Hindawi
International Journal of Aerospace Engineering
Volume 2022, Article ID 3396811, 11 pages
https://doi.org/10.1155/2022/3396811

Research Article

Fast and Accurate Hand Visual Detection by Using a Spatial-Channel Attention SSD for Hand-Based Space Robot Teleoperation

Qing Gao,[1,2] **Xin Zhang**,[3,4] **and Wenrao Pang**[1,2]

[1]*Institute of Robotics and Intelligent Manufacturing & School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China*
[2]*Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518129, China*
[3]*State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Science, Shenyang 110016, China*
[4]*Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China*

Correspondence should be addressed to Qing Gao; gaoqing@cuhk.edu.cn

Received 16 March 2022; Revised 7 April 2022; Accepted 15 April 2022; Published 4 May 2022

Academic Editor: Angelo Cervone

Space robot teleoperation is an important technology in the space human-robot interaction and collaboration. Hand-based visual teleoperation can make the operation more natural and convenient. The fast and accuracy hand detection is one of the most difficult and important problem in the hand-based space robot teleoperation. In this work, we propose a fast and accurate hand detection method by using a spatial-channel attention single shot multibox detector (SCA-SSD). The SSD framework is used and improved in our method by introducing spatial-channel attentions with feature fusion. To increase the restricted receptive field in shallow layers, two shallow layers are fused with deep layers by using feature fusion modules. And spatial attention and channel-wise attention are also used to extract more efficient features. This method can not only ease the computational burden but also bring more contextual information. To evaluate the effectiveness of the proposed method, experiments on some public datasets and a custom astronaut hand detection dataset (AHD) are conducted. The results show that our method can improve the hand detection accuracy by 2.7% compared with the original SSD with only 15 fps speed drops. In addition, the space robot teleoperation experiment proves that our hand detection method can be well utilized in the space robot teleoperation system.

1. Introduction

Due to the limited intelligence of space robots, space human-robot interaction plays an important role in the application of space tasks [1]. Teleoperation is one of a widely used space human-robot interaction method [2]. Teleoperation does not depend on the high intelligence capabilities of space robots. It can effectively combine the human decision-making ability with the space robot precise operation ability to improve the operation ability of space robots. There are some devices for space teleoperation. Some traditional devices, such as haptic feedback controllers [3–5], have stable and robust performance but lack of convenience. Some hand-based teleoperation devices [6], such as data gloves [7–9] and surface electromyography (SEMG) wristbands [10–12], have good convenience performance. However, because they are wearable devices, the performance on different people is very different. So, complex calibration work is required before using them. Hand-based visual teleoperation [6] is an emerging teleoperation method. It has the advantages of noncontact, natural, and convenience.

Hand detection is an important and difficult issue in the hand-based visual teleoperation. Because (1) space robot teleoperation needs real-time and robust operation. So, the hand detection should balance fast and accurate performances. (2) Complex backgrounds and changing illumination inside and outside the space station cabin make the astronaut hands

difficult to detect and locate. (3) Hand is a small object. Detection for small objects has always been a difficult problem in computer vision.

To deal with the above problems, a fast and accurate hand detection method is proposed in this paper. SSD framework [13] is used to design the hand detector since its good balance of speed and precision and ease of structural improvement. However, the SSD is not good at detecting small objects. Because it uses shallow layers to detect small objects, and shallow layers have enough contextual information but lack of semantic information. To address the lack of semantic information in the shallow layers, a multiattention module with feature fusion (MA-FF) is proposed to combine shallow layers with deep layers. The multiattention module extract channel attention features from deep and low-resolution feature maps and extract spatial attention features from high resolution layers, respectively. Then, the feature fusion module fuses these features to obtain new shallow layer feature maps with enough contextual and semantic information.

The main contributions and innovations are shown as follows. (1) A spatial-channel attention SSD (SCA-SSD) is proposed to deal with fast and accurate hand detection. The layers for object detection in the SSD structure are visualized to find out which layers play the most important role for small object detection. And these layers are improved and fused with deep layers. A multiattention module with feature fusion (MA-FF) is proposed. It includes a channel attention branch, a spatial attention branch, and a feature fusion branch. (2) A custom astronaut hand detection dataset (AHD) is designed. This dataset collects a large number of astronaut hand images and is used for hand detection verification for space robot teleoperation. (3) The experiments on hand detection datasets proves that the proposed SCA-SSD has fast and accurate hand detection performance, which is superior to some state-of-the-art method. And the experiments on the space robot teleoperation platform prove that the designed hand detector can be well used in the hand-based space robot teleoperation.

The rest of this paper is structured as follows. Section 2 reviews the prior work of hand-based robot teleoperation and hand detection methods. In Section 3, we first describe and visualize the original SSD and then elaborate the structure details of the proposed hand detection method. In Section 4, we provide the results of ablation experiments and comparative experiments on public datasets and a custom AHD dataset. And we also provide the application experiment on hand-based space robot teleoperation platform. Finally, we draw the conclusions and future work in Section 5.

2. Related Work

2.1. Hand-Based Robot Teleoperation. The hand-based robot teleoperation methods include contact and noncontact methods. The mainly contact methods include haptic feedback-based, sEMG-based, and data glove-based methods. Haptic feedback-based teleoperation [3–5] is a traditional teleoperation method. It transmits the 6-Dof position and orientation of human hand to the robot through the haptic feedback controller. For example, the da Vinci surgical telemanipulator [3] can transmit the dual-hand motion information of the chief surgeon through two main joysticks to control the instruments and a 3D high-definition endoscope. The principle of the sEMG-based teleoperation [10–12] is that when hand moves, the arm will generate corresponding motor neuron information, which can be obtained by decoding the sEMG signal. For example, Raspopovic et al. [10] used sEMG equipment to collect sEMG signal of hand gestures and used these gestures to control a dexterous hand. Data glove-based teleoperation [7–9] uses curvature sensors to collect the bending degrees of the fingers and the posture change of the entire human hand, to decode the movement of the hand. Fang et al. [7] designed a novel data glove to control a robotic hand-arm teleoperation system. The above contact teleoperation methods are lack of robustness for different people. The visual teleoperation is robust to different people due to its noncontact advantage [14–16]. For example, Li et al. [14] designed a mobile robot hand-arm teleoperation system by using vision and IMU. Handa et al. [15] designed a vision-based teleoperation method for a dexterous robotic hand-arm system. Table 1 shows the comparison and summary of the above hand-based robot teleoperation methods.

2.2. Hand Detection Methods. Traditional visual hand detection methods [17] mainly include skin color-based hand detection, motion flow information-based hand detection, and shape model-based hand detection. These methods only extract the shallow information of hands, which are subject to many conditions. Nowadays, deep learning-based hand detection methods can achieve better detection performance in complex environment [18–20]. Hand detection can be regarded as a kind of object detection. There are some typical deep learning-based object detectors, such as RCNN series [21, 22], YOLO series [23–25], and SSD series [13, 26, 27]. Among them, the SSD is a light weight one-stage network, which considers speed and accuracy trade-off and is easy to modify. For example, Gao et al. [18] designed a feature-map-fused SSD for robust real-time hand detection and localization. He also used SSD and body pose estimation for dual-hand detection [19]. Yu et al. designed a deep temporal model-based identity-aware hand detector by using the SSD framework for space human-robot interaction [20]. However, the SSD is stuck with the speed and accuracy dilemma for small object detection. Some useful methods and tricks are proposed to resolve this dilemma. DSSD [26] attempts to recover higher resolution features and adds with the primary features through shortcut connection. FSSD [28], DF-SSD [29], RSSD [30], and ESSD [31] provided many feature fusion methods to add more contextual information into shallow feature maps. Table 2 shows the comparison and summary of the above hand detection methods.

3. Spatial-Channel Attention SSD

In this section, first, the original SSD is introduced and visualized. Then, the proposed SCA-SSD is introduced, which includes the multi-attention module and feature fusion module.

3.1. SSD Introduction and Visualization. In this subsection, the SSD architecture is introduced first. And then, the

Table 1: Comparison and summary of the hand-based robot teleoperation methods.

Method	Brief methodology	Highlights	Limitations
Haptic feedback-based method [3–5]	This method transmits the 6-Dof position and orientation of human hand to the robot through the haptic feedback controller.	High accuracy and mature technology.	Contact method and only be used for 6-DOF control.
sEMG-based method [10–12]	When hand moves, the arm will generate corresponding motor neuron information, which can be obtained by decoding the sEMG signal.	High accuracy.	Contact method and lack of robustness.
Data glove-based method [7–9]	This method uses curvature sensors to collect the bending degrees of the fingers and the posture change of the entire human hand, to decode the movement of the hand.	High accuracy and mature technology.	Contact method and lack of robustness.
Vision-based method [14–16]	This method uses camera to capture the movements of human hands and maps the human hand movement information to the robot.	Strong noise and immature technology.	Noncontact method, strong robustness, and naturalness.

Table 2: Comparison and summary of the hand detection methods.

Method	Brief methodology	Highlights	Limitations
Traditional visual hand detection methods [17]	They mainly include skin color-based hand detection, motion flow information-based hand detection, and shape model-based hand detection.	Small amount of calculation, mature technology.	These methods only extract the shallow information of hands, which are subject to many conditions.
Deep learning-based hand detection methods [18–20]	They autonomously extract deep features of hand images through deep neural networks.	High detection performance.	They are stuck with the speed and accuracy dilemma for small object detection.

detection visualization in SSD is shown to find out which layers are suitable for improving.

3.1.1. SSD Architecture. The SSD [13] is one of the outstanding one-stage detectors with high speed and accuracy. The architecture is shown in Figure 1. The VGG-16 is used as its backbone, and several extra convolution layers on the top of the network are used for prediction and classification by filters directly. Unlike other detectors, SSD uses pyramidal multiresolution feature maps as convolutional detector input, which means it handles different scales in different resolution feature maps. The SSD brings significant improvement on speed because of its one-stage architecture. However, it cannot get a high detection accuracy on small object. Because the shallow layers for detection have much contextual information but less semantic information. While the deep layers for detection are reverse. Small object detection needs enough semantic and contextual information for its low resolution. So, feature maps with enough semantic and contextual information should be designed for hand detection.

3.1.2. Detection Visualization in SSD. To find out which layers are suitable for improving for small object detection, the results of feature maps for object detection in SSD are visualized. We select one convolution layer as the input of the detector and block other convolution layers which means we only use one specific convolution layer to detect objects. The results are shown in Figure 2, which shows that the small objects are easier detected in shallow layers (conv4_3 and conv7 layers), and large objects are easier detected in deep layers (conv8_2, conv9_2, and conv10_2 layers). Because the contextual information is vital to small object detection and shallow layers have enough contextual information. However, due to the lack of semantic information, there are some missing detection results of small objects in conv4_3 and conv7 layers. Once it misses the object in shallow layers, it has no chance to be detected in the subsequent deep layers. To increase the accuracy of small object detection, we propose the SCA-SSD. A multiattention module is employed on conv4_3 and conv7 layers and then fuses them with conv8_2 and conv11_2, separately. The details are presented below.

3.2. SCA-SSD Architecture. In this subsection, the overview of the SCA-SSD is introduced first. Then, the multiattention module and feature fusion module are introduced, respectively.

3.2.1. Overview of SCA-SSD. The architecture of our proposed SCA-SSD is introduced and shown in Figure 3. From the figure, we can see that the SCA-SSD reuses the multiscale and one-stage architecture of the original SSD. Two multiattention branch with feature fusion (MA-FF) modules are employed on the shallow layers conv4_3 and conv7, respectively. They use the multiattention modules to extract channel and spatial features and use feature fusion modules to fuse the two shallow layers (conv4_3 and conv7) with the deep layers (conv8_2 and conv11_2). Finally, the two new feature maps output from the MA-FF modules are mainly used for small object detection.

3.2.2. Multiattention Module. To address the lack of information in shallow layers, we propose a multiattention module with feature fusion, and the improved structure of conv4_3

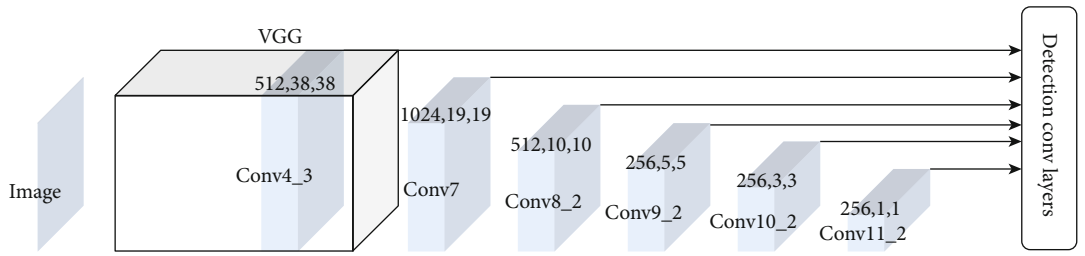

Figure 1: Original SSD architecture.

Figure 2: Detection results for each layer of SSD. In SSD, shallow layers are usually used to detect small objects and deep layers respond to objects in large scales.

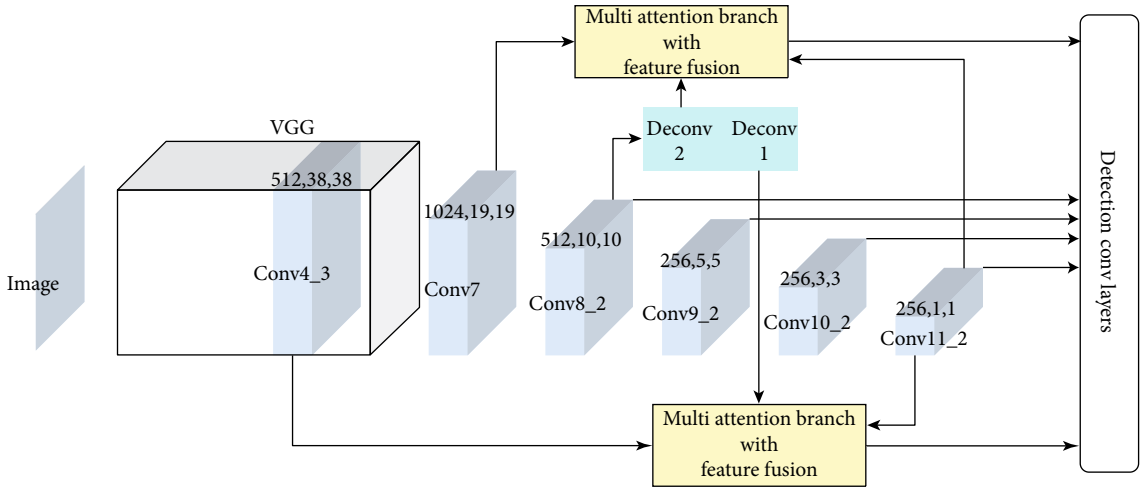

Figure 3: The architecture of the SCA-SSD.

is shown in Figure 4 as an example. The design of the attention module is inspired by bottleneck attention module (BAM) [32]. To be specific, first, the spatial attention branch Atts is employed after conv4_3 and conv7, respectively. After that, channel attention branch Att_c is employed after the conv11_2 whose resolution is 1×1 so that we can skip the global pooling operation in the squeeze stage coincidentally. Then, the Atts and Att_c are combined by element wise add operation to generate the cross resolution spatial-channel attention which terms Att_{sc}. Finally, the sigmoid is applied for Att_{sc} to obtain the weighted Att_{sc} and then multiply with feature maps from the corresponding feature map. For instance, as shown in Figure 4, weighted Att_{sc} is obtained from conv4_3 so that it multiplies and adds with conv4_3.

The spatial branch structure is shown in Figure 5(1). This branch follows encoder-decoder structure, but we do not down the resolution of the feature map to preserve more information. Each branch consists of a 1×1 convolution layer to reduce the dimensions of channels, and two 3×3 dilated convolution layers are employed for obtaining long-range information with a widely receptive field. Then, it will restore the number of channels as input by another 1×1 convolution layer. In practice, each convolution layer and dilated convolution layer are followed by a batch normalization and a ReLU activation function except for the last 1×1 convolution layer. Set the input feature map is F, the output Att_s can be expressed as

$$Att_s(F) = \sigma\left(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])\right), \quad (1)$$

where σ denotes the sigmoid function, $f^{7\times7}$ denotes a conclusion operation with the filter size of 7×7.

The structure of the channel branch is shown in Figure 5(2). In Att_c, in order not to affect the value of conv11_2 feature map, one 1×1 convolution layer following

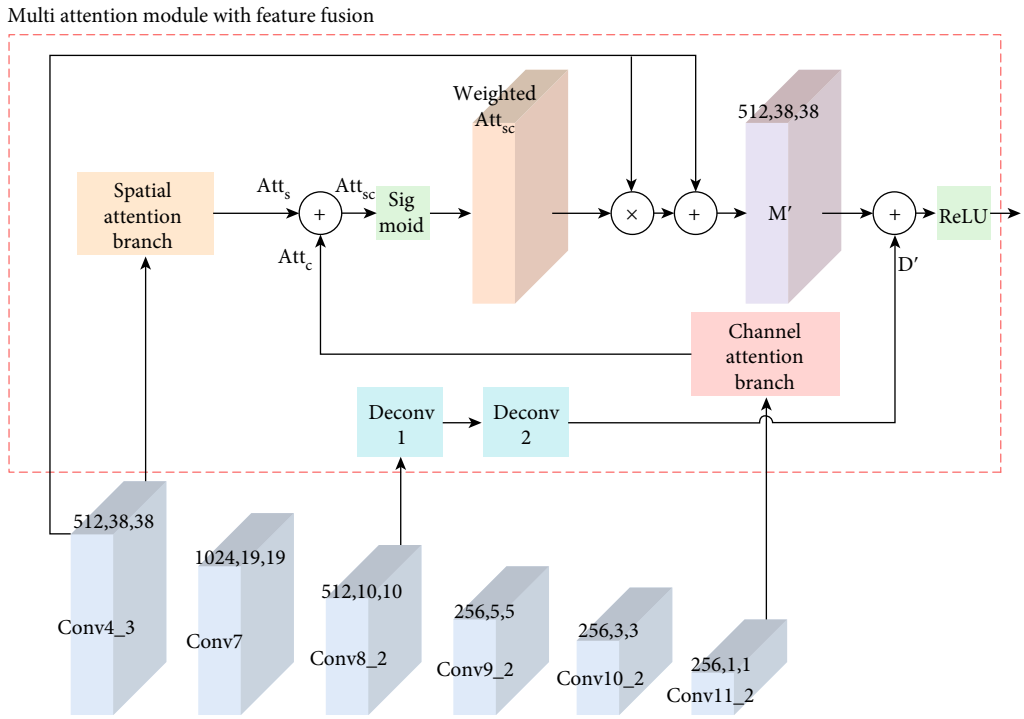

FIGURE 4: Multiattention module with feature fusion. The structure is the MA-FF on conv4_3 layer. In practice, the output of Deconv-1 adds with conv7, and the output of Deconv-2 adds with conv4_3. + means the symbol of element wise add operation, and × means the corresponding element multiply operation.

ReLU after the conv11_2 is employed. In excitation, two fully connected layers are used to blend different values in different channels. Then, they are expanded to match the sizes of conv4_3 and conv7. Set the input feature map is F, the output Att_c can be expressed as

$$Att_c(F) = \sigma(MLP(\text{AvgPool}(F)) + MLP(\text{MaxPool}(F))), \tag{2}$$

where σ denotes the sigmoid function.

3.2.3. Feature Fusion Module. Even though the multiattention module brings extra contextual information to shallow layers, the spatial attention branch still has a drawback. The context is encoded as an attention mask so that the value is limited between zero and one. By multiplying with input feature map, it can enhance the useful information for detection. However, context and long-range information are encoded as attention mask Att_{sc} which only provides weighted value. So, to capture more context, a feature fusion module which can be embedded within the multiattention module is proposed, and it is shown in Figure 4.

In the feature fusion module, two deconvolution layers are employed to restore the size of feature map from 10×10 to 19×19 and 38×38, so that it can match the size of conv7 and conv4_3. Our feature fusion module is inspired by DSSD [26], and two deconvolution layers are only used to avoid increasing much computational burden. In each Deconv-n block, it includes a deconvolution layer and a batch normalization (BN). After deconvolution, fusion operation is employed to merge a reweighted feature map $M^{'}$ with the output of deconvolution $D^{'}$. Follow the feature-fusion SSD [27], element-wise add is used as the fuse operation. It can be proved that element-wise add outperforms the concatenate operation in the feature-fusion SSD [27]. At the end of this module, the ReLU activation function is employed.

4. Experiments and Analysis

In this section, to compare performance with the state-of-the-art object detection methods, experiments are conducted on Pascal VOC dataset [33] first. Then, experiments are conducted on the Oxford hands dataset [34] to demonstrate the effectiveness of our proposed method on public hand detection datasets. After that, the AHD dataset will be introduced, and experiments will be conducted on this dataset to prove the performance of astronaut hand detection. The mean average precision (mAP) is adopted as evaluation metric to evaluate our model prediction performance.

We implement the MA-SSD based on PyTorch [35]. The data augmentation method is followed with SSD [13], and the VGG-16 is used as the pretrained backbone. All experiments are performed on 4 NVIDIA RTX 2080 Ti GPU.

4.1. Experiments on the Pascal VOC Dataset

4.1.1. Training. In training stage, the batch size is set to 32, and the learning rate is set to $1 \times 10 - 3$ with a warm-up phase at the first 500 times iteration. However, the experiment resulted

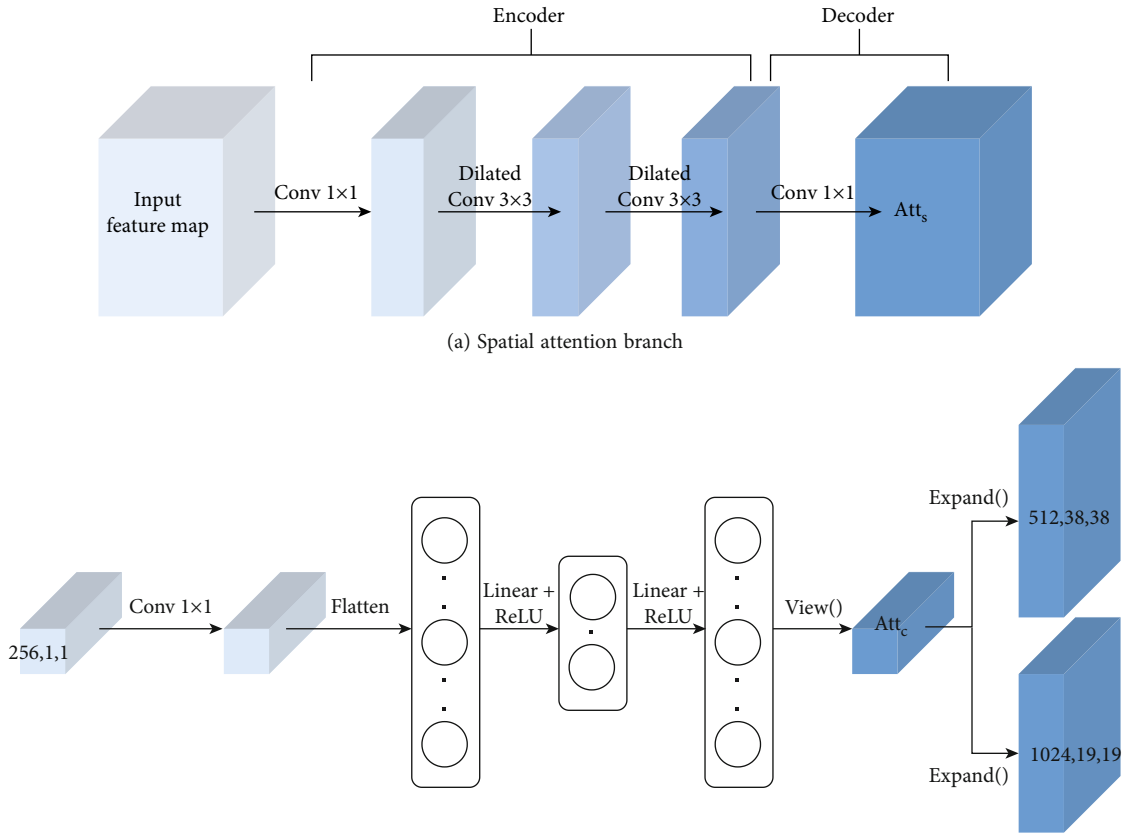

Figure 5: Spatial branch and channel branch in the MA-FF module.

the default learning rate is too small. Instead, the learning rate is set to $4 \times 10 - 3$ with a warm-up phase at first 2800 iterations. Learning rate should be increased from $1 \times 10 - 6$ with warm-up factor as 0.03333 gradually. The training step is set to 140 k iterations totally, and the learning rate is divided by 10 at 84 k and 112 k iterations which is different from original SSD [13] but similar to RFB-Net [36]. Following the trick in RFB-Net, the number of prior boxes in conv4_3 is increased to 6.

4.1.2. Introduction of the Pascal VOC Dataset. The objects in the Pascal VOC 2007 dataset include 4 categories and 20 subcategories, which are vehicle (car, bus, bicycle, motorbike, airplane, boat, and train), household (chair, sofa, dining table, TV, bottle, and potted plant), animal (cat, dog, cow, horse, sheep, and bird), and person. These images are collected from flickr and Microsoft Research Cambridge (MSRC) dataset. The dataset includes 9,963 images containing 24,640 annotated objects.

4.1.3. Comparative Experiments. To demonstrate the performance of the proposed SCA-SSD, some other state-of-the-art methods are compared. The results are shown in Table 3. For a fair comparison, the updated SSD [37] is used as our baseline, which can get a 77.7% mAP on the VOC test dataset. It is slightly higher than that of the original SSD [13], which mAP is 77.2%. By employing multiattention and fusion modules on the SSD, it achieves a 79.9% mAP, which is 2.7% higher than that of the original SSD [13] and 2.2% higher than that of the baseline [37]. The SCA-SSD brings significant improvement into SSD with the least impact on speed. It is only 15 FPS slower than the original SSD. And the mAP of the SCA-SSD is even higher than that of the SSD512, which has a higher input resolution (512×512) than that of the SCA-SSD (300×300). We also compare the results of the proposed SCA-SSD with some state-of-the-art object detection methods like faster-RCNN [21], YOLO v4 [25], R-FCN [38], and StairNet [39]. From Table 1, we can see that the performance of the SCA-SSD is higher than most of the state-of-the-art methods both on accuracy and seed. In addition, we also show the results of some SSD-series methods like DSSD [26] and FSSD [28]. To the best of our knowledge, our SCA-SSD achieves the best performance within SSD-series methods. It proves that the proposed SCA-SSD can achieve a great performance for object detection both on speed and accuracy.

4.2. Experiments on Oxford Hands Dataset

4.2.1. Introduction of the Oxford Hands Dataset. The hand detection is different with normal object detection. It has small size and changeable shape. To better prove the performance of the SCA-SSD for hand detection, the experiments on hand detection dataset are also conducted. The Oxford hands dataset [34] which is a public hand detection dataset is used for training and testing. In the dataset, a total of 13050 hand

Table 3: Comparison of object detection methods on VOC 2007 and VOC 2012 test dataset. In this table, the SCA-SSD is compared with some state-of-the-art methods and other SSD-based methods to illustrate the promising performance on object detection.

Method	Backbone	Input size	mAP	Device	FPS
Fast RCNN [21]	ResNet-101	600 × 1000	76.4	K40	2.4
YOLO v4 [25]	DarkNet-19	352 × 352	78.2	Titan X	81
R-FCN [38]	ResNet-101	600 × 1000	79.5	K40	5.8
StairNet [39]	VGG-16	300 × 300	78.8	Titan X	30
SSD300 [13]	VGG-16	300 × 300	77.2	2080Ti	119
Baseline [37]	VGG-16	300 × 300	77.7	2080Ti	119
DSSD321 [26]	ResNet-101	321 × 321	78.6	Titan X	9.5
FSSD300 [28]	VGG-16	300 × 300	78.8	1080Ti	65.8
FA-SSD300 [40]	ResNet-101	300 × 300	78.3	—	34.7
FF-SSD300 [27]	VGG-16	300 × 300	78.9	—	43
Shift SSD300 [41]	VGG-16	300 × 300	78.3	Titan X	77
ESSD++300 [24]	VGG-16	300 × 300	79.2	—	52
DF-SSD300 [29]	Dense-32-S-1	300 × 300	78.9	Titan X	11.6
RSSD300 [30]	VGG-16	300 × 300	78.5	—	35
SCA-SSD300	VGG-16	300 × 300	79.9	2080Ti	104

instances are annotated. Hand instances larger than a fixed area of bounding box (1500 sq. pixels) are considered "big" enough for detections and are used for evaluation. This gives around 4170 high-quality hand instances. In each image, all the hands that can be perceived clearly by humans are annotated.

4.2.2. Ablation Experiment. To understanding SCA-SSD structure deeper and better, several ablation experiments are conducted to show the effectiveness of each module of the network on hand detection. The results are summarized in Table 4. In this experiment, first, we add channel attention and spatial attention models on the baseline structure, respectively. The mAP can increase 1.6% and 1.3% compared with the baseline method. And the speeds only drop by 3FPS. It proves that the proposed channel attention and spatial attention models are effective in hand detection. Second, we take the feature fusion module away from the SCA-SSD, which terms as SCA-SSD w/o fusion. The result decreases from 44.6% to 43.8% compared with the SCA-SSD w/fusion, which indicates the feature fusion module is effective in hand detection. The feature fusion module can improve 0.8% of mAP but it has little impact on the speed of inference, the speed still keeps on over 100 FPS (104FPS). So, the ablation experiment results show that the proposed channel attention, spatial attention, and feature fusion modules are effective to improve the performance of hand detection.

Table 4: Ablation experiment results on the Oxford hand dataset.

Model	mAP	FPS
Baseline [37]	40.2	119
w/ channel	41.8	116
w/ spatial	41.5	116
SCA-SSD w/o fusion	43.8	114
SCA-SSD w/fusion	44.6	104

Table 5: The verification results of the SCA-SSD hand detector on the AHD dataset.

IoU	Area	mAP
0.50 : 0.95	All	0.69
0.50	All	0.88
0.75	All	0.82
0.50 : 0.95	Small	0.56
0.50 : 0.95	Medium	0.62
0.50 : 0.95	Large	0.81

4.3. Experiments on AHD Dataset

4.3.1. AHD Dataset. To further verify the effectiveness of the designed SCA-SSD hand detector in hand-based space robot teleoperation, the experiment on the space environment images should be conducted. Since there is no such hand detection dataset, we customize a set of astronaut hand images in various intra/extravehicular activities from some sci-fi movies and YouTube resource. We named it AHD dataset. The dataset includes a total of 2000 images and more than 4000 instances. All hands in the images are labelled as "hand."

4.3.2. Verification Experiment and Visualization. The AHD dataset is just used for verification. The hand detector trained on the Oxford hands dataset is verification on the AHD dataset, and the results are shown in Table 3. From Table 5, we can see that when the IoU is 0.50 : 0.95, the hand detect accuracy is 0.69. And when the IoU is 0.50, the hand detection accuracy is 0.88. It is proved that the SCA-SSD

Figure 6: Some hand detection results on the AHD dataset.

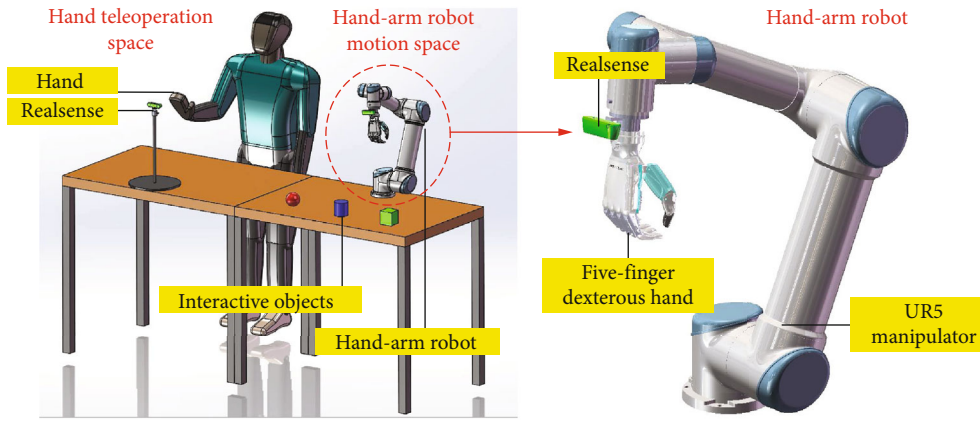

Figure 7: Space robot teleoperation platform. (a) The platform includes a hand teleoperation space and a hand-arm robot motion space. (b) The hand-arm robot includes a UR5 manipulator, a five-finger dexterous hand, and a RealSense.

hand detector can achieve good performance on the AHD dataset. And when the hand areas are small, medium, and large, the hand detection accuracies are 0.56, 0.62, and 0.81, respectively. It is proved that the SCA-SSD hand detector can achieve good performance on hands with various areas.

To better show the results of the hand detection for astronaut's hand, some of the result images are visualized as follows. From Figure 6, we can see that the proposed SCA-SSD hand detector can detect astronaut's hands in various scenes.

4.4. Experiments on Space Robot Teleoperation Platform. The SCA-SSD hand detector is utilized in a designed space robot teleoperation platform, which is shown in Figure 7. The teleoperation platform includes a hand teleoperation space and a hand-arm robot motion space. A RealSense camera can

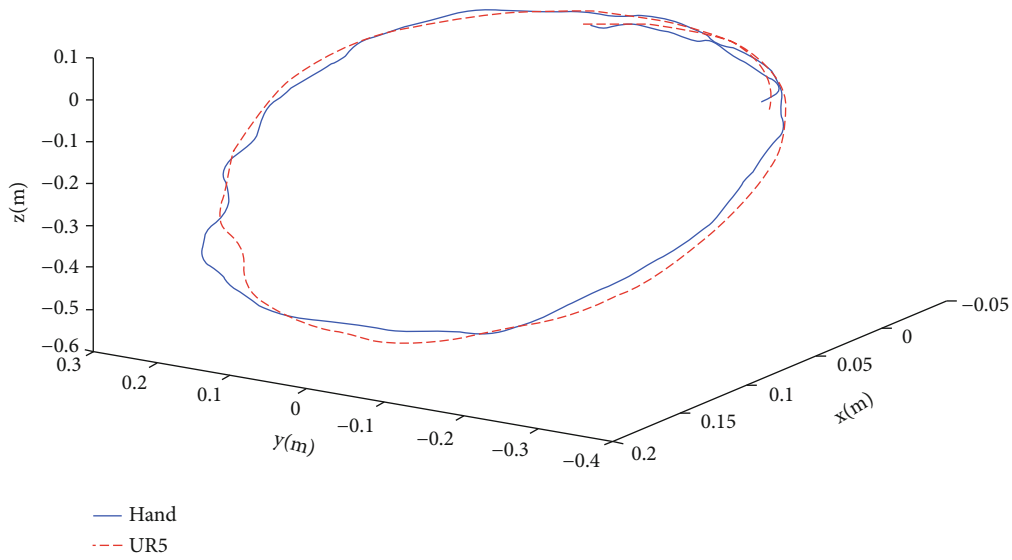

Figure 8: Motion trajectories of hand and the end effector of UR5.

capture the astronaut's hands in real time. After that, the SCA-SSD hand detector can detect hands on the RGB images, and then, the 2D hand positions can be mapped to the corresponding depth images to obtain the 3D hand positions. Then, the real-time hand positions in the hand teleoperation space can be transferred to the hand-arm robot motion space by using the following mapping relationship equation.

$$\begin{cases} x_i^R = x_{i-1}^R + \lambda\left[x_i^H - x_{i-1}^H\right], \\ y_i^R = y_{i-1}^R + \lambda\left[y_i^H - y_{i-1}^H\right], \\ z_i^R = z_{i-1}^R + \lambda\left[z_i^H - z_{i-1}^H\right], \end{cases} \tag{3}$$

where the (x^R, y^R, z^R) is the position of the end effector of the robot, and the (x^H, y^H, z^H) is the hand position in the camera coordinate system. (x_i^R, y_i^R, z_i^R) is the hand position in i-th frame, and $(x_{i-1}^R, y_{i-1}^R, z_{i-1}^R)$ is the hand position in $(i-1)$-th frame. λ is a scale factor, and we set $\lambda = 1$ in the teleoperation experiment.

By collecting the motion trajectories of hand in the camera coordinate system and robot end effector in the robot coordinate system, the trajectories are shown in Figure 8.

From Figure 8, we can see that the end effector of the robot can track the movement trajectory of the hand very well. And the maximum error is only 9.3 mm.

5. Conclusion and Future Work

In this work, a fast and accurate hand detection method was proposed by using a spatial-channel attention single shot multibox detector (SCA-SSD). And the proposed hand detector was utilized in a hand-based space robot teleoperation system. Specifically, two shallow layers were fused with deep layers by using feature fusion modules to increase the restricted receptive field in shallow layers. And spatial attention and channel-wise attention were also used to extract more efficient features. This method can not only ease the computational burden but also bring more contextual information. The comparative experiment, ablation experiment, and verification experiment have proved the good performance of the proposed SCA-SSD hand detector. Finally, the experiment on space robot teleoperation platform has demonstrated that the proposed SCA-SSD hand detector can be applied well in the space robot teleoperation. There are some limitations of the proposed hand detection and teleoperation method. First, the proposed method is only trained on public datasets, and due to the small sizes of the public datasets, the generalization ability of hand detection is not strong. Second, only the detection and localization of hands cannot control the space robots well, and the subsequent recognition of hand gestures and poses is also required.

In the future, hand gesture recognition methods need further research to realize space robot teleoperation for complex tasks. In addition, skeleton-based hand detection and pose estimation also require further research to achieve more precise teleoperation.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62006204 and 62103407) and partly supported by the Shenzhen Outstanding Scientific and Technological Innovation Talents Training Project (RCBS20210609104516043).

References

[1] K. Hambuchen, J. Marquez, and T. Fong, "A review of NASA human-robot interaction in space," *Current Robotics Reports*, vol. 2, no. 3, pp. 265–272, 2021.

[2] M. Shahbazi, S. F. Atashzar, and R. V. Patel, "A systematic review of multilateral teleoperation systems," *IEEE Transactions on Haptics*, vol. 11, no. 3, pp. 338–356, 2018.

[3] C. Freschi, V. Ferrari, F. Melfi, M. Ferrari, F. Mosca, and A. Cuschieri, "Technical review of the da Vinci surgical telemanipulator," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 9, no. 4, pp. 396–406, 2013.

[4] A. Bolopion and S. Régnier, "A review of haptic feedback teleoperation systems for micromanipulation and microassembly," *IEEE Transactions on Automation Science and Engineering*, vol. 10, no. 3, pp. 496–502, 2013.

[5] Y. Liang, G. Du, C. Li, C. Chen, X. Wang, and P. X. Liu, "A gesture-based natural human-robot interaction interface with unrestricted force feedback," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[6] R. Li, H. Wang, and Z. Liu, "Survey on mapping human hand motion to robotic hands for teleoperation," *IEEE Transactions on Circuits and Systems for Video Technology*, p. 1, 2021.

[7] B. Fang, D. Guo, F. Sun, H. Liu, and Y. Wu, "A robotic hand-arm teleoperation system using human arm/hand with a novel data glove," in *In 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 2483–2488, Zhuhai, China, 2015.

[8] L. Dipietro, A. M. Sabatini, and P. Dario, "A survey of glove-based systems and their applications," *Ieee transactions on systems, man, and cybernetics, part c (applications and reviews)*, vol. 38, no. 4, pp. 461–482, 2008.

[9] C. Mizera, T. Delrieu, V. Weistroffer, C. Andriot, A. Decatoire, and J.-P. Gazeau, "Evaluation of hand-tracking systems in teleoperation and virtual dexterous manipulation," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1642–1655, 2019.

[10] S. Raspopovic, M. Capogrosso, F. M. Petrini et al., "Restoring natural sensory feedback in real-time bidirectional hand prostheses," *Science Translational Medicine*, vol. 6, no. 222, p. 222ra19, 2014.

[11] X. Lv, C. Dai, H. Liu et al., "Gesture recognition based on sEMG using multi-attention mechanism for remote control," *Neural Computing and Applications*, pp. 1–11, 2022.

[12] S. E. Ovur, X. Zhou, W. Qi et al., "A novel autonomous learning framework to enhance sEMG-based hand gesture recognition using depth information," *Biomedical Signal Processing and Control*, vol. 66, p. 102444, 2021.

[13] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *In European conference on computer vision, Computer Vision – ECCV 2016*, pp. 21–37, Springer, Cham, 2016.

[14] S. Li, J. Jiang, P. Ruppel et al., "A mobile robot hand-arm teleoperation system by vision and imu," in *In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10900–10906, Las Vegas, NV, USA, 2021.

[15] A. Handa, K. V. Wyk, and W. Yang, "Dexpilot: vision-based teleoperation of dexterous robotic hand-arm system," in *In 2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9164–9170, Paris, France, 2020.

[16] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: learning a robotic hand imitator by watching humans on Youtube," 2022, http://arxiv.org/abs/2202.10448.

[17] M. Kölsch and M. A. Turk, "Robust hand detection," *Current Robotics Reports*, vol. 4, pp. 614–619, 2004.

[18] Q. Gao, J. Liu, and Z. Ju, "Robust real-time hand detection and localization for space human–robot interaction based on deep learning," *Neurocomputing*, vol. 390, pp. 198–206, 2020.

[19] Q. Gao, J. Liu, Z. Ju, and X. Zhang, "Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9663–9672, 2019.

[20] J. Yu, H. Gao, D. Zhou, J. Liu, Q. Gao, and Z. Ju, "Deep temporal model-based identity-aware hand detection for space human-robot interaction," *IEEE Transactions on Cybernetics*, pp. 1–14, 2021.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances In Neural Information Processing Systems*, vol. 28, 2015.

[22] Z. Cai and N. Vasconcelos, "Cascade r-cnn: delving into high quality object detection," in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, Salt Lake City, 2018.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Silicon Valley, 2016.

[24] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, http://arxiv.org/abs/1804.02767.

[25] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, http://arxiv.org/abs/2004.10934.

[26] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: deconvolutional single shot detector," 2017, http://arxiv.org/abs/1701.06659.

[27] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, "Feature-fused SSD: fast detection for small objects," *In Ninth International Conference on Graphic and Image Processing (ICGIP 2017)*, vol. 10615, p. 106151E, 2018.

[28] Z. Li and F. Zhou, "FSSD: feature fusion single shot multibox detector," 2017, http://arxiv.org/abs/1712.00960.

[29] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: an improved SSD object detection algorithm based on DenseNet and feature fusion," *IEEE access*, vol. 8, pp. 24344–24357, 2020.

[30] J. Jeong, H. Park, and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," 2017, http://arxiv.org/abs/1705.09587.

[31] J. Leng and Y. Liu, "An enhanced SSD with feature fusion and visual reasoning for object detection," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6549–6558, 2019.

[32] J. Park, S. Woo, J. Lee, and I. S. Kweon, "Bam: bottleneck attention module," 2018, http://arxiv.org/abs/1807.06514.

[33] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[34] A. Mittal, A. Zisserman, and P. H. Torr, "Hand detection using multiple proposals," *In Bmvc*, vol. 2, no. 3, p. 5, 2011.

[35] A. Paszke, S. Gross, F. Massa et al., "Pytorch: an imperative style, high-performance deep learning library," *Advances In Neural Information Processing Systems*, vol. 32, 2019.

[36] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *In Proceedings of the European conference on computer vision (ECCV)*, pp. 385–400, 2018.

[37] C. Li, *High Quality, Fast, Modular Reference Implementation of SSD in PyTorch*, 2018, https://github.com/lufficc/SSD.

[38] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: object detection via region-based fully convolutional networks," *Advances In Neural Information Processing Systems*, vol. 29, 2016.

[39] S. Woo, S. Hwang, and I. S. Kweon, "Stairnet: top-down semantic aggregation for accurate one shot detection," in *In 2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 1093–1102, Lake Tahoe, NV, USA, 2018.

[40] J. Lim, M. Astrid, H. Yoon, and S. Lee, "Small object detection using context and attention," in *In 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 181–186, Jeju Island, Korea (South), 2021.

[41] L. Fang, X. Zhao, and S. Zhang, "Small-objectness sensitive detection based on shifted single shot detector," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 13227–13245, 2019.