applied sciences

MDPI

Article

Visual Sorting Method Based on Multi-Modal Information Fusion

Song Han, Xiaoping Liu * and Gang Wang

School of Modern Post, Beijing University of Posts and Telecommunications, Beijing 100876, China; hansong@bupt.edu.cn (S.H.); wg58977@bupt.edu.cn (G.W.)

* Correspondence: liuxp@bupt.edu.cn

Abstract: Visual sorting of stacked parcels is a key issue in intelligent logistics sorting systems. In order to improve the sorting success rate of express parcels and effectively obtain the sorting order of express parcels, a visual sorting method based on multi-modal information fusion (VS-MF) is proposed in this paper. Firstly, an object detection network based on multi-modal information fusion (OD-MF) is proposed. The global gradient feature is extracted from depth information as a self-attention module. More spatial features are learned by the network, and the detection accuracy is improved significantly. Secondly, a multi-modal segmentation network based on Swin Transformer (MS-ST) is proposed to detect the optimal sorting positions and poses of parcels. More fine-grained information of the sorting parcels and the relationships between them are gained by adding Swin Transformer models. Frequency domain information and depth information are used as supervision signals to obtain the pickable areas and infer the occlusion degrees of parcels. A strategy for the optimal sorting order is also proposed to ensure the stability of the system. Finally, a sorting system with a 6-DOF robot is constructed to complete the sorting task of stacked parcels. The accuracy and stability the system are verified by sorting experiments.

Keywords: multi-modal; self-attention; Swin Transformer; depth estimation; robot sorting

Citation: Han, S.; Liu, X.; Wang, G. Visual Sorting Method Based on Multi-Modal Information Fusion. *Appl. Sci.* **2022**, *12*, 2946. https://doi.org/10.3390/app12062946

Academic Editor: Giovanni Boschetti and João Miguel da Costa Sousa

Received: 12 February 2022
Accepted: 10 March 2022
Published: 14 March 2022

1. Introduction

With the vigorous development of e-commerce and the rising labor costs, more and more e-commerce and logistics companies are building automated logistics sorting centers. The degree of automation is getting higher and higher. However, there are still some defects in the automated sorting process where human assistance is required. For example, the disorderly stacking of express parcels requires manual sorting and handling, which greatly limits the efficiency of express sorting and transportation. For such scenarios, industrial robots are used to replace manpower in our solution, and a visual sorting method based on multi-modal information fusion (VS-MF) is proposed. The proposed strategy could realize the detection of the optimal sorting position, pose and order for disorderly stacked express parcels.

Robotic visual sorting systems [1–3] generally use the environmental information collected by 3D vision systems or RGB-D cameras as the original input source. Then, one or more potential sorting positions are predicted by visual algorithms. The grasping posture of the robot is gained according to the object posture or the image features. Finally, the robot achieves the grasping through trajectory planning. The key tasks in the whole process are improving the detection accuracy of sorting position and pose for each parcel, and determining the optimal sorting order of multi-object scene. The completion effect of these two tasks will influence the accuracy safety of the final sorting and the efficiency of the sorting system.

Deep learning methods have achieved great success in various vision tasks. Many researchers have introduced deep learning frameworks into the field of robot sorting, forming methods based on two different types of vision tasks. One is the visual sorting system based

on object detection. Ulrich Viereck [4] used a deep neural network to learn the method of closed-loop controller for machine sorting, and obtained the accurate grasping posture of the object by training the convolutional neural network to learn the distance function, with which a better grasping effect was achieved. Xuedan DU [5] detected and classified the target objects in the image through a deep learning-based object detection algorithm, and then performed the grasping position detection. The robot grasping position was determined by performing a box searching on each image. In similar work, many scholars have carried out special modeling trained for grasping rectangle detection. Kumra [6] proposed a multi-modal detection model based on deep learning. Color and depth information are used to detect objects. After being trained and tested on the Cornell dataset, this method achieved an accuracy of 89.21%, while ensuring the real-time performance of the system. Similarly, Zhang Hanbo [7] proposed a fully convolutional visual grasping detection network based on directed anchor boxes, which realizes real-time detection of grasping. The aforementioned methods not only achieve high accuracy and a high recall rate, but also ensure the real-time performance of the system to a certain extent. However, the mentioned methods are quite difficult to accurately detect the targets when objects are heavily stacked since the semantic understanding of complex scenes is lacking. Thus, they are quite difficult to accurately detect when objects are heavily stacked. Recently, some other methods for robot grasping were proposed based on 3D object detection [8–11]. The grasping pose and position of the object can be determined effectively, though the real-time performance cannot be guaranteed. In addition for objects whose shape and texture features are not obvious enough, the estimation of the optimal grasping pose is difficult to train effectively.

In another type of research, segmentation-based deep learning model is used to detect grasping regions. Zeng [12] proposed a robotic sorting system for known items and novel items in complex environments, which won the first place in the 2017 Amazon Picking Competition. The system uses multiple installed RGB-D cameras as data sources. Additionally, it utilizes deep neural networks based on ResNet-101 [13] and FCN [14] to segment pixel-wise grasping candidate regions. Then, the regression scores of the candidate regions are calculated. At last, a series of optimization strategies is proposed to select the best grasping point. The system can achieve 96.7% gripper grasping accuracy and 92.4% suction grasping accuracy, and can sort new items well. Similarly, Nguyen [15] employs two connected deep neural networks, one as an object detector and one for detection of functional regions of objects. It was proved that combining an object detector and using CRF [16] for post-optimization can achieve a high detection effect. It was experimentally verified on a full-scale humanoid robot. On this basis, Thanh-Toan Do [17] proposed AffordanceNet, which borrows the idea of the famous instance segmentation algorithm Mask R-CNN [18]. An end-to-end network structure is used to implement object classification and the segmentation of functional areas. This type of system more directly determines the grasping point or the segmentation of the candidate grasping area. As these segmentation algorithms mostly adopt a heavy network structure, they often need to balance the accuracy and real-time performance in complex scenes.

For the scene of express parcel sorting, Xing [19] proposed a robot sorting method based on a deep neural network in complex scenes, extracting more detailed candidate regions by fusing shallow feature multi-layer and final feature maps. A keypoint-based cascaded optimal sorting position detection network was proposed to detect parcels in real time. The model based on key point detection in this study was designed in three dimensions. The positions and poses of objects can be estimated through key points, but there is still room for improvement in detection accuracy. Song [20] used a pruning strategy to propose a new lightweight network model architecture, which can quickly detect the sorting position of stacked parcels in complex scenes without losing accuracy. At the same time, the proposed multi-task network model improves the detection accuracy and gets the sorting pose at the same time. However, though the above-mentioned algorithm for parcels detection obtains the sorting pose by regressing key points, the designed key points are relatively close, and the pose estimated according to the key points will be largely limited by the accuracy of the sensor. Differently from the above methods, this paper

propses a multi-modal segmentation network based on the Swin Transformer to obtain the candidate sorting area. Then, the optimal sorting position and pose are accurately estimated according to the point cloud features of the sorting area.

Recently, in order to overcome the limitations of convolution networks on local information interaction, many segmentation networks have borrowed the transformer model [21]. Robin Strudel [22] presents the semantic segmentation problem as a sequential to sequential problem. The network makes full use of context information at each stage of the model, which proves that it can produce very competitive results on all kinds of segmentation datasets. The research object of this paper is the parcels with similar color and stacked with each other, and the local information is not conducive to the detection of optimal grasping area and the prediction of occlusion degree. Therefore, this paper introduces the Swin Transformer [23,24] model into the robot visual grasping task to make full use of global information and improve the detection accuracy.

Another task of this study was to plan the optimal sorting order for multi-object scenes. In this respect, Panda [25] tries to learn the contact relationship between objects and derives the contact relationship between the target and surrounding objects through three simple interactions (support from below, support from the side and containment). The objects are sorted in order according to the support relationship. Rosman [26] proposed a method to learn spatial relationships between objects using segmented point clouds. Such methods can effectively ensure the safety and stability of grasping, but the 3D modeling of each object and the mechanical analysis of the contact surface are extremely time-consuming and cannot meet the real-time requirements of the industry. Zhang Hanbo [7] proposed a visual manipulation relationship network (VMRN) to perform real-time prediction of manipulation relationships. In the visual manipulation relationship network, object detection and end-to-end training for manipulation relationship prediction are completed. However, for the research scenario of this paper, most express parcels are boxes with similar colors and similar shapes but different sizes. They are stacked and occluded by each other. It was necessary to design a novel network structure to detect the stacking of parcels and formulate a strategy that is more suitable for judging the sorting order of stacked parcels.

In order to realize detection of the optimal sorting position and pose, and gain the sound sorting order of stacked parcels, we propose a visual sorting method based on multi-modal information fusion (VS-MF). In summary, The novelties of the proposed VS-MF are as follows:

1. A new object detection network based on multi-modal information fusion (OD-MF) is proposed to detect parcels in stacked scenes, which not only achieves high detection accuracy, but also meets the real-time requirements for robots performing sorting tasks.
2. A novel multi-modal segmentation network based on Swin Transformer (MS-ST) is proposed, which obtains the pickable region and the occlusion degree of the parcels. As is known, it is the first network to complete these two tasks at the same time. Then, the optimal sorting position and pose of parcels can be calculated.
3. A novel strategy for the optimal sorting order of parcels is proposed, which can effectively ensure the stability of robot system during the sorting process.

The subsequent parts of this paper are organized as follows: Section 2 introduces the entire framework of VS-MF. Then, OD-MF, MS-ST and the strategy for the optimal sorting order are explained in detail. Section 3 presents experiments to verify the accuracy and effectiveness of the method. The conclusion and future work are presented in Section 4.

2. Visual Sorting Method for Stacked Parcels in Complex Scenes

2.1. Overall Framework

In order to solve the sorting problem of heavily stacked parcels, VS-MF is proposed. The method can accurately obtain the sorting position and pose of express parcels, and at the same time obtain the sorting order of express parcels within the view of the camera. The overall framework of that is shown in Figure 1. First of all, the color and depth images

collected by the RGB-D camera are input into the OD-MF and gain the result of object detection quickly and accurately. Secondly, MS-ST network performs semantic segmentation for each parcel to obtain the estimation results of pickable region and occlusion degree. Thirdly, the point cloud information of the pickable region and occlusion degree is sent into the strategy for the optimal sorting order module to obtain the optimal sorting order. The optimal sorting position and pose are calculated at the same time. Finally, the sorting position, pose and order are sent to the robot system. The robot selects each parcels after coordinate transformation and trajectory planning.

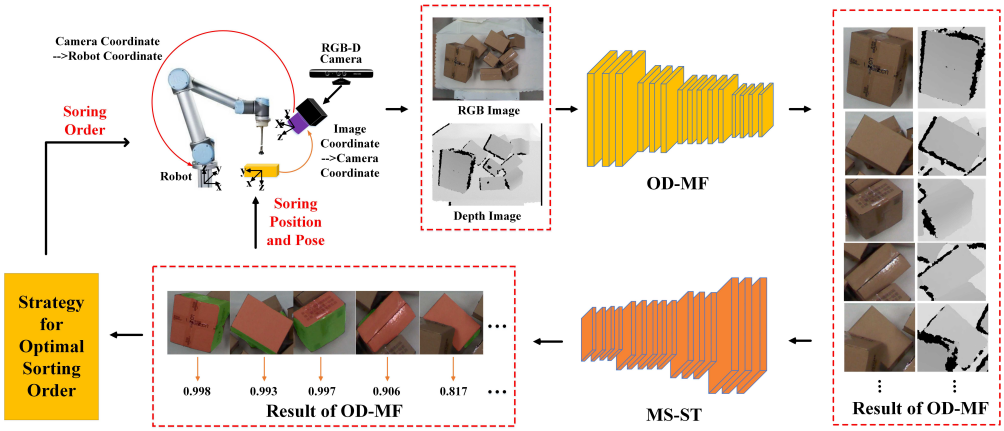

Figure 1. The overall framework of VS-MF consisting of OD-MF, MS-ST, a strategy for the optimal sorting order module and a robot system.

2.2. Object Detection Network Based on Multi-Modal Information Fusion

The situation of express parcel sorting is quite complex. Parcels are usually stacked disordered and are shielded from each other. Directly using the existing network to detect express parcels will ignore the relationships between boxes. Therefore, in order to better detect express parcels in a smart logistics sorting system, this paper proposed OD-MF. The network completes the supplementary information of the RGB and depth image, establishing a multi-object 3D geometric model. At the same time, the global gradient feature extracted by the depth information is used as a self-attention module to learn the implicit spatial feature information, thereby improving the object detection accuracy. When we locate the position and category of the target using our eyes, we often analyze the target from multiple dimensions, just like the parcels in the complex scene shown in Figure 2. The red bounding boxes represent two occluded objects, and the blue boxes represent two occluded region of those. The optimal sorting position, pose and order of the target can be located through analyzing the occlusion and dependence relationship between the targets.

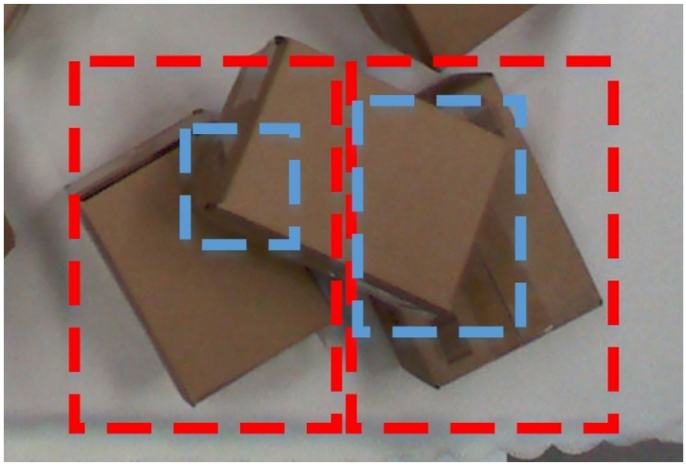

Figure 2. Stacked parcels occlude each other in complex scenarios.

Humans use multi-dimensional information fusion subconsciously. They can learn and acquire explicit and implicit knowledge autonomously. For neural networks, shallow features are generally defined as explicit knowledge, while deep features are defined as implicit knowledge. Commonly multi-dimensional information fusion involves fusing shallow features with deep features, so that the neural network can obtain multi-dimensional information. For the OD-MF proposed in this paper, the original information of RGB image and depth image is used as explicit knowledge. The global gradient information extracted from depth map is put into the network as implicit knowledge. The feature maps hidden in the neural network is also defined as implicit knowledge, which is unable to be observed directly. As shown in Figure 3, this paper proposes a network model to integrate multi-dimensional information of explicit knowledge and implicit knowledge. By learning a unified expression, the multi-modal information can be complemented, and the global gradient feature extracted from the depth information is added as a self-attention module to learn the spatial features. Subsequent experiments proved that the object detection network designed in this paper with multi-modal information fusion effectively improves the detection performance in complex scenes.

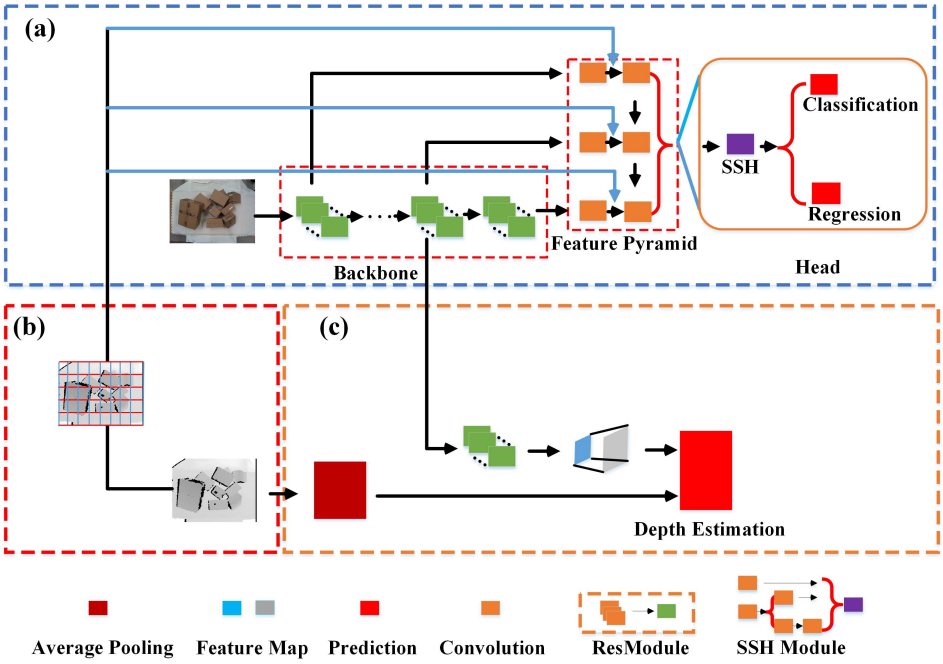

Figure 3. The overall system of OD-MF. (**a**) Modal's architecture (**b**) Depth self-attention (**c**) Depth estimation.

The model's architecture is shown in Figure 3a. Due to the real-time and accuracy requirements of the intelligent logistics sorting system in practice, this paper uses YOLOv5s [27,28] as the basic model. YOLOv5s is a smaller version of YOLO V5, which is the latest product of the YOLO architecture series. The detection accuracy of this network model is high, and the inference speed is fast. Thus, it is suitable for deployment in the architecture of our method to implement real-time detection. However, the detection performance of YOLOv5s in complex scenes with mutual occlusion is limited. Therefore, SSH [29], which is maturely applied to face detection successfully, was added to the object detection head of the model. The SSH module not only has low memory and size invariance, but also improves the receptive field of the module during training, which can effectively extract global information.

Depth self-attention is shown in of Figure 3b. The use of explicit information can accurately and efficiently guide model training. We perform channel fusion using multi-layer shallow and deep feature maps. In order to fuse the deep features of the objects, the global gradient features extracted by the depth information are added as the self-attention module

to learn the spatial feature information, extracting more detailed features. The extraction process of the global gradient feature is as follows. First, the depth map is normalized to reduce the influences of different sensors and the environment. Then, the gradients of the normalized pixels are calculated for four directions, and each pixel takes the maximum gradient value (MAG) to form an implicit information map that describes the orientation and location features of the target. Finally, the map is divided into $M \times N$ regions. The average gradient in each region is selected as the gradient feature. The global gradient feature is a vector of $M \times N$ dimensions, which contains the position and gradient of the depth map. It supervises network training at different scales, which can not only improve the recall rate of targets, but also improve the convergence speed of network training.

Depth estimation is shown in Figure 3c. Depth images are collected by different depth cameras, but the relative positions and poses of objects can be reflected by the gradient distribution of the depth images. Therefore, in the training process, multi-dimensional information of the color map and depth map is used as input to create supplementary information. However, unlike most other methods, we propose a new model structure. In the training stage, depth estimation is carried out by using depth map information as a supervision signal to achieve the fusion of explicit knowledge and implicit knowledge. This can improve the detection accuracy of the network significantly. In the inference stage, RGB maps can be used for detection alone to reduce dependence on depth cameras or depth data.

2.3. Multi-Modal Segmentation Network Based on Swin Transformer

Accurately obtaining the optimal sorting position and pose of parcels is the core of VS-MF. In order to improve the detection accuracy, we use the frequency domain signal and the depth signal as the supervision signal, and propose MS-ST. The network segments each parcel into two regions: a pickable region and an unpickable region. At the same time, the occlusion degree of the object to be sorted is estimated for the subsequent sorting order decision. As shown in Figure 4, the red region represents the pickable region; the green region represents the unpickable region; the black region represents an irrelevant region, such as the background; and the purple region represents the occlusion region of the sorting plane.

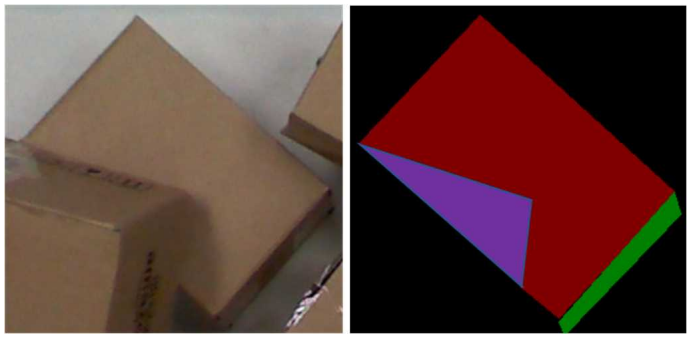

Figure 4. Schematic diagram of the pickable, unpickable, background and occlusion regions on a parcel.

Due to the various positions and poses of express parcels, in order to obtain the long-range semantic information interaction, we builds the network using Swin Transformer modules. Unlike other methods using Transformer, Swin Transformer only calculates multi-head self attention in each local windows, which has linear computational complexity with respect to image size. Thus, it can fully meet the requirements of robot sorting tasks in industrial environments. In addition, Swin Transformer adopts a hierarchical design and the shifted window approach, which improves the network's ability to understand multi-scale and long-range information at the same time. Thus, it is quite suitable for the tasks of segmentation and occlusion degree estimation in this paper. The framework of MS-ST is shown in Figure 5. During the training process, the input is the color image from the result of OD-MF. Then, the network adds multi-modal supervision branches to estimate

the depth image and spectrum image. The relationship and constraints between multiple tasks can be established in the process of training the model. The output is the segmentation result of the pickable region and unpickable region of the parcels. The occlusion degree of the target is estimated at the same time. In this way, the precision of target segmentation and the degree of occlusion can be improved.

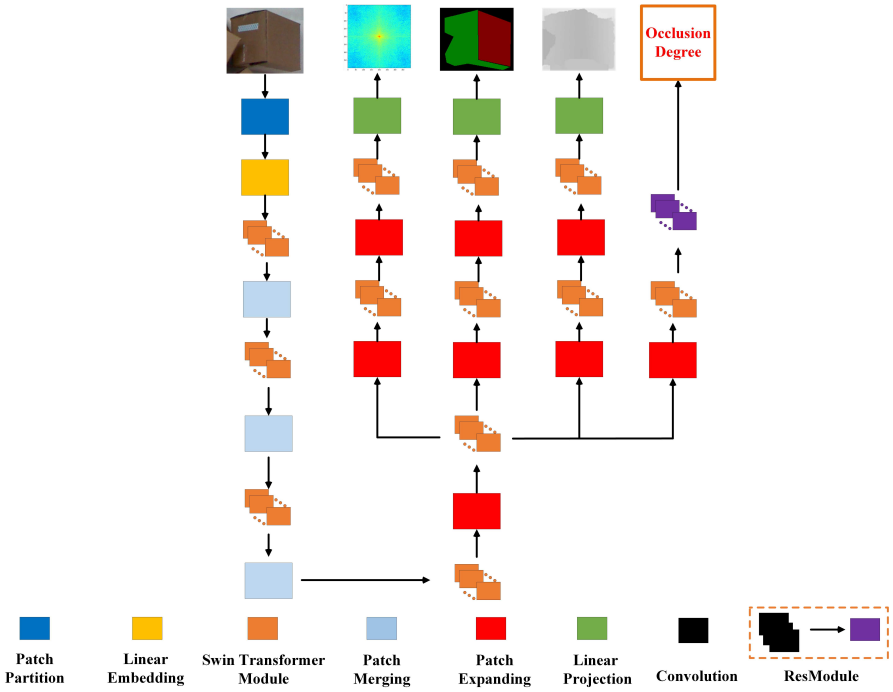

Figure 5. The overall framework of MS-ST, using multi-modal supervision branches to estimate the depth image and spectrum image.

The direct segmentation of the pickable region and the unpickable region will ignore the fine-grained information of the sorting parcel and the interactions with other parcels. Therefore, the model architecture we designed takes the multi-modal information, the Fourier-transformed spectrum, the depth image from depth estimation and the occlusion degree, as the supervision signals to improve training speed and detection accuracy.

First, the depth image is used as a supervision signal while the segmentation is performed. By increasing the depth information of the object, the rich gradient information of the object to be sorted is captured. The segmentation accuracy of sorting region detection is improved. In Figure 6, Origin donates the original image of the target to be sorted, and Depth donates the depth map. The depth image not only contains the three-dimensional locations information of parcels, but also reflects the occlusion relationships between parcels. In this paper, depth estimation is introduced into the model as a supervision signal to improve the model's ability to understand the position relationships between objects.

Second, the fine-grained frequency domain details are very important for the model to distinguish the pickable regions from the unpickable regions. As shown in Figure 6, FFT donates the Fourier spectrogram after converting the original image to grayscale, and FFT_Shift donates the spectrum after centering the image. The frequency domain images of sorting objects with different forms can reflect their edge features. Additionally, the energy domain in the spectrogram can reflect the distribution law of the sorting region. It provides more fine-grained information for the segmentation task. In this paper, spectral information is extracted as a supervision signal to enhance the representational capability of spatial details.

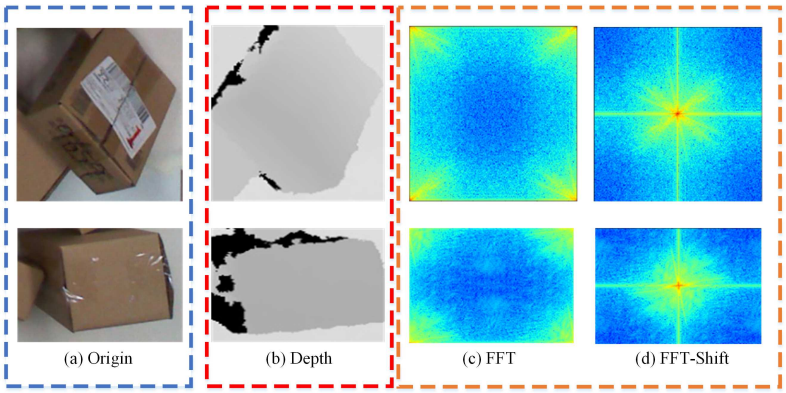

Figure 6. Multi-modal supervision signal in MS-ST.

Third, the occlusion degree of the sorting region is a key factor in judging the sorting order. The subsequent experiments demonstrate that using the occlusion degree as a supervision signal can also effectively improve the segmentation accuracy of the target regions. As shown in Figure 7. Origin donates the original image of the object to be sorted. Red represents the pickable region, and green represents the unpickable region. Mask represents the pickable and unpickable region of the object to be sorted. Additionally, Complete represents the complete region of sorting plane. Before the network training, we labeled the complete regions of stacked parcels in our dataset based on our perception. The occlusion degree is the result of dividing the number of pixels in the pickable region by the number of pixels in the complete region.

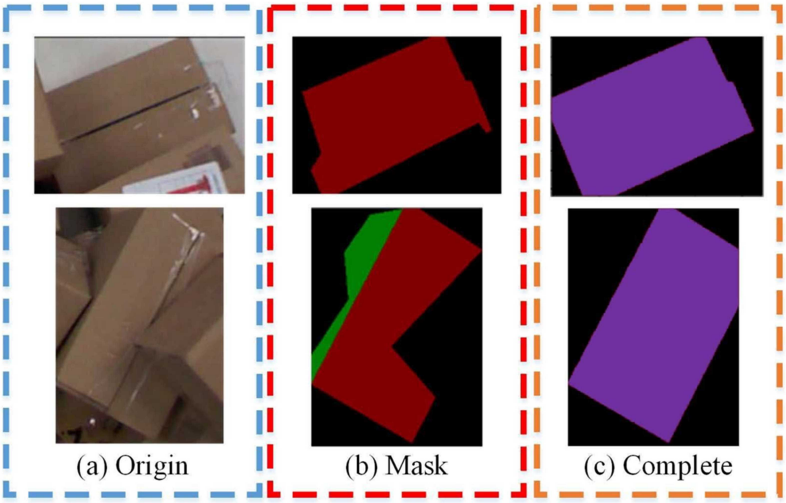

Figure 7. Sorting regions of different objects.

We calculate the geometric center of the sorting region output by MS-ST as the optimal sorting position. Then, with combining the depth information of the scene, the 3D point cloud of the sorting region is gained. We use the RANSAC algorithm [30] to get the fitting plane of the sorting region, and select the reverse sorting vector of the normal vector of the sorting plane (pointing to the inside of the parcel) to characterize the robot's sorting pose. We use RANSAC for plane fitting, because the RANSAC method is more robust to the error data of depth detection, compared with the least squares method.

2.4. Strategy for Optimal Sorting Order

After obtaining the sorting position and pose, it is necessary to judge the sorting order of multiple targets. For stacked parcels, the two other methods, RSDNN [19] and VSMT [20], design a strategy for sorting order, considering different aspects. The former one designs the strategy based on the distance between the parcel's position and robot, and

the closer parcels are sorted by robot preferentially. The latter sets the priority according to the confidence of the object detection, and preference is given to those with high confidence. We designed the sorting order judgment strategy to better satisfy the stability of the system. In the robot sorting task, the decision of the sorting order needs to consider the positional relationship between the sorting parcel and other parcels. We believe that parcels that are in an unstable state should be prioritized, so that less damage is caused to the environment and other parcels. The stability of the system can be considered from the following three aspects: the interdependence of the parcels, the heights of the parcel and the inclination of the parcels. Therefore, we formulated the following decision strategy.

First, considering the relationship between the parcels, the parcels with less occlusion are preferentially selected. Second, considering the heights of the parcels, priority is given to sorting the parcels nearest to the top. In order to speed up the calculation, we use the height of the sorting point to represent the height of the parcel. Third, considering the incline degree of the parcel, we preferentially choose a larger incline degree. We use the normal vector of the parcel sorting region to characterize the incline degree of the parcel.

The estimation of the occlusion degree is used as the quantitative score of the degree of occlusion of the parcel S_{OD}. The normal vector of the sorting plane $\boldsymbol{N} = (n_x, n_y, n_z)$ is obtained via RANSAC algorithm. According to Equation (1), the sine value of the included angle between the plane to be picked and the horizontal plane is obtained as the quantized score of the inclination degree S_{NV}. We use the depth data of the sorting point to reflect the height of the object, and then quantify the height score S_H according to Equation (2). The $h_{\max}$ and $h_{\min}$ are the maximum and minimum heights of the object sorting position in the robot's working space. In this study, they were set to 1.23 m and 0.68 m. Finally, the three scores are weighted and averaged to get the final quantitative score of the sorting priority S, as calculated in Equation (3), where K_{OD}, k_{NV} and k_H represent the weights of the three influential factors, respectively. Through experiments, we selected $K_{OD} = 2$, $k_{NV} = 1$ and $k_H = 1$, which can ensure the stability and rapidity of the system in terms of parcel picking order.

$$S_{NV} = \frac{{n_x}^2 + {n_y}^2}{\sqrt{{n_x}^2 + {n_y}^2 + {n_z}^2}} \tag{1}$$

$$S_H = \frac{h_{depth} - h_{\min}}{h_{\max} - h_{\min}} \tag{2}$$

$$S = \frac{k_{OD}S_{OD} + k_{NV}S_{NV} + k_H S_H}{k_{OD} + k_{NV} + k_H} \tag{3}$$

3. Experiment and Analysis

In the training and testing process of OD-MF and MS-ST, we used a GTX1080TI GPU with 11 GB of video memory, 32 GB of memory, and an i78700k, 6-core, 12-thread CPU in the experiment. The deep learning framework used for training was pytorch 1.7.0, and the python version was 3.6. In order to simulate the express parcel sorting scene in the real scene and verify the correctness and effectiveness of VS-MF, an experimental platform for robot sorting based on RGB-D vision was built. It mainly includes a UR-5 6-DOF robot, a RGB-D camera and a sucker actuator, as shown in Figure 8. The sorting objects were box-shaped express parcels of different sizes. The parcels were randomly stacked in the view range of the RGB-D camera.

3.1. Experiment on OD-MF

In the process of training the object detection model, we chose YOLOv5s as a baseline. First, the YOLOv5s model was trained. Additionally, the weights pre-trained on ImageNet were used to initialize the model parameters. Then, fine-tuning was performed on the parcel dataset made by ourselves. The laboratory data came from the data collected by the KinectV1 and KinectV2 cameras. The training set had 7500 images, the validation set had 2500 images and the test set had 2500 images. The hyperparameters required for model

training are shown in Table 1. During the training process, the diversity of data was increased by horizontal and vertical inversion, rotation and other enhancement methods. During training, the batch size is set to 16, and the popular SGD optimizer with momentum was set to 0.9. The weight decay of 0.0005 was used to optimize our model for back propagation.

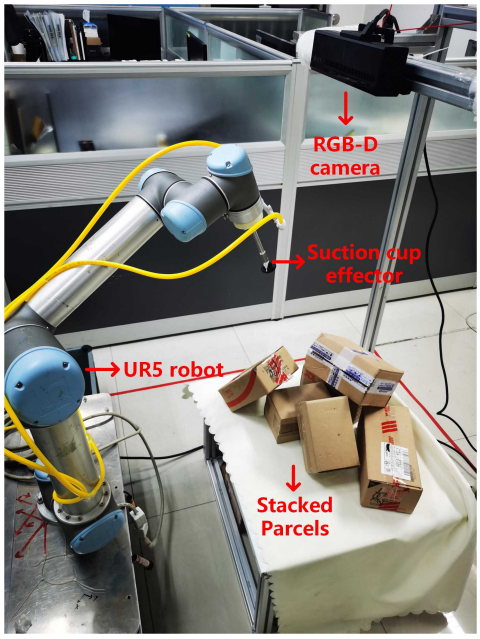

Figure 8. The robot sorting experimental platform mainly includes a UR-5 6-DOF robot, a RGB-D camera and a sucker actuator.

Table 1. The object detection network's hyperparameter setting.

Hyperparameters	Value
batch size	16
base learning rate	0.01
learning rate decay	9000, 14,000
rate of learning rate change	0.1
maximum iterations	16,000
Momentum	0.9
weight decay	0.0005

In order to verify the superiority of the OD-MF proposed in this paper, the improvements in the network by adding depth map estimation and self-attention module are compared. The experimental results are shown in Table 2 below.

Table 2. Comparison of the precision, recall and FPS of different network structures on our parcel dataset.

Model	Precision (%)	Recall (%)	FPS
YOLOv5s (Baseline)	91.75	92.33	16.13
+ MMI	98.28 (+6.53)	95.56 (+3.23)	13.89
+ Self-Attention	93.23 (+1.48)	96.05 (+3.72)	15.15
+ MMI and Self-Attention	98.57 (+6.82)	96.62 (+4.29)	13.51

It can be seen in Table 2 that compared with the original YOLOv5s model, the accuracy of parcel detection can be improved by 6.53% with multi-modal information (MMI). On the other hand, adding the self-attention module alone improved the accuracy by only 1.2%, and the recall was improved by about 4%. This proves that the method proposed

in this paper can more easily supplement and integrate explicit knowledge and implicit knowledge. At last, OD-MF with MMI and the self-attention module improved accuracy by 6.82%. That is, the network significantly improved the detection accuracy of parcels. The reason is that OD-MF can better learn the interaction information in the image by using the depth image as the estimated supervision signal and the self-attention module. Through these experiments, it can be concluded that OD-MF can effectively solve the problem of detection of stacked parcels, and improve precision and recall without losing speed.

On the other hand, it was also found in the experiment that the gradient distribution of the obtained depth map is biased and cannot be directly used as the global gradient feature of the image. The relative positions of the camera and the target shifted during the process of collecting data. To solve this problem, we used normalization and calculated the maximum gradient (MAG) in the process of acquiring global gradient characteristics of the depth map. The influences of the two improvements on the experimental results were proved by experiments. As shown in Table 3, the OD-MF using normalization and MAG could effectively improve the recall rate of objects.

Table 3. The influences of the normalization and MAG.

Model	Recall (%)	FPS
YOLOv5s + Self-Attention	96.62	13.51
+ normalization	96.85 (+0.23)	13.33
+ normalization and MAG	97.24 (+0.62)	13.15

3.2. Experiment on MS-ST

In order to accurately judge the optimal sorting position and pose, we propose the MS-ST, which estimates the occlusion degree of the target while obtaining the optimal sorting region. In the network, Fourier transform and depth estimation were used as supervision signals to complete the training of the model. As shown in Figure 9, the MS-ST algorithm accurately segmented the sound sorting region of parcels and estimated the occlusion degree of the parcel, which presented as a score range from 0 to 1. After calculating the 3D point cloud of the sorting area, the RANSAC algorithm was used to fit the sorting plane of the parcels. It can be seen that even if there are noise points coming from the sensors, the fitting algorithm can still accurately calculate the sorting plane. That is, it can prove the method, which is use for gaining the sorting pose, has strong robustness.

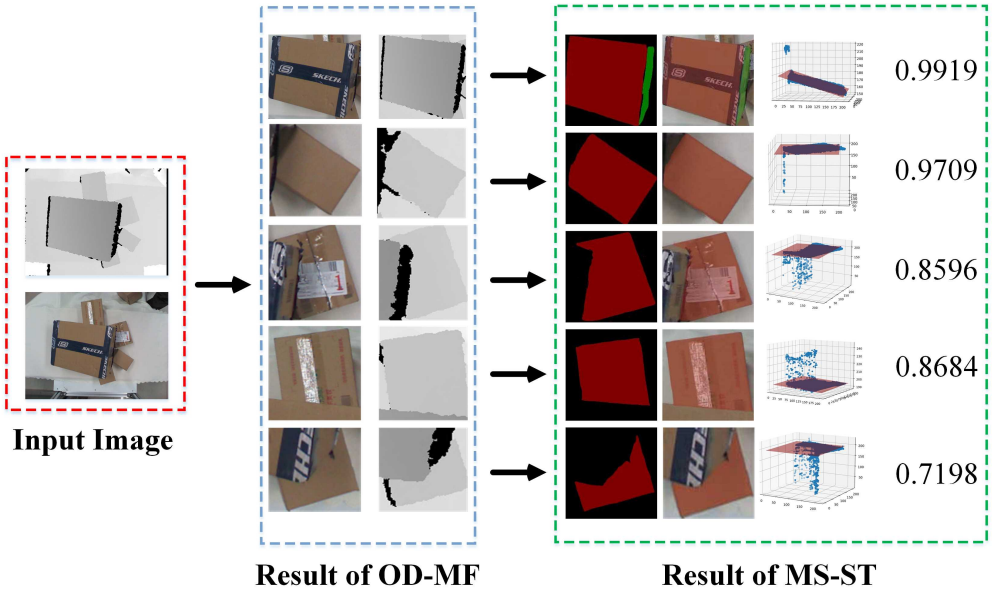

Figure 9. The result of MS-ST shows the accurate segmentation of the sorting region of parcels and the estimation of the occlusion degree.

We conducted experiments on the improvements generated by the added multi-modal supervision signal on the network detection accuracy. The experimental results are shown in Table 4 below. Adding different supervision signals improved both pixel accuracy (PA) and mean intersection over union (MIOU). As shown in Table 4, the PA of the pickable region improved 3.12% by adding a FFT signal. The edge information of the target could be effectively obtained by using the frequency domain features as the supervision signal. On the other hand, using the depth image as a supervision signal could effectively capture the rich gradient information of the objects to be sorted. Additionally, the MIOU was raised to 95.41% significantly, by adding a depth signal alone. After adding FFT and depth at the same time, the segmentation accuracy of the sorting region was improved by 7.79%. It was proven that the signals of the two modalities complete the supplementary information in different dimensions, which improved the segmentation accuracy of the network.

Table 4. The PA and MIOU on the parcel dataset with different supervision signals.

Model	PA of Pickable Region (%)	PA of Unpickable Region (%)	MIOU (%)
MS-ST (Baseline)	87.24	90.57	89.11
+FFT	90.36 (+3.12)	91.15 (+0.58)	90.57 (+1.46)
+Depth	94.15 (+6.91)	96.71 (+6.14)	95.41 (+6.30)
+FFT and Depth	95.03 (+7.79)	96.94 (+6.37)	95.97 (+6.86)

After the sorting plane was obtained, the geometric center of the segmented point cloud on the plane was calculated as the sorting position. The normal vector of the sorting plane was used as the sorting pose. In order to further prove the detection accuracy of this algorithm for sorting position and pose, a comparative experiment was conducted with two other methods [19,20] for parcel sorting. The experimental results are shown in Table 5 below.

Table 5. Position and pose detection precision comparison among MS-ST, VSMT and RSDNN on the parcel dataset.

Model	MAE of Position	MAE of Pose
RSDNN	14.97 ± 0.61	0.59 ± 0.07
VSMT	13.76 ± 0.42	0.42 ± 0.05
MS-ST	12.68 ± 0.36	0.23 ± 0.02

As can be seen in Table 5, MS-ST improved the detection accuracy to a certain extent for sorting position compared with the other two methods. In the detection of sorting poses, MS-ST greatly improved the accuracy compared with the others. In addition, according to the standard deviations obtained from multiple groups of experiments, our method has better stability, especially in pose detection. This is because the RSDNN proposed by the authors of [19] and the VSMT proposed in the paper [20] adopt a method based on key point detection, and the normal vector of sorting plane is determined by those key points. When encountering a parcel with a large degree of inclination, there will be a certain degree of error in the estimation of the key points' positions. At the same time, due to the small number of key points and the short distances between them, the normal vector is easily affected by sensor noise, which could cause large detection error of pose. Differently from those methods, MS-ST is based on the Swin Transformer architecture, and it achieves accurate segmentation of the sorting region by fusing multi-modal information. Then, at least 70% of the points on the plane, which are more than 5000, are selected as interior points of the RANSAC algorithm to estimate the sorting plane. Thus, it greatly reduces the estimation error of the normal vector of the sorting plane. At the same time, the noise points are summarized as outliers and do not participate in the calculation, which makes the method more robust.

3.3. Experiment on Strategy for Optimal Sorting Order

In the experimental scenario, we placed 6–8 parcels on the platform each time. If the robot with suction completed all sorting tasks, it was considered a successful sorting. On the other hand, the sorting strategy proposed in this paper not only needs to ensure the success rate of sorting, but also needs to ensure the stability of the system during the process. A stable sorting should do not affect the position and posture of other parcels as much as possible. We used the position change of other parcels, before and after the current sorting, to reflect the impact of this sorting on the system stability. We used the results of the two object detections to calculate the MAP as the error criterion. The MAP in Table 6 was used to characterize the effect of each sorting on the system stability. We recorded the positions of other parcels in the two frames of images before and after sorting, and calculated the IOU of bounding boxes of the corresponding parcels. The parameter overlap was set to 0.7. When the IOU exceeded the set overlap, the selection was considered to be stable. By 300 tests, the MAP in Table 6 could be obtained. The higher the MAP accuracy, the better the stability of the system during the sorting process. At last, according to the result of the experiment, we selected the appropriate weight and determined the strategy.

We chose the strategy in RSDNN and the strategy in VSMT as comparisons. Three-hundred sorting experiments were carried out for each method. It can be seen in Table 6, in terms of sorting success rate that this method performed better compared with the other methods. Its best sorting success rate was 94.3%. At the same time, our strategy achieved significantly better MAP compared with the others. When the parameters were $k_{OD} = 2$, $k_{NV} = 1$ and $k_H = 1$, the effect was the most outstanding, reaching 93.3%. We proved the stability of the system using this sorting order decision strategy. Finally, we selected those parameters to set of weights for the strategy.

Table 6. The performances of different strategies for sorting order in sorting success rate and system stability.

Model	MAP	Sorting Success Rate
RSDNN	73.4	79.5
VSMT	76.2	89.7
VS-MF ($k_{OD} = 1, k_{NV} = 1, k_H = 1$)	89.2	92.0
VS-MF ($k_{OD} = 1, k_{NV} = 1, k_H = 2$)	87.1	90.3
VS-MF ($k_{OD} = 1, k_{NV} = 2, k_H = 1$)	91.8	93.7
VS-MF ($k_{OD} = 2, k_{NV} = 1, k_H = 1$)	93.3	94.3
VS-MF ($k_{OD} = 2, k_{NV} = 2, k_H = 1$)	91.7	93.7
VS-MF ($k_{OD} = 1, k_{NV} = 2, k_H = 2$)	90.0	92.3
VS-MF ($k_{OD} = 2, k_{NV} = 1, k_H = 2$)	88.1	91.7

3.4. Robot Sorting Experiment

With the VS-MF method proposed in this paper, the sorting position, pose and order of stacked parcels within the view of the camera can be quickly and accurately obtained. We conducted comprehensive testing on the effectiveness and stability of this method in a robot system. We constructed the robotic sorting system shown in Figure 9, simulating a real sorting scenario in a logistics environment. After performing "hand-eye calibration" on the robot and RGB-D camera [30], the target sorting position and pose information were converted into the base coordinate system of the robot. Then, the robot was guided to sort the stacked parcels in the optimal order planned in our method.

Finally, 50 rounds of sorting experiments were carried out, and the number of express parcels in each round of sorting scenarios was not less than six; that is, the sorting test experiment was carried out on not less than 300 express parcels. The stacking parcel sorting process is shown in Figure 10.

(**a**) (**b**) (**c**) (**d**)

Figure 10. The sorting experiment for stacked parcels. (**a**) Sorting first parcel (**b**) Sorting second parcel (**c**) Sorting third parcel (**d**) Sorting fourth parcel.

In the individual experiments, our solution achieved more accurate results in detecting the sorting position sand poses of parcels compared with solutions proposed in [19,20]. The sorting point was closer to the central area of the parcel, and the pose was also better guaranteed to be perpendicular to the sorting plane. In the process of sorting, other methods were more likely to preferentially sort parcels in obscured or low positions due to unreasonable sorting order. This could lead to the failure of sorting and the destruction of the sorting environment due to excessive force between parcels. From the results of sorting experiments, we found that the method proposed in this paper could achieve a sorting success rate of 94.3%, which was a 4.6% improvement compared to the sorting experiment in [20], showing the robustness and stability of the system. Our solution could perform better than other methods because multi-modal information was fused in our method to detect sorting positions and poses reasonably, and a more suitable strategy for the sorting order of stacked parcels is included. We proved that VS-MF can be effectively applied to the task of robot sorting of stacked parcels in complex logistics scenarios.

4. Conclusions

In this paper, a new visual sorting method based on multi-modal information fusion was proposed for express parcels. First, an object detection method based on multi-modal information fusion was proposed, which improved the detection accuracy of stacked parcels in complex scenes. Secondly, a novel multi-modal segmentation network based on Swin Transformer was proposed, which accurately obtained the pickable regions and the occlusion degrees of the parcels. Finally, a novel strategy for the optimal sorting order of parcels was proposed to ensure the stability of robot system during the sorting process.

The experimental results show that, compared with other existing methods, the proposed method in this paper greatly improves the detection accuracy of parcels' sorting positions and poses, achieving a 4.6% improvement in success rate of sorting task in the same environment compared to another method. It fully guarantees the stability of the sorting system and the success rate of robot sorting. The whole system using the VS-MF method can significantly reduce the labor cost of the storage center and improve the automation degree of the storage center. Since the proposed method can only guarantee the sorting of square parcels, the sorting methods for other shapes of parcels need further research. We found it will cause some unsuccessful sorting in the experiment when the sorting positions appeared on the tapes or labels of parcels. This is another area for improvement of our method we should focus on in future research. In the future, we will further optimize the proposed method to implement efficient sorting for more types of express parcels. At the same time, we will further improve the detection accuracy and speed of the method, and apply it to an actual logistics sorting center to solve the problem of automatic sorting for stacked parcels.

Author Contributions: Conceptualization, X.L., G.W. and S.H.; methodology, S.H. and X.L.; software, S.H.; validation, G.W.; writing—original draft preparation, S.H.; writing—review and editing, X.L. and G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sansoni, G.; Bellandi, P.; Leoni, F.; Docchio, F. Optoranger: A 3D pattern matching method for bin picking applications. *Opt. Lasers Eng.* **2014**, *54*, 222–231. [CrossRef]
2. Wu, C.H.; Jiang, S.Y.; Song, K.T. CAD-based pose estimation for random bin-picking of multiple objects using a RGB-D camera. In Proceedings of the 2015 15th International Conference on Control, Automation and Systems (ICCAS), Busan, Korea, 13–16 October 2015.
3. Song, K.T.; Wu, C.H.; Jiang, S.Y. CAD-based Pose Estimation Design for Random Bin Picking using a RGB-D Camera. *J. Intell. Robot. Syst.* **2017**, *87*, 455–470. [CrossRef]
4. Viereck, U.; Pas, A.T.; Saenko, K.; Platt, R. Learning a visuomotor controller for real world robotic grasping using simulated depth images. In Proceedings of the Conference on Robot Learning (CoRL), Mountain View, CA, USA, 13–15 November 2017.
5. Du, X.; Cai, Y.; Lu, T.; Wang, S.; Yan, Z. A robotic grasping method based on deep learning. *Robot* **2017**, *39*, 820–837.
6. Kumra, S.; Kanan, C. Robotic grasp detection using deep convolutional neural networks. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017.
7. Zhang, H.; Lan, X.; Zhou, X.; Tian, Z.; Zhang, Y. Visual Manipulation Relationship Network for Autonomous Robotics. In Proceedings of the 2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids), Beijing, China, 6–9 November 2018.
8. Sundermeyer, M.; Marton, Z.; Durner, M.; Triebel, R. Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. *Int. J. Comput. Vis.* **2020**, *128*, 714–729. [CrossRef]
9. Hodan, T.; Haluza, P.; Obdržálek, Š.; Matas, J.; Lourakis, M.; Zabulis, X. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) IEEE, Santa Rosa, CA, USA, 24–31 March 2017; pp. 880–888.
10. Park, K.; Patten, T.; Vincze, M. Pix2pose: Pixelwise coordinate regression of objects for 6d pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 7668–7677.
11. Wei, J.; Liu, H.; Yan, G.; Sun, F. Robotic grasping recognition using multi-modal deep extreme learning machine. *Multidimens. Syst. Signal Process.* **2017**, *28*, 817–833. [CrossRef]
12. Andy, Z.; Shuran, S.; Kuan-Ting, Y.; Elliott, D.; Alberto, R. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016
14. Long, J.; Evan, S.; Trevor, D. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,Boston, MA, USA, 8–10 June 2015.
15. Nguyen, A.; Kanoulas, D.; Caldwell, D.G.; Tsagarakis, N.G. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017.
16. Lafferty, J.; Andrew, M.; Pereira, F.C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001.
17. Do, T.; Anh, N.; Ian, R. Affordancenet: An end-to-end deep learning approach for object affordance detection. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018.
18. Kaiming, H.; Georgia, G.; Piotr, D. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
19. Han, X.; Liu, X.-P. Robotic sorting method in complex scene based on deep neural network. *J. Beijing Univ. Posts Telecommun.* **2019**, *42*, 22–28.
20. Han, S.; Liu, X.-P.; Han, X.; Wang, G.; Wu, S. Visual sorting of express parcels based on multi-task deep learning. *Sensors* **2020**, *20*, 6785. [CrossRef] [PubMed]
21. Vaswani, A.; Shazeer, N.; Parmar, N. Attention is all you need. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 5998–6008.
22. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 March 2021.

23. Alexey, D.; Lucas, B.; Alexander, K.; Dirk, W. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the ICLR, Vienna, Austria, 3–7 April 2021.
24. Liu, Z.; Lin, Y.; Cao, Y. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 March 2021.
25. Panda, S.; Hafez, A.; Jawahar, C. Learning support order for manipulation in clutter. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013.
26. Rosman, B.; Ramamoorthy, S. Learning spatial relationships between objects. *Int. J. Robot. Res.* **2011**, *30*, 1328–1342. [CrossRef]
27. Yan, B.; Fan, P.; Lei, X.; Liu, Z.; Yang, F. A real-time apple targets detection method for picking robot based on improved YOLOv5. *Remote Sens.* **2021**, *13*, 1619. [CrossRef]
28. Najibi, M.; Samangouei, P.; Chellappa, R.; Davis, L.S. Ssh: Single stage headless face detector. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 21 October 2017.
29. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
30. Horaud, R.; Fadi, D. Hand-eye calibration. *Int. J. Robot. Res.* **1995**, *14*, 195–210. [CrossRef]