\times w} \tag{3}$$

With the generated grasp in the image space, the robot can execute grasping by applying the transformation from the image coordinate to the robot reference frame as follows:

$$G_{\boldsymbol{r}} = \boldsymbol{T_{rc}}(\boldsymbol{T_{ci}}(G_i)) \tag{4}$$

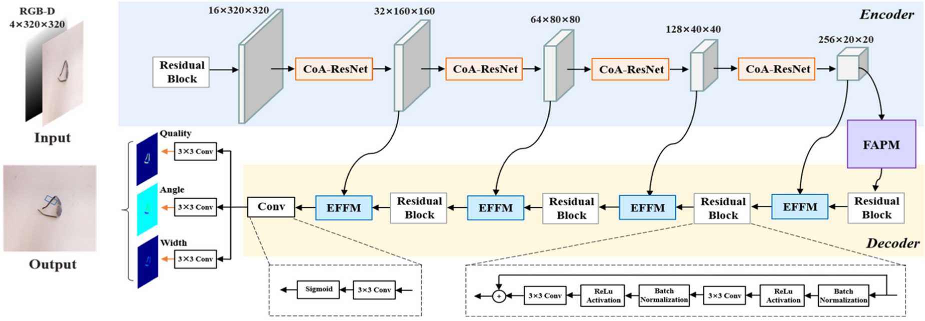

Fig. 1. The overall architecture of AAGDN. The coordinate attention residual modules (CoA-ResNet) in the first row gradually encode the feature maps to capture high-level semantic information, and act as the feature encoder. The feature augmentation pyramid module (FAPM) is responsible for enhancing critical features and reducing the loss of information. The effective feature fusion modules (EFFM) in the last row decode the feature maps to generate the heatmaps with the same resolution as the input, and can guarantee complete and efficient information propagation by considering the importance of different-level features during fusion.

where, $\boldsymbol{T_{ci}}$ is the transform matrices of the 2-D image space to the 3-D camera space, $\boldsymbol{T_{rc}}$ converts the camera frame to the robot frame with the camera calibration results.

IV. Method

A. Network Architecture

The overall network architecture of AAGDN is shown in Fig. 1, which is mainly composed of the CoA-ResNet, EFFM and FAPM. The grasp detection network takes a 4-channel image as input and generates pixel-wise grasp poses by grasp quality, angle and width heatmaps. Firstly, given the input RGB-D image with the size of $4 \times 320 \times 320$, we feed it into the encoder to gradually extract semantic information and encode it into high-level features to generate heatmaps, where the semantic information refers to the valuable features contained in the image and the information that they reflect about the properties the object to be grasped. Afterwards, we send the extracted features into the FAPM to reduce the loss of features and enhance important features. Then, the decoder with the EFFM is employed to fuse multi-level features and produce grasp configurations for every pixel. EFFM consists of two inputs: the output F_{high} of the previous layer and the feature F_{low} that have the same-number channels from encoder. Each EFFM is followed by a residual block to decrease the channels. One convolution layer with the kernel scale of 1×1 is added after the whole upsampling process. The final network layer contains 4 task-specific convolutional filters $(f_{conv}^{q}, f_{conv}^{\cos(2\theta)}, f_{conv}^{\sin(2\theta)}, f_{conv}^{w})$ with the kernel size of 3×3, and the output grasp of the whole model can be depicted as:

$$g_i = \max_{q} f_{conv}^{i}\left(x\right), i = q, \cos(2\theta), \sin(2\theta), w \qquad (5)$$

where x is the feature from the last layer of decoder.

It is worth noting that we adopt $\cos(2\theta)$ and $\sin(2\theta)$ heatmaps to represent the grasping angle θ, which can eliminate discontinuity around $\pm\frac{\pi}{2}$ and maintain a unique mapping between θ and $[-\frac{\pi}{2}, \frac{\pi}{2}]$, and the final grasping angle can be calculated by $\theta = (\arctan(\sin 2\theta / \cos 2\theta))/2$. Based on the optimal grasp position in the quality heatmap, we can further obtain the ultimate grasp angle and grasp width from $\mathbf{\Theta}, \mathbf{W}$ heatmaps.

B. CoA-ResNet: Coordinate Attention Residual Network

Recent works have shown that the employed attention could significantly enhance the capabilities of grasp detection networks [4], [11]. However, existing studies only consider reweighing the importance of each channel but the rich position information in each channel is generally neglected. To address this problem, we propose to introduce the coordinate attention [25] to this task and construct the CoA-ResNet to extract effective position information and generate attention weights, which reflect the importance of the corresponding position in feature maps. This approach can enhance the spatial expressive ability of the learned features and help our model more accurately locate the optimal grasp position.

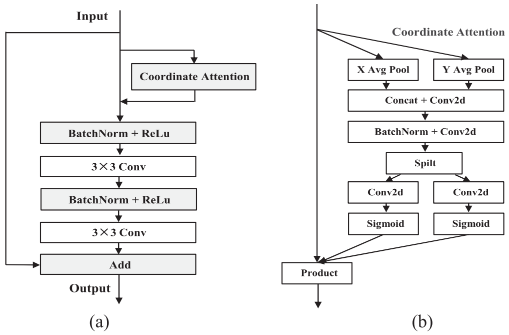

Fig. 2. The proposed CoA-ResNet module. (a) The detailed structure of CoA-ResNet. (b) The illustration of coordinate attention block.

Fig. 2(a) illustrates the structure of CoA-ResNet. The first component of CoA-ResNet is the coordinate attention block. The second part is a full pre-activation residual unit suggested in [26]. The detailed structure of coordinate attention block is given in Fig. 2(b), which contains two stages: coordinate information embedding and generation. By these two stages, we encode the position information of the image into the feature maps to enhance the grasping-position attention and spatial sensitivity of features so that the feasible grasp regions can be accurately highlighted. Concretely, given the input feature $\mathbf{X}$, two pooling kernels are first employed to encode features along the horizontal and vertical coordinates, which yields a pair of direction-aware feature maps. The feature map at c-th channel with the vertical coordinate h is as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \qquad (6)$$

Similarly, the feature map at c-th channel with the horizontal coordinate w is written as:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(w, i) \qquad (7)$$

With aggregated position information, the feature maps $z_c^h(h)$ and $z_c^w(w)$ are concatenated and then fed into a shared 1×1 convolution for coordinate attention generation. Then, we spilt

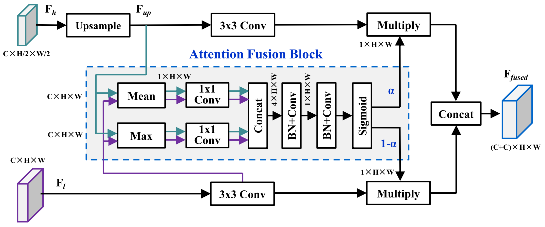

Fig. 3. The structure of Effective feature fusion module. EFFM first uses the attention fusion block to produce the fusion weight map, and then fuses the input multi-level features by element-wise product and concatenation.

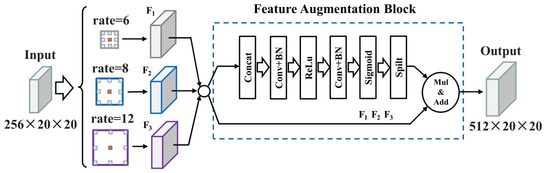

Fig. 4. The structure of feature augmentation pyramidal module. FAPM refines the grasp detection performance by expanding the receptive fields and calculating the weighted sum with the feature augmentation block.

the intermediate feature map $\boldsymbol{f}$ along the spatial direction into two tensors $\boldsymbol{f}^w$ and $\boldsymbol{f}^h$, and another two 1×1 convolutions F_h and F_w are applied to separately transform $\boldsymbol{f}^w$ and $\boldsymbol{f}^h$ to tensors with the same channel numbers as input $\mathbf{X}$:

$$\mathbf{g}^i = \sigma(F_i(\boldsymbol{f}^i)), i = h, w \tag{8}$$

where σ denotes the sigmoid function.

Finally, the direction-aware tensor $\mathbf{g}^i$ is expanded and the output feature $\mathbf{Y}$ of coordinate attention block can be denoted as:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{9}$$

C. Effective Feature Fusion Module

Fusing Multi-level features can enrich the feature representation and increase the ability of network to understand grasp semantic information. Most grasp detection networks adopt a direct concatenation or addition operation to combine the multi-level features. However, the low-level features contain abundant spatial details while high-level features are rich in semantic information. Simply aggregating would weaken the effectiveness and cause the loss of important features, making the detector unable to focus on the most relevant grasping features. To conquer this issue, we propose an effective feature fusion module by applying the attention fusion map, which permits each pixel to select separate information based on the inputs.

The detailed structure of the proposed EFFM is illustrated in Fig. 3. Given the feature maps F_{high} and F_{low}, we first adopt bilinear interpolation for upsampling F_{high} to the same size as F_{low}. Afterwards, we utilize the channel-wise mean and maximum pooling to aggregate information, and two 1×1 refining convolutions are employed to encode the aggregated features. Then, the output four feature maps are concatenated together and fed into two 7×7 convolutional layers (with BN) followed by a sigmoid function to predict the fusion weight map $\alpha \in R^{1 \times H \times W}$, where each element represents the importance of the corresponding pixel in feature maps contributing to the final information propagation. In this case, the attention fusion block can focus on the pixels that contain important semantic information in the high-level features and critical spatial detail information in the low-level features. The formulation of the attention fusion block can be depicted as:

$$F_{cat} = Concat(Mean(F_{up}), Max(F_{up}),\ Mean(F_{low}), Max(F_{low})) \tag{10}$$

$$\alpha = Sigmoid(Conv(F_{cat})) \tag{11}$$

With the generated two weight maps, we further employ the element-wise product between fusion masks and the input feature maps. Finally, we perform concatenation for the attention-weighted features and outputs the fused feature:

$$F_{fused} = Concat(f_{bilinear-interpolation}(F_{up}) \cdot \alpha, F_{low} \cdot (1 - \alpha)) \tag{12}$$

Compared to prior methods, the EFFM applies the weight map for different-level features to focus on the most relevant features, which bridges the representation gaps when fusing them and enables effective information propagation.

D. Feature Augmentation Pyramidal Module

As the encoder network deepens, some features may be ignored and gradually lost during encoding, which causes some critical grasp-related information can not to be restored when upsampling in the decoder and limits the quality of grasp detection, especially for small-size objects. To address this issue, a feature augmentation pyramidal module is developed in this letter to act as the bridge between the encoder and decoder to enhance important information as needed and refine the grasp detection performance.

As illustrated in Fig. 4, the input feature is first fed to three parallel atrous convolutions with different rates. Then, these extracted multi-scale features are fed into the feature augmentation block to generate weights for each input. Specifically, we use concatenation, convolution and activation functions for these features and then effectively aggregate them to the output by calculating the weighted sum. In this letter, we adopt the 3×3 atrous convolution with rate parameters from [6], [12], [18] to generate the multi-scale context information. With the FAPM, we ensure the integrity of spatial and contextual information, reduce the loss of detailed features and enhance grasp-related information.

E. Loss Function

We aim to approximate the mapping from input images $I = \{I_1, \ldots, I_n\}$ to grasp prediction $\hat{G}$. For a dataset consisting of objects and grasp label $L = \{L_1, \ldots, L_n\}$, we would like to minimize the difference between the generated heatmaps $\hat{G}$ and the ground truth L. We adopt the Smooth L1 loss as the loss function, which is robust and stable for training our model. Therefore, the entire loss function *Loss* can be expressed as:

$$\begin{aligned} &Loss(\hat{G}, L) = \sum_i^N \sum\nolimits_m Loss_{smothL1}\left(\hat{G}_i^m - L_i^m\right) \\ &m \in \{Q, W, \sin(2\Theta), \cos(2\Theta)\} \end{aligned} \tag{13}$$

with$Loss_{smothL1}$:

$$Loss_{smothL1} = \begin{cases} 0.5\left(\hat{G}_i^m - L_i^m\right)^2, & \text{if } \left|\hat{G}_i^m - L_i^m\right| < 1 \\ \left|\hat{G}_i^m - L_i^m\right| - 0.5/\sigma^2, & \text{otherwise} \end{cases} \tag{14}$$

where N is the number of samples, σ is the hyperparameter controlling the smooth area and we set it to 1 in our work.

V. Experiments

We conduct abundant experiments for evaluating the performance of AAGDN. Specifically, we test the AAGDN on two single-object datasets [3], [27] and one multi-object dataset [6], perform comparison studies with the state-of-the-art methods, study the contributions of proposed modules, and then validate the physical grasping capability on a Franka robot.

A. Datasets and Experimental Setup

1) Dataset: We adopt the Cornell [3] and Jacquard [27] grasp datasets for training and testing our model. The Cornell dataset consists of 885 images of 224 objects, and we employ online data augment methods to extend the dataset. The Jacquard Dataset contains 54000 images of 11000 objects, which is sufficient for the training process. Both of them are divided into the 80% training set and 20% testing set, 80% of the training set is adopted for training, and another 20% is used for validating. Moreover, to emphasize the highest confidence in the center grasp position, we adopt the Gaussian-based grasp represen-tation method [28] to encode the ground-truth grasps.

2) Network Training: The presented AAGDN is implemented using PyTorch 1.14 and CUDA 10.1, and the whole system is employed on Ubuntu 18.04. The number of network parameters is not large, and there are no high requirements for computational cost. We train the AAGDN on an NVIDIA RTX 2080Ti GPU with an Adam optimizer. The batch size is 8, and the learning rate is set to decrease as the training progress with an initial value of 0.001, where these parameters have been proved as the optimal settings by extensive tests. Before training, the input image is normalized by subtracting its mean and dividing the standard deviation. At training each step, a batch of images from the training set is randomly sampled, and we adopt the ground truth as the target values for heatmaps to train our model.

3) Evaluation Metric: We adopt the widely used Jaccard index and Angle threshold as the evaluation metric for fair comparisons with other methods. A grasp is considered to be correct if it satisfies the following conditions:

a) The orientation angle difference between the generated grasp rectangle and the ground truth is less than 30°.

b) The intersection over union (IoU) score between the generated grasp $\boldsymbol{G}_i$ and the ground truth $\boldsymbol{G}_{gt}$ is larger than 0.25, which means:

$$\mathrm{IoU} = \frac{|\boldsymbol{G}_{\mathrm{i}} \cap \boldsymbol{G}_{\mathrm{gt}}|}{|\boldsymbol{G}_{\mathrm{i}} \cup \boldsymbol{G}_{\mathrm{gt}}|} > 0.25 \tag{15}$$

B. Experiments and Analysis on Cornell Dataset

The comparison studies are conducted with the well-known GR-ConvNet [9], the latest Efficient Grasp [28] and the state-of-the-art SE-ResUNet [4] on the Cornell grasp dataset to visualize the advantages of our method. Fig. 5(a) describes the grasp detection results for unseen objects. The top row shows the generated grasp rectangles, and the other three rows are the grasp quality, angle and width heatmaps, respectively.

It can be seen that compared with other methods, the proposed AAGDN can ignore irrelevant features and more precisely detect the optimal grasp area, especially when the grasp position is challenging. Moreover, the generated grasp configurations by AAGDN such as the grasp position and width are more appropriate. For example, for a pair of scissors in Fig. 5, our approach predicts the middle position as the grasp point, and for a small-size object, our model can also generate suitable grasp width by considering the narrow space, as shown in the eighth column of Fig. 5(a). The comparison results show the AAGDN knows where to grasp and how to adjust the grasp width for different objects better.

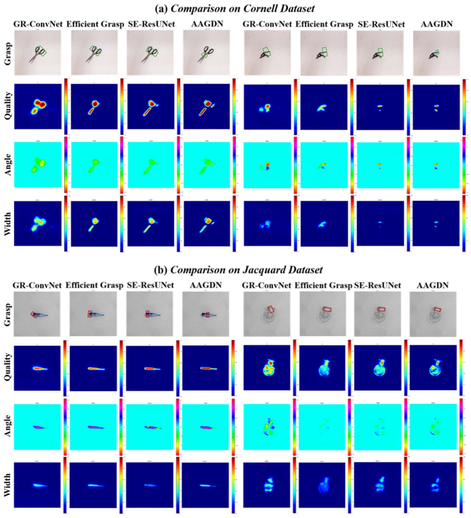

Fig. 5. Comparison results on the Cornell and Jacquard datasets. (a) Comparison on Cornell dataset. (b) Comparison on Jacquard dataset.

TABLE I
The Accuracy Comparison on Cornell Dataset

Method	Year	Accuracy (%)		Time (ms)
		IW	OW	
SAE [5]	2015	73.9	75.6	1350
GG-CNN [8]	2018	73.0	69.0	19
AlexNet, MultiGrasp [12]	2015	88.0	87.1	76
GR-ConvNet [9]	2020	97.7	96.6	20
RSEN [11]	2021	96.4	-	5
Det_seg [18]	2021	98.2	-	63
TF-Grasp [10]	2022	98.0	96.7	42
Efficient Grasp [28]	2022	97.8	-	6
SE-ResUNet [4]	2022	98.2	97.1	25
AAGDN	**2023**	**99.3**	**98.8**	**18**

TABLE II
The Accuracy Comparison on Jacquard Dataset

Authors	Method	Year	Accuracy (%)
Depierre et al. [27]	Jacquard	2018	74.2
Zhou et al. [16]	FCGN, ResNet-101	2018	91.8
Morrison et al. [8]	GG-CNN	2018	84
Kumra et al. [9]	GR-ConvNet	2020	94.6
Cao et al. [11]	RSEN	2021	94.8
Ainetter et al. [18]	Det_seg_Refine	2021	92.95
Wang et al. [10]	TF-Grasp	2022	94.6
Cao et al. [28]	Efficient Grasp	2022	95.6
Yu et al. [4]	SE-ResUNet	2022	95.7
Ours	**AAGDN**	**2023**	**96.2**

In addition, the proposed AAGDN is compared with several representative methods in recent years under the same metric. We adopt a cross-validation setup by image-wise (IW) and object-wise (OW) to divide datasets, and the detection accuracy on Cornell dataset is summarized in Table I. The proposed AAGDN achieves an accuracy of 99.3%, which is better than current algorithms. The average inference speed is 18 ms per image, which guarantees real-time grasping on real robots.

To further evaluate the robustness of our method, we compare the AAGDN with the approaches of [6], [16], [17] and [29] in Table III under different evaluation indexes, including the Jaccard index and Angle threshold described in Section V-A. The importance of grasp position has been emphasized in previous sections. Note that our model can still maintain 98.4% accuracy when the Jaccard index comes to 0.4, which means high requirements for the grasping-position accuracy. We can see from the results that our approach achieves the best under all varying evaluation indexes for IW and OW splitting. Furthermore, with

TABLE III
ACCURACY UNDER VARYING JACCARD INDEX AND ANGLE THRESHOLD ON THE CORNELL DATASET

Method	Splitting	Jaccard index					Angle threshold				
		20%	25%	30%	35%	40%	10°	15°	20°	25°	30°
Guo et al. [17]	IW (%)	93.8	93.2	91.0	85.3	-	79.0	90.7	95.1	96.6	97.5
Song et al. [29]		-	95.6	94.9	91.2	87.6	-	-	-	-	-
Chu et al. [6]		-	96.0	94.9	92.1	84.7	-	-	-	-	-
Zhou et al. [16]		98.3	97.7	96.6	95.5	-	86.4	94.4	97.2	97.7	97.7
Ours		**99.5**	**99.3**	**99.3**	**99.2**	**98.4**	**93.7**	**95.6**	**98.2**	**99.3**	**99.3**
Guo et al. [17]	OW (%)	85.1	82.8	79.3	74.1	-	79.0	90.5	95.1	95.9	96.4
Song et al. [29]		-	97.1	97.1	96.4	93.4	-	-	-	-	-
Chu et al. [6]		-	96.1	92.7	87.6	82.6	-	-	-	-	-
Zhou et al. [16]		97.7	96.6	93.8	91.5	-	85.3	93.2	95.5	96.0	96.6
Ours		**98.9**	**98.8**	**98.2**	**96.7**	**95.3**	**94.6**	**97.4**	**98.1**	**98.6**	**98.8**

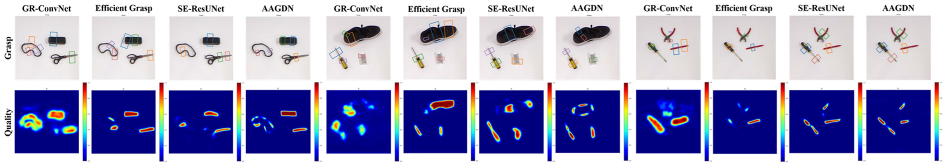

Fig. 6. Comparison results on the multi-object dataset.

the Jaccard index increases and angle threshold decreases, the success rates of [6], [16], [17] and [29] decrease rapidly, but our method remains a high detection accuracy. The results prove the stable and robust grasp detection ability of AAGDN.

C. Experiments and Analysis on Jacquard Dataset

We also perform the comparison experiments on the Jacquard dataset for unseen objects, where Fig. 5(b) shows the results. As shown, our approach can generate grasp poses with higher confidence in quality heatmaps for simple shaped objects, such as a bottle. For objects with complicated structures, our model can also accurately capture the most suitable grasping position with appropriate grasp angle and width. The above advantages may attribute to our model learning more efficient detailed feature information of the object, which are critical for grasping tasks. The experiments indicate that our method behaves more applicable and robust for objects with different shapes.

Table II gives the accuracy of AAGDN compared with current algorithms on the Jacquard dataset. Our approach achieves 96.2% accuracy and outperforms other methods, indicating that the proposed method with augmented attention mechanisms improves the grasp performance.

D. Comparison Studies on the Multi-Object Dataset

To validate the accuracy and robustness for multiple objects, we compare the AAGDN with GR-ConvNet, Efficient Grasp and SE-ResUNet on a multi-object dataset [6], where all objects are unseen before. The generated grasp poses and quality heatmaps are presented in Fig. 6. From the first row we can find the current methods generate inaccurate grasp positions in some cases, and even occasionally predict grasp rectangles for the background. Moreover, their generated grasp angle and width are not suitable enough, which may lead to poor robustness and failed grasp when employed on a real robot. What's more, the quality heatmaps in the second row show that the whole object is always considered as the grasp area when using the current methods, which means they are unable to accurately catch the area that is easy for grasping. In comparison, our method has a better comprehension of grasping scenes and predicts grasp success rate more exactly for different positions of the object.

The detection results indicate that AAGDN refines the performance. We speculate that the previous methods may pay little attention to the information of grasp position, leading to poor spatial expression ability and loss of some key features. Instead, the proposed method dexterously adopts CoA-ResNet, EFFM and FAPM during the whole information propagation. Benefiting from them, the grasp model can establish accurate relationships between the grasps and the features like the shape, outline and position of objects, which is critical for successful grasping.

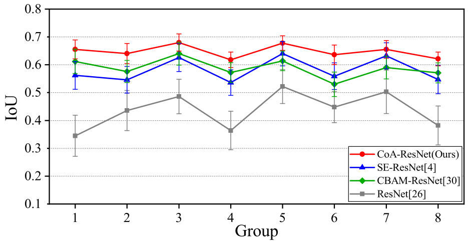

Fig. 7. Effectiveness evaluation for the proposed CoA-ResNet.

E. Effectiveness Evaluation of Proposed Modules

To further investigate the contributions and effectiveness of the proposed CoA-ResNet and EFFM, we conduct specific evaluations. For each evaluation, we conduct 8 groups of experiments, and each group randomly selects 100 unseen objects from the Jacquard dataset. Except for the module used for comparison, all models adopt the same network architecture and stages to ensure fairness.

1) Effectiveness of the CoA-ResNet Module: We adopt traditional ResNet [26], the latest SE-ResNet [4] and CBAM-ResNet [30], and the proposed CoA-ResNet as feature extractors to test the detection accuracy for various objects. To better show the grasping-position accuracy, we propose a novel evaluation index, which is equal to the IoU score between the predicted grasping rectangle and the ground truth when the angle error is less than 20°. In this case, the evaluation index can reflect the position accuracy of generated grasp area. The average accuracies of 8 groups are presented in Fig. 7. As shown, the proposed CoA-ResNet can more accurately locate the optimal grasp area and outperform the other methods on all object groups. Moreover, the CoA-ResNet achieves more robust and stable position accuracy between 0.6 and 0.7 for different groups,

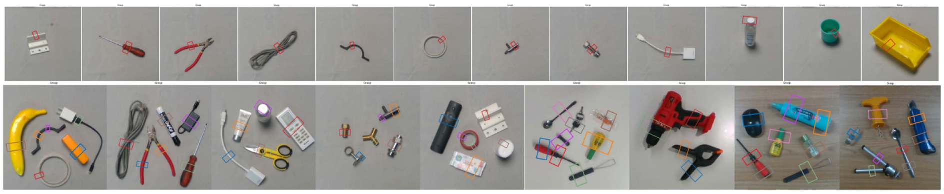

Fig. 8. Real-word performance for single-object and multi-object scenes.

TABLE IV
ABLATION STUDY OF THE ATTENTION MECHANISMS IN PROPOSED MODULES

Coordinate Attention **Attention Fusion Block** **Feature Augmentation Attention**			√			√		√	√
				√		√	√		√
					√		√	√	√
Accuracy (%)	**IW**	95.4	97.2	97.1	96.2	99.1	97.9	98.4	**99.3**
	OW	93.6	96.1	95.8	95.1	98.4	96.5	97.3	**98.8**

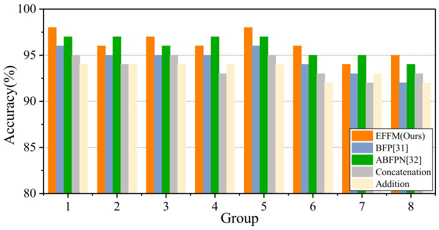

Fig. 9. Effectiveness evaluation for the proposed EFFM.

TABLE V
ABLATION STUDY OF PROPOSED MODULES ON CORNELL DATASET

Network	**Accuracy (%)**	
	IW	**OW**
CoA-ResNet	97.2%	96.1%
CoA-ResNet + FAPM + Add	97.8%	96.8%
CoA-ResNet + FAPM + Concat	98.4%	97.3%
CoA-ResNet + EFFM	99.1%	98.2%
ResNet + FAPM + EFFM	97.9%	96.5%
SE-ResNet + FAPM + EFFM	99.0%	97.9%
CoA-ResNet + FAPM + EFFM	**99.3%**	**98.8%**

TABLE VI
THE PHYSICAL GRASPING RESULTS IN CLUTTERED SCENES

Method	**Physical grasp**	**Success Rate (%)**
GG-CNN [8]	167/200	83.5
GR-ConvNet [9]	172/200	86.0
Efficient Grasp [28]	352/400	88.0
SE-ResUNet [4]	369/400	92.3
Ours	**544/575**	**94.6**

which means that our method is only slightly affected by the physical properties of objects. The results demonstrate that the proposed CoA-ResNet module can pay more attention to the grasping-position information and help the detector locate the optimal grasp area more precisely.

2) *Effectiveness of the EFFM:* We also perform experiments on the impact of the EFFM to explore whether the proposed method can better integrate multi-level features, which is detailed in Fig. 9. We adopt the widely used addition and concatenation approaches, as well as the latest feature fusion methods [31], [32] in computer vision as a contrast. Since the EFFM aims to improve the feature fusion manner and enhance the effectiveness of learned grasp-related features, we adopt the average accuracy on unseen objects as the evaluation index. From the Fig. 9 we can find that our proposed EFFM achieves better grasp accuracy over the other methods. Specifically, we consider that compared with the addition and concatenation method, EFFM can benefit from the adaptive response of features and focus on the most relevant grasping features. And compared to the feature fusion method in [31], the grasp model can obtain more complete and effective feature information to predict grasp poses. Moreover, it is remarkable that the ABFPN in [32] also obtains hopeful results in our experiments; however, it needs high time cost for each detection, which may not be suitable for this real-time grasping task. The results prove that the EFFM performs better at fusing multi-level features to improve the grasp detection accuracy.

F. Ablation Studies

We implement ablation studies based on the Cornell dataset to understand the role of the attentional mechanism in each module by removing the attention blocks. The details of experimental results are shown in Table IV. Moreover, we also investigate the contribution of each component in AAGDN, as shown in Table V. Concretely, with the coordinate attention, the proposed CoA-ResNet can focus on the shape and position information of feasible grasp area and help our model locate the position that is easy to grasp. The developed FAPM enhances multi-scale features as needed and reduces the loss of useful information for grasping. Our proposed EFFM effectively integrates the different-level features and solves the information asymmetry problem. The results prove that above improvements contribute to the high performance of final predictions and achieve promising results.

G. Robotic Grasping

In this part, we perform extensive robotic grasp-and-place experiments. The robotic platform and screenshots of physical grasping in cluttered scenes are shown in Fig. 10. We adopt the Franka Emika robot and the RealSense SR305 camera, which is mounted with the eye-in-hand manner to get a clear view of the graspable objects. We employ the well-trained grasp model into an independent thread, which communicates with the camera and other robot threads through the ROS topic mechanism to subscribe images and publish grasp poses. In each grasp, the detector AAGDN predicts the optimal grasp pose and then transforms it into the robot base frame. At last, the robot approaches the target object by motion planning and closes the gripper to grasp.

We adopt a wide range of objects for grasping, including household objects, 3D-printed and metal objects. All of these objects have never been seen before for our model and are set at random in diverse positions and orientations. The single-object and multi-object grasp detection results of some representational objects are shown in Fig. 8. Besides, we conduct extensive robot grasping experiments in cluttered scenes with other learning-based methods, and the success rates are presented in Table VI. The robot with AAGDN successfully completes 544 grasps,

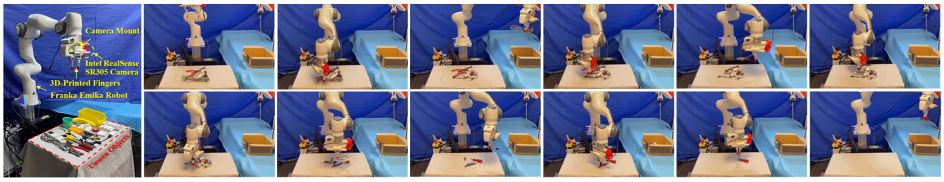

Fig. 10. The experimental setup and screenshots of robot grasping experiments in typical cluttered scenes.

obtaining a 94.6% grasp success rate. The results indicate that the proposed method with augmented attention refines the grasp performance and also achieves better grasp results on real robots.

VI. Conclusion

This letter proposes a high-performance grasp detector, AAGDN, to predict the optimal grasp pose for unknown objects. Benefiting from the developed modules, the AAGDN can focus on the grasping-position features and guarantee complete and effective information propagation. This way, the robot can accurately locate the optimal grasp position and generate suitable grasp angles and widths for different objects. We evaluate the proposed AAGDN on public grasp datasets and it outperforms the current state-of-art algorithms. Moreover, we conduct extensive robotic grasping experiments in various scenes, which further prove that the AAGDN can predict and perform accurate grasps. In sum, the proposed method can learn more efficient grasp-related features and achieve higher accuracy than existing methods. Further research will be extended to developing the universal attention-based grasping approach for other grippers, for example, the three-finger gripper and five-finger dexterous hand.

References

[1] Q. Bai, S. Li, J. Yang, Q. Song, Z. Li, and X. Zhang, “Object detection recognition and robot grasping based on machine learning: A survey,” *IEEE Access*, vol. 8, pp. 181855–181879, 2020.

[2] P. Ni, W. Zhang, X. Zhu, and Q. Cao, “PointNet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 3619–3625.

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

[4] S. Yu, D.-H. Zhai, Y. Xia, H. Wu, and J. Liao, “SE-ResUNet: A novel robotic grasp detection method,” *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 5238–5245, Apr. 2022.

[5] I. Lenz and A. Saxena, “Deep learning for detecting robotic grasps,” *Int. J. Robot. Res.*, vol. 34, no. 4/5, pp. 705–724, 2015.

[6] F.-J. Chu, R. Xu, and P. A. Vela, “Real-world multi-object, multi-grasp detection,” *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 3355–3362, Oct. 2018.

[7] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis.*., 2015, pp. 1440–1448.

[8] D. Morrison, P. Corke, and J. Leitner, “Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach,” in *Proc. Robot. Sci. Syst.*, 2018.

[9] 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.*, 2020, pp. 9626–9633.

[10] S. Wang, Z. Zhou, and Z. Kan, “When transformer meets robotic grasping: Exploits context for efficient grasp detection,” *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 8170–8177, Jul. 2022.

[11] H. Cao, G. Chen, Z. Li, J. Lin, and A. Knoll, “Residual squeeze-and-excitation network with multi-scale spatial pyramid module for fast robotic grasping detection,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13445–13451.

[12] J. Redmon and A. Angelova, “Real-time grasp detection using convolutional neural networks,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2015, pp. 1316–1322.

[13] 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.*, 2019, pp. 4768–4775.

[14] J. Mahler et al., “Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” in *Proc. Robot. Sci. Syst.*, 2017.

[15] H. Zhang, X. Zhou, X. Lan, J. Li, Z. Tian, and N. Zheng, “A real-time robotic grasping approach with oriented anchor box,” *IEEE Trans. Syst. Man Cybern., Syst.*, vol. 51, no. 5, pp. 3014–3025, May 2021.

[16] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, “Fully convolutional grasp detection network with oriented anchor box,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 7223–7230.

[17] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, “A hybrid deep architecture for robotic grasp detection,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 1609–1614.

[18] 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. Automat.*, 2021, pp. 13452–13458.

[19] B. Olshausen, C. Anderson, and D. Van Essen, “A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information,” *J. Neurosci.*, vol. 13, no. 11, pp. 4700–4719, 1993.

[20] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, “BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation,” *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, 2021.

[21] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, “ICNet for real-time semantic segmentation on high-resolution images,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

[22] H. Li, P. Xiong, H. Fan, and J. Sun, “DFANet: Deep feature aggregation for real-time semantic segmentation,” in *Proc. IEEE/Comput. Vis. Found. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9522–9531.

[23] X. Li, H. Zhao, L. Han, Y. Tong, S. Tan, and K. Yang, “Gated fully fusion for semantic segmentation,” in *Proc. Assoc. Adv. Artif. Intell. Conf. Artif. Intell.*, 2020, pp. 11418–11425.

[24] D. Morrison, P. Corke, and J. Leitner, “Learning robust, real-time, reactive robotic grasping,” *Int. J. Robot. Res.*, vol. 39, no. 2/3, pp. 183–201, 2020.

[25] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *Proc. IEEE/CVF Comput. Vis. Found. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[27] A. Depierre, E. Dellandrea, and L. Chen, “Jacquard: A large scale dataset for robotic grasp detection,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 3511–3516.

[28] H. Cao, G. Chen, Z. Li, Q. Feng, J. Lin, and A. Knoll, “Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation,” *IEEE/Amer. Soc. Mech. Engineers Trans. Mechatron.*, early access, Dec. 16, 2022, doi: 10.1109/TMECH.2022.3224314.

[29] Y. Song, L. Gao, X. Li, and W. Shen, “A novel robotic grasp detection method based on region proposal networks,” *Robot. Comput.-Integr. Manuf.*, vol. 65, 2020, Art. no. 101963.

[30] K. Ma, C. A. Zhan, and F. Yang, “Multi-classification of arrhythmias using ResNet with CBAM on CWGAN-GP augmented ECG Gramian Angular Summation Field,” *Biomed. Signal Process. Control*, vol. 77, 2022, Art. no. 103684.

[31] T. Zhang, Y. Cao, L. Zhang, and X. Li, *Efficient Feature Fusion Network Based On Center and Scale Prediction For Pedestrian Detection*. Freehold, NJ, USA: Vis. Comput., 2022, pp. 1–8.

[32] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, “A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection,” *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 3507014.