Article

HTC-Grasp: A Hybrid Transformer-CNN Architecture for Robotic Grasp Detection

Qiang Zhang [1], Jiangwei Zhu [1], Xueying Sun [1,*] and Mingmin Liu [2]

[1] School of Automation, Jiangsu University of Science and Technology, No. 666 Changhui Road, Zhenjiang 212100, Jiangsu, P.R. China;
[2] Central Research Institute, SIASUN Robot & Automation Co., LTD., NO.16 Jinhui Street, Hunnan District, Shenyang 110168, P.R. China;
* Correspondence: sunxueying@just.edu.cn;

Abstract: Accurately detecting suitable grasp areas for unknown objects through visual information remains a challenging task. Drawing inspiration from the success of the Transformer in vision detection, the hybrid Transformer-CNN architecture for robotic grasp detection, known as HTC-Grasp, is developed to improve the accuracy of grasping unknown objects. The architecture employs an external attention-based hierarchical Transformer as an encoder to effectively capture global context and correlation features across the entire dataset. Furthermore, a channel-wise attention-based CNN decoder is presented to adaptively adjust the weight of the channels in the approach, resulting in more efficient feature aggregation. The proposed method is validated on the Cornell datasets and Jacquard datasets, achieving an image-wise detection accuracy of 98.3% and 95.8% on each dataset, respectively. Additionally, the object-wise detection accuracy of 96.9% and 92.4% on the same datasets are achieved based on the method. A physical experiment is also performed using the Elite 6Dof robot, with a grasping accuracy rate of 93.3%, demonstrating the proposed method's ability to grasp unknown objects in real scenarios. The results of this study indict that the proposed method outperforms other state-of-the-art methods.

Keywords: Robotic Grasp; Transformer; attentional mechanism

1. Introduction

In the nearest decade, the advancement of artificial intelligence has made smart robots increasingly important in industries such as smart factories and healthcare [1,2]. Among the tasks performed by these robots, grasping objects is a fundamental ability that enables them to carry out more complex operations [3,4]. Vision-based automated grasping, where the robot uses visual sensors to identify the best gripping position for an object, is crucial for their intelligence and automation. However, despite the advancements in the field, most of the current methods are still limited to models of known objects or trained for known scenes, making the task of grasping unknown objects with high accuracy a significant challenge [5].

Currently, most grasp detection methods for vision robots rely on Convolutional Neural Networks (CNNs) [6–10]. Despite their popularity, CNNs have limitations in handling grasping tasks. They are designed to process local information through their small convolutional kernels and have difficulty capturing global information due to limited filter channels and convolution kernel sizes. The convolutional computation method used by CNNs also makes it challenging to capture long-distance dependency information during information processing.

The Transformer architecture has seen great success in the field of vision lately [11,12]. The Transformer's self-attention mechanism provides a more comprehensive understanding of image features compared to CNNs. The Transformer has the ability to effectively

 doi:10.20944/preprints202302.0382.v2

capture global information through its self-attentive mechanism, which makes it a more representative model.

While the self-attention mechanism of the Transformer is useful for capturing information within a single sample, it may not fully leverage the potential connections between different samples. In the task of grasping, the features of the grasping target are often correlated, and the background features of similar scenes are consistent. Thus, considering the potential connections between different samples can lead to a more robust feature representation. To address this challenge, the proposed HTC-Grasp incorporates external attention into the transformer block to enhance the representation of correlations between different images.

Moreover, the multi-scale feature fusion mechanism introduces a significant amount of noisy features, which can negatively impact grasp detection performance. To mitigate this issue and improve the role of effective features, the proposed framework incorporates a residual connection-based channel attention block in the decoder. This approach enables efficient learning of discriminative channel-wise features.

The original contributions of this research are outlined below:

1. A highly robust hierarchical Transformer-CNN architecture for robot grasp detection is developed that integrates local and global features.

2. In this architecture, the external attention based hierarchical Transformer is proposed as an encoder to effectively capture global context and the correlation features across the whole data. Furthermore, a channel-wise attention based CNN decoder is provided to adaptively adjust the weight of the channels, thus providing a more efficient feature aggregation.

3. Extensive experiments are conducted on both public datasets and real-world object grasp task to validate the performance of the HTC-grasp approach. The results, both qualitative and quantitative, show that the HTC-grasp outperforms state-of-the-art methods and can detect stable grasps with high accuracy.

The proposed approach can improve the environmental adaptability of robots and can be applied in logistics centers for picking up goods, automated garbage sorting, robotic assistance for household tasks, etc.

2. Related Works

The representation of object grasping is crucial for robot grasp detection. Jiang et al. [13] proposed an efficient method that describes the grasping position using a rectangular representation, using a 5-dimensional vector to describe the position, height, width, and rotation angle of the grasp in the image. Morrison et al. [14] introduced a grasp location description method, which gives the gripping position and posture by predicting the gripping quality of each pixel. These two models are widely used in robot grasp detection tasks.

Current grasp detection models can be broadly categorized into two different types: cascade methods and one-stage methods. Cascade approaches perform the entire grasp prediction process in stages, including the extraction of target features, generation of candidate regions, and evaluation of the optimal gripping position. Lenz et al. [15] created the Cornell dataset and proposed a two-stage cascade detection model to learn this five-dimensional grasp. The first stage uses a neural network to extract grasp prediction features. The second phase refines the predicted grasp parameters to output the optimal grasp location. Zhou et al. [16] presented a model that predicts multiple grasping poses using an oriented anchor box. Zhang et al. [17] introduced ROI-GD approach, which uses ROI features to detect grasps instead of the whole image. Laili et al. [18] presented a region-based approach to locate grasping point pairs. A consistency-based method is used to train the grasp detector with less labelled training data.

In the last few years, the development of one-stage detection approaches for object grasping has gained popularity due to their simple and efficient structure. The one-stage approach trains a grasp detection model to directly output the grabbing location. Previous

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

works, such as Redmon et al. [19], used AlexNet to directly process the input image and predict the grasp location. Kumra et al. [20,21] built a grasp network based on ResNet that extracts features from RGB and depth images to output both classification and regression results for the optimal grasp location. Mahler et al. [22] put forward a grasp quality evaluation network using image segmentation and a corresponding point cloud for grasp prediction. Morrison et al. [14] used convolutional layers for encoding and decoding to perform pixel-level grasp prediction of feature maps. Yu et al. [23] presented a U-Net like architecture with channel attention modules to better utilize features. Wu et al. [24] introduced an anchor-free grasp detector which employs a fully convolutional network. This approach frames grasp detection as two separate tasks: regression of the closest horizontal or vertical rectangle, and classification of the grasp angle. The CNN based grasping target detection algorithms discussed above have made significant progress. However, the use of convolutional kernels, which primarily focus on local spatial information, can limit the ability to capture global information correlations, potentially hindering further improvements in detection accuracy.

Recently, the transformer has gained traction in computer vision because of its ability to capture global information, overcoming the limitations of CNN models. The transformer has shown excellent performance in vision tasks such as object detection, classification and tracking [25,26] through its self-attention mechanism and pyramid-like structure. In 2022, Wang et al. [27] used the SWIN Transformer to extract features with impressive results. The self-attentive mechanism, while useful, has a limitation in that it focuses only on the information contained within a single sample and ignores the connection cross the whole dataset, which may negatively impact the robustness of feature representation.

To further improve the accuracy of robot grasp detection, a hybrid Transformer-CNN architecture for robotic grasp detection is proposed in this article. With the help of external attention mechanism, long distance spatial correlation can be learned. Global context between data samples can improve the robustness of the feature implicitly. In order to aggregate the extracted multiscale features, up-sampling and skip connection are introduced to the decoder. The channel attention modules based on SE-block further assign adaptive weights to each feature channel to enhance the feature representation.

3. Method

3.1 Grasp Task Representation

The vision grasping tasks typically involve collecting visual images of the target object using sensors such as RGBD cameras. These images are processed by a model to determine the optimal grasp position. When the robot is equipped with parallel grippers, the grasping parameters p can be represented as a 5-dimensional tuple.

$$p = \{x, y, \theta, w, h\} \tag{1}$$

where (x, y) represents the horizontal and vertical coordinates of the center of the grasp box, (w, h) represents the width and height of the grasp box, and θ is the rotation angle of the gripper box with respect to the horizontal axis.

An alternative representation for high-precision, real-time robot grasping was introduced in [14]. In this representation, the grasp is redefined for 2DoF robotic grasping tasks as follows:

$$P = \{Q_g, \Theta_g, W_g\} \in \mathbb{R}^{3\times W\times \mathrm{H}} \tag{2}$$

Where P is a 3-dimensional tensor. The first dimension, Q_g represents the grasping quality of each point in the image; the second dimension, Θ_g, denotes the orientation angle of the finger gripper; and the third dimension, W_g, represents the opening width of the finger gripper. Each pixel, with a specific width $W_{g_{i,j}}$ and angle $\Theta_{g_{i,j}}$, corresponds to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

the width and orientation angle of the finger gripper at that particular position. W and H represent the length and width of the feature map.

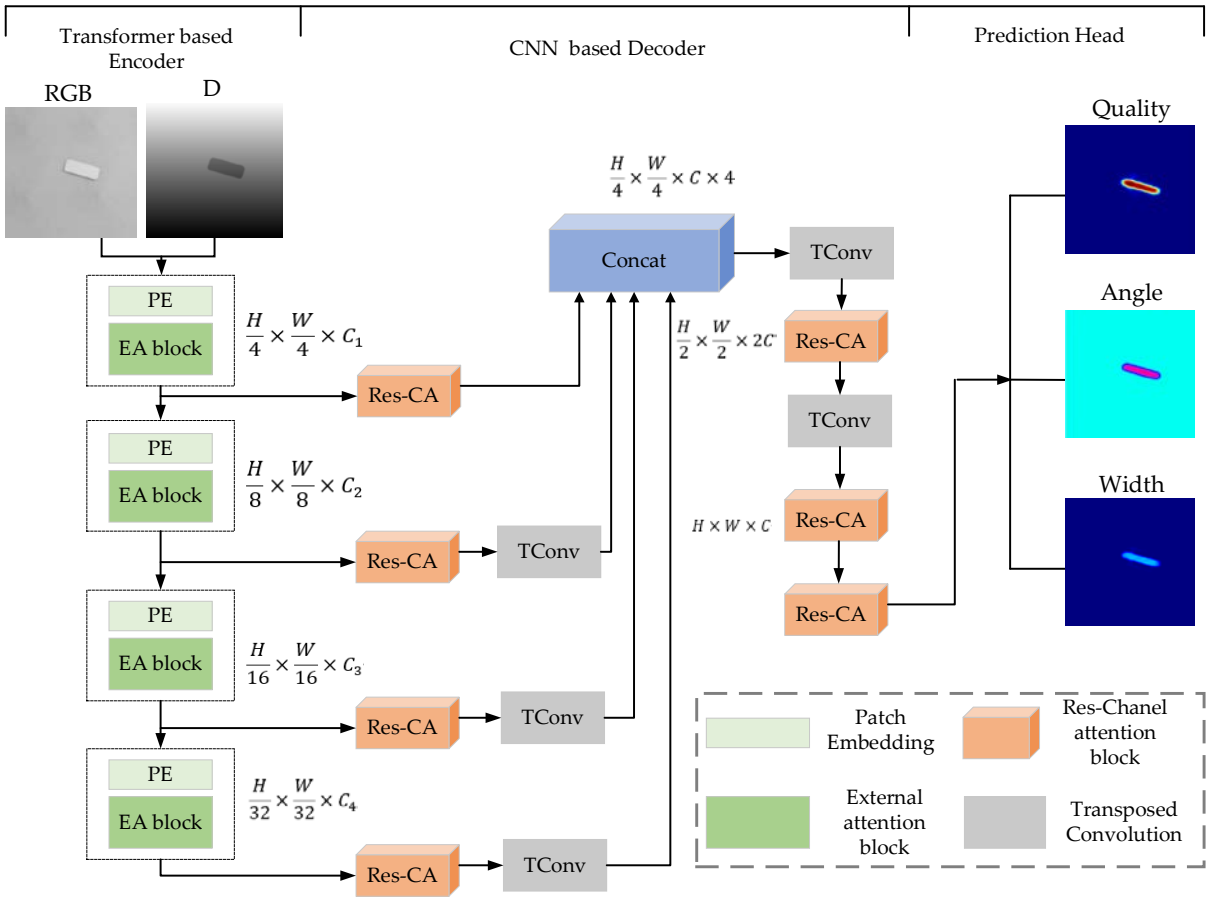


Figure 1. Overall network architecture of HTC-Grasp. W and H represent the width and height of the feature map, respectively. C represents channel of the feature map.

3.2 Grasp Overview

The overall architecture of HTC-Grasp is illustrated in Figure 1. The architecture of the HTC-Grasp consists of three parts: the Transformer based encoder, the CNN based decoder, and the prediction head. The encoder is built using hierarchical transformers with a pyramidal structure to extract multi-scale features. The decoder, made up of transposed convolution layers with res-channel attention blocks, fuses the previously obtained multi-scale features. Finally, the fused features are used by four sub-task networks to predict grasp heatmaps, including the map of quality score, the angle (in $\sin 2\theta$ and $\cos 2\theta$ form) and the width.

The specific process is as follows. Using an RGB-D image as an input, the size of which is $H \times W \times 4$, it is first divided into blocks with 4×4 pixels for each block. These blocks are then used as inputs to the transformer blocks, which output multi-level feature images with resolutions of $\{1/4, 1/8, 1/16, 1/32\}$ of the original image. These multi-level features are then up-sampled to $56 \times 56 \times C_i$ based on the transparent convolutional layers. By employing channel-wise concatenation, the four level of features are aggregated. To make the width and height of the features consistent with the original image, the two deconvolution modules further up-sample the features. Two Res-channel attention blocks are also employed to improve the robustness of the features at the end of the decoder. The prediction head then can predict quality, angle and width heatmaps. The details of the proposed encoder-decoder design are explored in the subsequent sections.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

3.3 HTC-grasp Architecture

3.3.1. Hierarchical Transformer Encoder

A layered Transformer architecture is adopted to facilitate the generation of multi-scale feature maps in this article. Multi-scale features generated by the hierarchical Transformer encoder enhance the performance of the model. The feature encoder of the proposed method comprises four stages, each designed to generate feature maps at a different scale. The structure of each stage is similar and consists of a transformer block and a patch embedding layer.

To be more specific, an image with resolution $H \times W \times 4$ is fed into Patch Embedding stages to get a hierarchical feature image F_i with the resolution of $\frac{H}{2^{i+1}}, \frac{W}{2^{i+1}}, C_i$, where i ranges from 1 to 4. Considering that uniform partitioning will make the obtained patches have no overlapping parts and weaken the connection between patches, overlapping parts between each patch in the partitioning are preserved intentionally. Then the images patches are fed into the encoder to obtain multi-scale features.

The Transformer blocks are used to extract features. Self-attention is the most important module of each Transformer block. The original self-attention mechanism generates three matrices: the query matrix $Q \in \mathbb{R}^{N \times d_k}$, the key matrix $K \in \mathbb{R}^{N \times d_k}$, and the value matrix $V \in \mathbb{R}^{N \times d_v}$. Here, N represents the number of patches, and d_k and d_v signify the feature dimensions of Q and K, and V, respectively. The calculation of self-attention is as follows:

$$Attention = softmax(\frac{QK^T}{\sqrt{d}})V \tag{3}$$

The high computational complexity of the self-attention presents a significant drawback to the real-time applications. Additionally, self-attention can only model correlations within individual samples, ignoring the correlations across the entire dataset. To overcome these limitations, the multi-head external attention (MEA) [28] mechanism is introduced into the Transformer blocks.

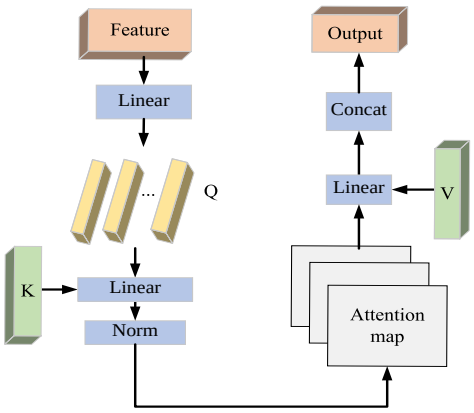
MEA mechanism is incorporated to improve the efficiency of the transformer layer. This mechanism is represented by the following equation:

$$h_i = ExternalAttention(F_i, M_k, M_v) \tag{4}$$

$$F_{out} = MultiHead(F, M_k, M_v) \tag{5}$$

$$F_{out} = Concat(h_1, \dots, h_H)W_o \tag{6}$$

where h_i stands for the ith multihead, Hd symbolizes the total number of multiheads, and W_o is a linear transformation matrix that has equal output and input dimensions. The structure of MEA is shown in Figure 2.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

Figure 2. The architecture of external attention block

3.3.2. Grasp Decoder

The grasp decoder is designed with a combination of convolutional layers and Res-channel attention blocks. As can be seen in Figure 3, the Res-channel attention block is a combination of a ResNet block and a channel attention block. The ResNet block is made up of 3 convolutional blocks. The kernel sizes of these 3 convolutional blocks are set o 1×1, 3×3 and 1×1, respectively. The channel attention block, on the other hand, utilizes global average pooling (GAP) to reduce the number of participants contained in the features. This block is then composed of 2 fully connected layers and 1 ReLU unit, which utilizes global information to selectively emphasize important features and reduce the emphasis on less relevant features.

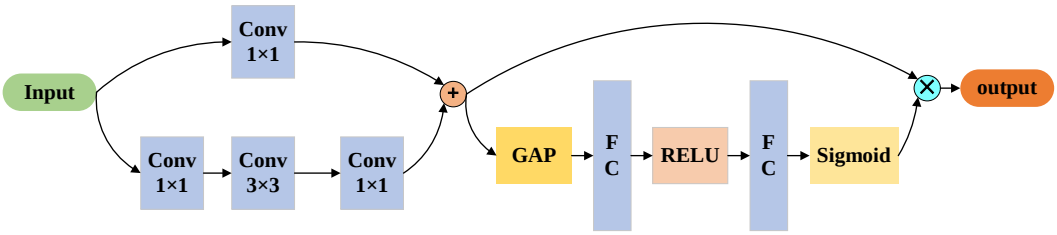


Figure 3. Res-Channel attention block

Specifically, the decoder involves three key steps. To begin with, the multilevel features F_i from the encoder are fed through the up-sample block, which increases the resolution to $1/4 \times 224 \times 224$, and then these features are concatenated. Next, a CNN layer is utilized to merge the resulting features, and this is followed by two upsampling layers that increase the resolution to 224×224. Finally, the fused features are utilized to make predictions regarding the grasp heatmaps.

3.3.3. Loss Function

In this study, the task of robot grasp detection is defined as a regression problem and the smooth $L1$ loss function is adopted as the optimization objective. The advantage of this loss function is that it is robust to outliers and can provide stability during training.

$$L_{reg}(\hat{T}_k, T_k) = \sum_{k \in \{q, \sin 2\theta, \cos 2\theta, w\}} Smooth(\hat{T}_k - T_k) \tag{7}$$

The Smooth $L1$ loss is defined as follows

$$Smooth_{L1}(x) = \begin{cases} 0.5x^2, & if|x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \tag{8}$$

In this work, the predicted grasping parameters $\hat{T}_k$ and the ground truth T_k are defined as follows: q represents the grasping quality, θ stands for the rotation angle of the gripper, and w represents the opening width.

4. Experiments and Results

4.1 Dataset

In this work, experiments are conducted on the Cornell and the Jacquard datasets to fully validate the HTC-grasp method.

(a) Datasets

The Cornell dataset is a dataset for robot grasp detection, which includes 240 distinct objects. It consists of 885 color images and 885 depth images. To ensure the best results from the transformer structure, which requires a substantial amount of data, data augmentation approaches like image rotation, scaling, and random cropping are applied to the Cornell dataset in the experiments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

The Jacquard dataset consists of 54485 diverse scenes for 11619 different objects. It provides RGB images, 3D point cloud data, and grasp annotations for each scene. Given the massive size of the Jacquard dataset, no data transformations are performed on it in this work.

(b) Implementation details

In this ariticle, the model was constructed using the Pytorch framework on the Ubuntu 20.04 platform. For training, an NVIDIA RTX 3090Ti GPU and an Intel Core i9-12900K CPU are utilized. In the data augmentation process for the Cornell dataset, each 640x480 image undergoes rotation, scaling, and random cropping, resulting in an image of size 224x224. During each training step, image samples were randomly selected from the training dataset, with 200 batches of size 32 in each epoch, and 100 epochs are trained in total. AdamW is employed as the optimizer and the initial learning rate is set to 0.0001.

HTC-Grasp is parameterized with the following configuration. The channel numbers for stages 1 to 4 are set to $C_1 = 32, C_2 = 64, C_3 = 128, C_4 = 256$, respectively. Head numbers in the self-attention layer for each block is set to 1, 2, 4, and 8, respectively. The number of encoder layers in stages 1 to 4 is set to $L_1 = L_2 = L_3 = L_4 = 256$. The number of channels in the decoder layer is set to C=256.

In this work, each dataset is structured into two parts, with 90% used for training and 10% for testing. To evaluate the performance of the method, both image-wise and object-wise detection accuracy were used. Image-wise split randomly assigns the entire dataset into a training set and a test set, to assess the network's ability to generalize to previously seen objects when they appear in new positions and orientations. Object-wise split, on the other hand, divides the dataset based on object instances, ensuring that objects in the test set do not show up in the training process, thereby testing the network's ability to generalize to unknown objects.

(c) Evaluation index

The predicted grasping box is considered correct if it meets the following two criteria.

(1) The angle difference between the predicted results and the ground truth must be within 30°

(2) The IOU index, which is defined in equation (9), must be greater than 0.25.

$$IOU(R^*, R) = \frac{|R^* \cap R|}{|R^* \cup R|} \tag{9}$$

4.2 Comparison Studies

To evaluate the performance of HTC-grasp against other grasp detection methods, the same evaluation metric is used to compare the results on both the Cornell and Jacquard datasets.

Table 1. The comparison results on Cornell Dataset

Authors	Method	Input	Accuracy (%)		Time (ms)
			IW	OW	
Lenz [15]	SAE	RGB-D	73.9	75.6	1350
Redmon [19]	AlexNet	RGB-D	88	87.1	76
Kumra [20]	ResNet-50x2	RGB-D	89.2	88.9	103
Morrision [14]	GG-CNN	D	73	69	19
Chu [29]	ResNet-50	RGB-D	96	96.1	120
Asif [8]	GraspNet	RGB-D	90.2	90.6	24
Kumra [21]	GR-CNN	RGB-D	97.7	96.6	20
Wang [27]	TF-Grasp	RGB-D	97.99	96.7	41.6
Ours	HTC-Grasp	RGB-D	**98.3**	**96.9**	**5.4**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

The comparison study starts with the evaluation on the Cornell dataset. The grasp position can be determined using the quality heatmap, with the best grasp position being the pixel with the highest quality score and the grasping box being determined by the angle and width corresponding to the best grasp position. Figure 4 presents the results of GR-CNN , TF-Grasp [27] and the proposed HTC-Grasp for unseen objects on the Cornell Dataset. The statistical results indicate that HTC-Grasp has a higher grasp quality as compared to the GR-CNN and TF-Grasp methods.

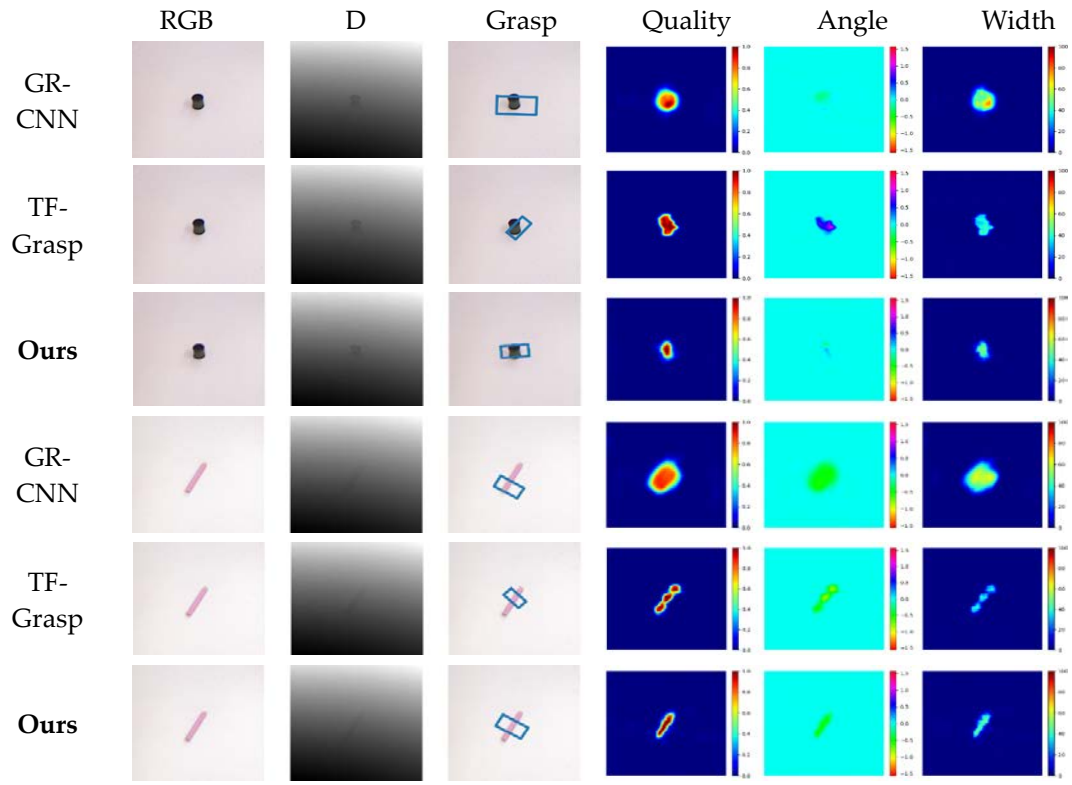


Figure 4. Comparison of predicted heatmaps on Cornell Dataset

For the classical method experimental results presented in Table 1, the data reported in their original paper are selected. Table 1 illustrates the performance of HTC-Grasp compared to existing algorithms on the Cornell dataset. HTC-Grasp sur-passes other algorithms with accuracy rates of 98.3% and 96.9% on Image-wise split (IW) and Object-wise split (OW) test, respectively. Furthermore, the proposed model, utilizing the NVIDIA RTX 3090Ti GPU, processes a single frame in approximately 5.4 ms, fulfilling the requirement for real-time processing.

Table 2. The statistical results on Jacquard Dataset

Authors	Method	Input	Accuracy (%)	
			IW	OW
Morrision [14]	GG-CNN	D	84	-
Kumra [21]	GR-CNN	RGB-D	92.6	87.7
Wang [27]	TF-Grasp	RGB-D	94.6	-
Ours	HTC-Grasp	RGB-D	**95.8**	**92.4**

Comparative experiments using the Jacquard dataset are also conducted. Figure 5 dis-plays some examples of the predicted heatmaps and predicted grasps of GR-CNN, TF-Grasp, and HTC-Grasp. The results indicate that HTC-Grasp exhibits a higher

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

grasping quality compared to GR-CNN and TF-Grasp methods. Table 2 presents the performance of HTC-Grasp on the Jacquard dataset in comparison to several classic algorithms. HTC-Grasp outperformed the other algorithms with an accuracy of 95.8% and 92.4% for Image-wise split (IW) and Object-wise split (OW) test on the Jacquard dataset.

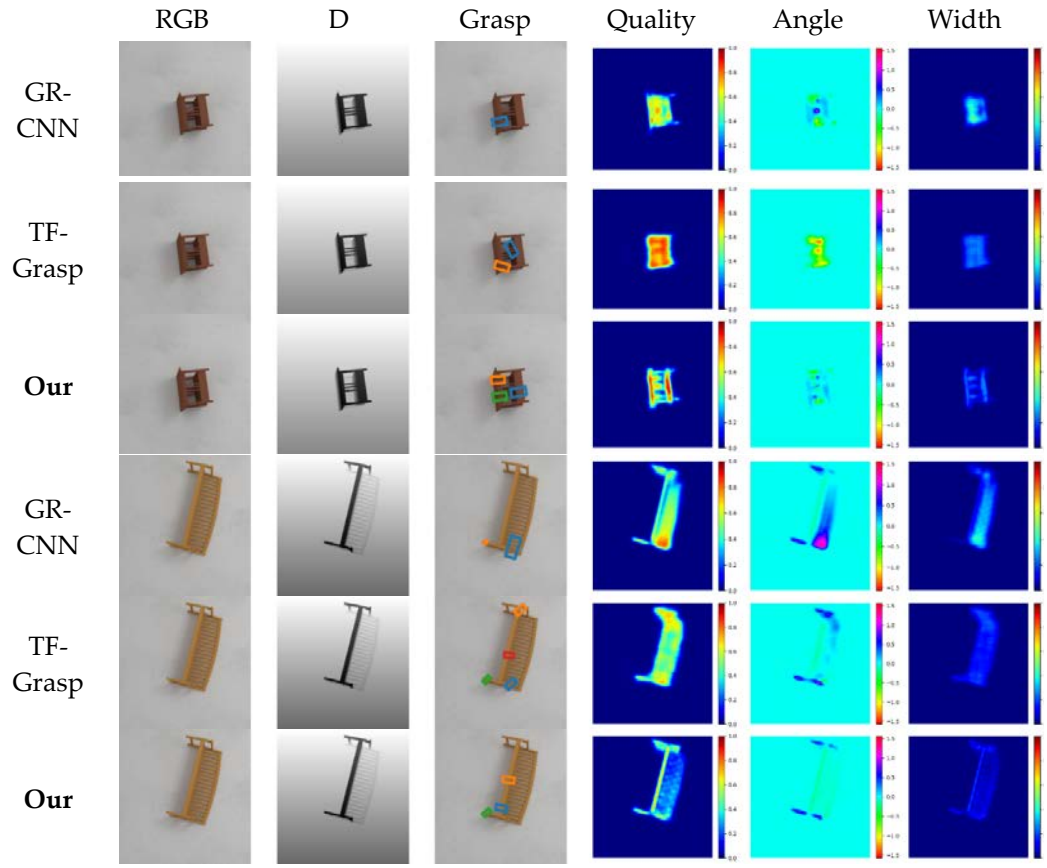


Figure 5. Comparison of predicted heatmaps on Jacquard Dataset

Qualitative comparison results for the Cornell and Jacquard datasets are demonstrated in Figures 4 and 5. It can be observed that:

1) As shown in the first and third rows of Figures 4 and 5, the GR-CNN method which is solely based on CNNs has a low prediction quality in the central region of easily grasped objects. The background predictions by GRCNN are close to the actual grasping poses, indicating that grasp pose detection is vulnerable to environmental interference. This is due to the absence of an attention mechanism in the GR-CNN network, leading to its poor performance.

2) In comparison to the Transformer-based TF-Grasp model, the proposed HTC-Grasp provides more precise predictions of grasp quality and retains more de-tailed shape information. This is achieved by incorporating an external attention mechanism in the encoder module, which enhances the network's capability to encode global context and differentiate semantics. Furthermore, a Residual Channel attention module is introduced into the decoder module, which allows the network to learn and determine the significance of each feature channel, thereby improving the utilization of valuable features and reducing the impact of redundant features.

Experimental results demonstrate that the HTC-grasp approach can accurately identify suitable grasp locations and effectively differentiate graspable regions with a high level of confidence. As seen in the third and sixth rows of Figure 4, the center of the object is highlighted with a high score close to 1, while the edges of the object are marked with a lower score. Similarly, in the third and sixth rows of Figure 5, the protruding parts of the object that are easily graspable are precisely marked with a high score, and the model

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

effectively captures both global information and fine-grained features such as the exact lo-cation and shape of the object.

To further evaluate HTC-grasp's efficiency, experiments using a test set of images captured by ourselves without additional training are conducted. The results shown in Figure 6 indicate that the proposed method can accurately identify grasp regions in an unseen real-world environment.

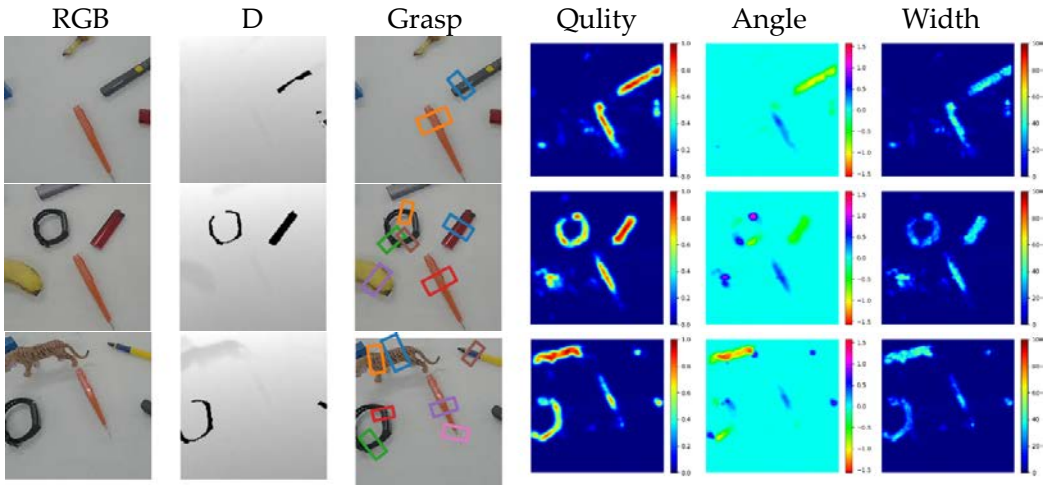


Figure 6. Test result of HTC-Grasp in the real-world multiple objects environment

4.3 Ablation Studies

To validate the impact of external attention and channel attention on the proposed grasp detection model, experiments on the same Cornell and Jacquard datasets are conducted. HTC-Grasp model is compared to versions without external attention and channel attention, respectively.

The results are shown in Table 3 and indicate that incorporating external attention in the encoder and channel attention in the decoder leads to improved performance. The external attention mechanism in the transformer effectively combines global features, leading to better results. Additionally, the Res-Channel attention blocks enhance the weight of effective feature maps, resulting in improved performance. The results demonstrate that both the external attention and Res-Channel attention contribute to the accuracy of the final grasp box predictions.

Table 3. The comparison results on Jacquard Dataset

	With external attention	With channel attention	Accuracy(%)
Cornell Dataset	√		97.2
		√	97.6
	√	√	**98.3**
Jacquard Dataset	√		94.2
		√	94.7
	√	√	**95.8**

4.4 Grasping in realistic scenarios

4.4.1. Experimental setup

In the grasping experiments, an Elite EC66 robot, a soft parallel gripper and an Orbbec Femto-W RGB-D camera are utilized as the experimental setup. As shown in Figure 7, the camera is positioned in a fixed location, and the image streams are captured by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

it. With the help of the Fin Ray effect inspired soft gripper, the robot can grasp different shapes of targets adaptively without force control. For robot motion planning and control, the open-source ROS is employed as the software platform. In order to verify the adaptability of the HTC-grasp algorithm to different shapes of grasping targets, more than 40 household objects are used in the experiment. As shown in Figure 8, these targets to be grasped such as rectangular, cylindrical, circular and other objects of various shapes and sizes.

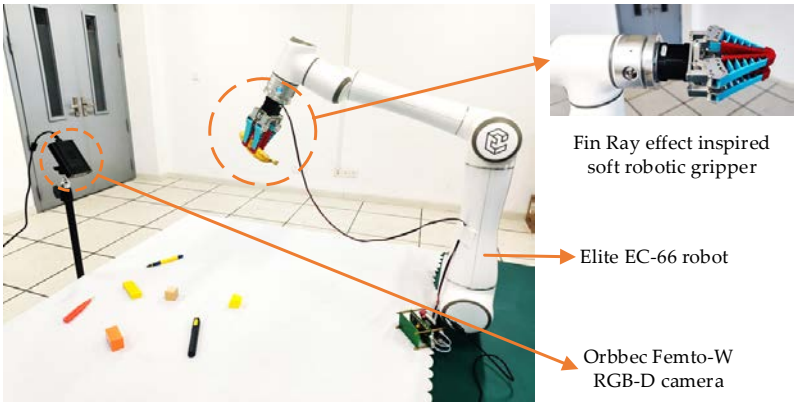


Figure 7. Realistic scenario experiment setup. The experimental equipment includes the Elite EC-66 robot, the Fin Ray effect inspired soft parallel gripper and the Orbbec Femto-W RGB-D camera.

Figure 8. Household objects used in the experiment.

4.4.1. Experiment results

The RGB-D camera captures video streams and fed them into the proposed model to obtain the best grasping pose. Subsequently, the robot's end actuator approaches the target according to the motion planning method, and the gripper is closed to grasp the target. The end actuator is then able to lift the object to another location.

Figure 9 illustrates the grasp process. A total of 180 grips were carried out to grasp household objects and the robot successfully grasped 168 times with an accuracy rate of 93.3%. The detailed experimental results are presented in Table 4 and demonstrate the effectiveness of HTC-Grasp method in real-world robot grasping tasks. Some successful and unsuccessful grasp examples are shown in appendix A and B.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2

According to the experiment results, the proposed method has excellent detection results in most cases. However, when faced with poor lighting conditions, transparency, reflections, etc., the accuracy of the proposed algorithm detection decreases.

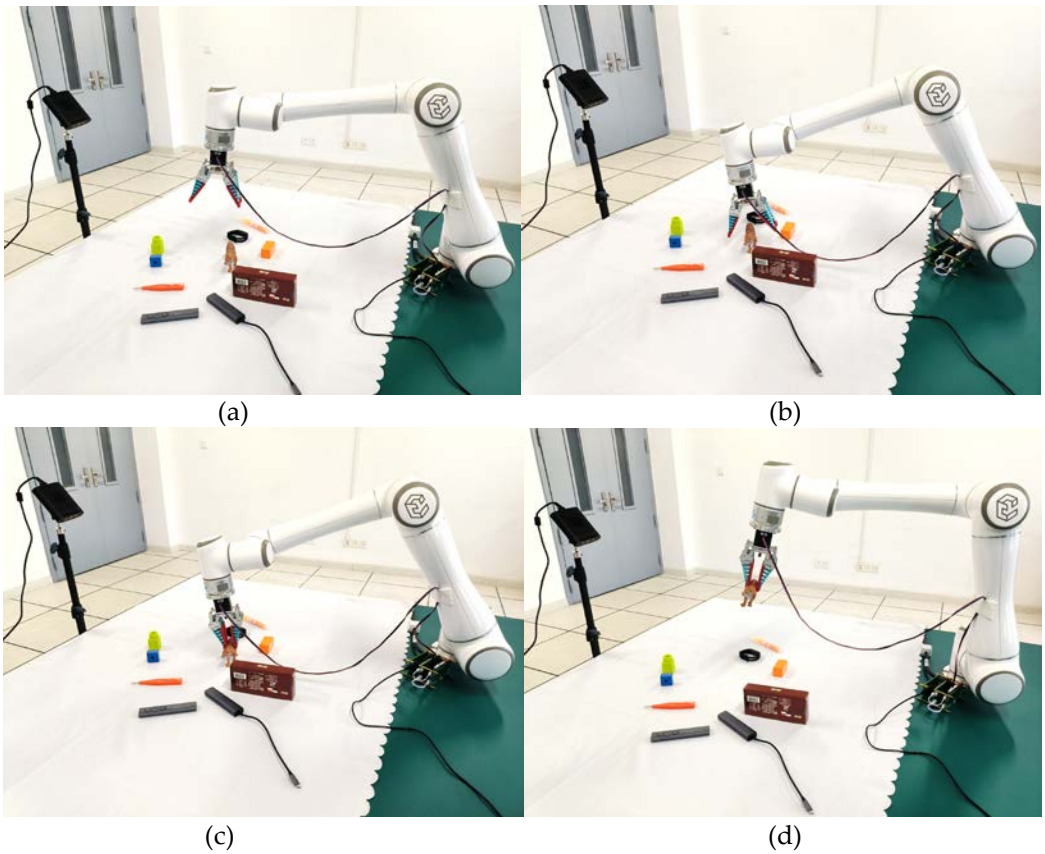


Figure 9. Example of the robotic grasp process. (a) shows the initial state of the robot. (b) illustrates the robot's gripper has moved to the target to be grasped. (c) shows the state of the object being grasped. (d) demonstrates the target being moved to another location

Table 4. Grasp success rates in robotic grasping experiments

Authors	Physical grasp	Success rate
Lenz [15]	89/100	89.0%
Morrison [14]	110/120	92.0%
Chu [29]	89/100	89.0%
Wang [27]	152/165	92.1%
Ours	168/180	**93.3%**

5. Conclusion

This article proposes a novel hierarchical hybrid architecture that combines Transformer and convolutional neural network (CNN) for visual grasping in robotics. Specifically, the proposed architecture enhances the conventional CNN by incorporating an external attention-based hierarchical Transformer as the encoder to capture global context and generate more informative feature representations. Additionally, a channel-wise attention mechanism is introduced to adaptively adjust channel weights for efficient feature aggregation. The proposed architecture, HTC-Grasp, is evaluated on two benchmark datasets, namely Cornell and Jacquard, and found that the proposed approach consistently outperformed existing state-of-the-art methods, leading to significant improvements in grasping accuracy.

 doi:10.20944/preprints202302.0382.v2

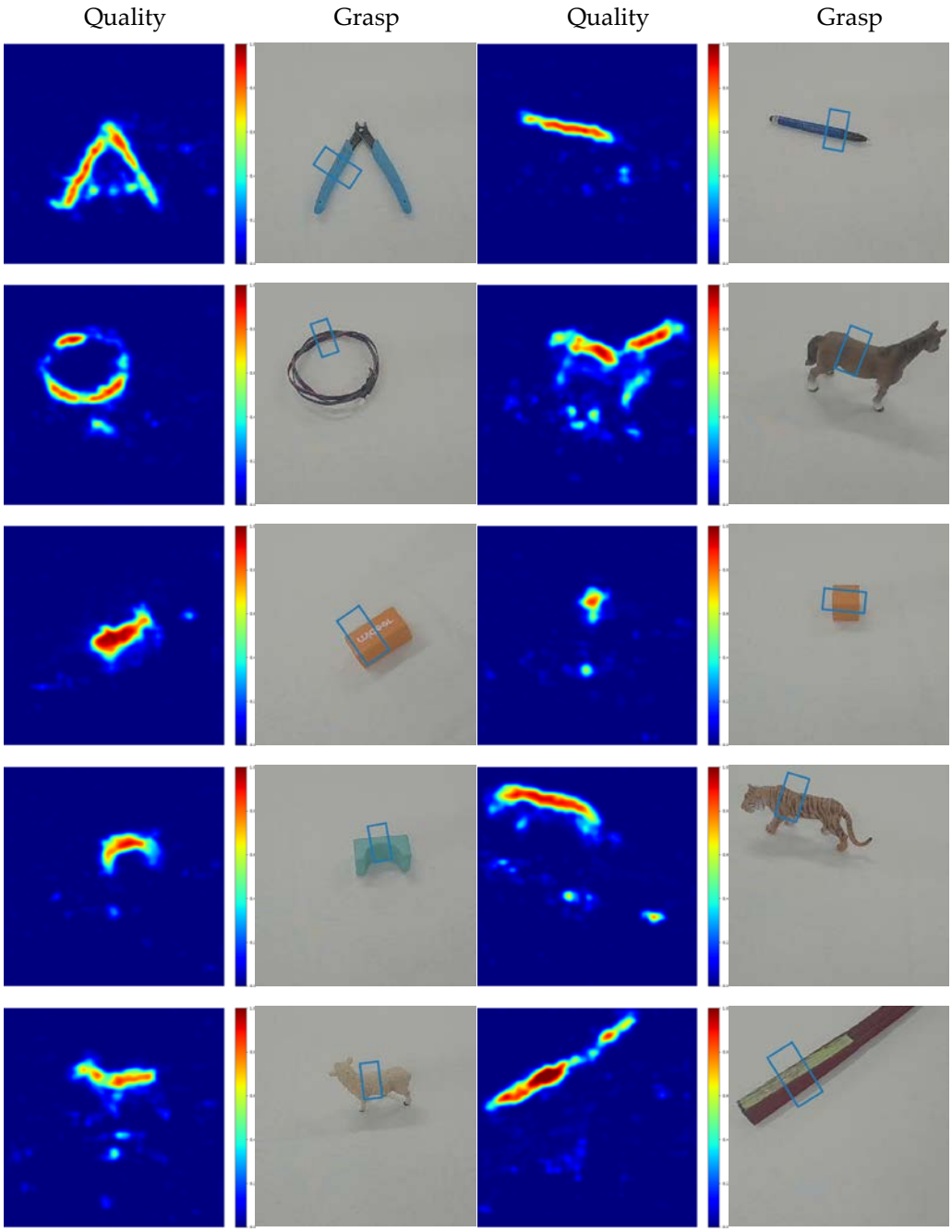
Author Contributions: Conceptualization, Q. Zhang; methodology, Q. Zhang and J. Zhu; software, Q. Zhang and J. Zhu; writing—original draft preparation, J. Zhu. and X. Sun.; writing—review and editing, Q. Zhang and M. Liu; visualization, X. Sun. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 61903162) and Jiangsu Province's "Double Innovation Plan": Research and development of flexible cooperative robot technology for intelligent manufacturing.

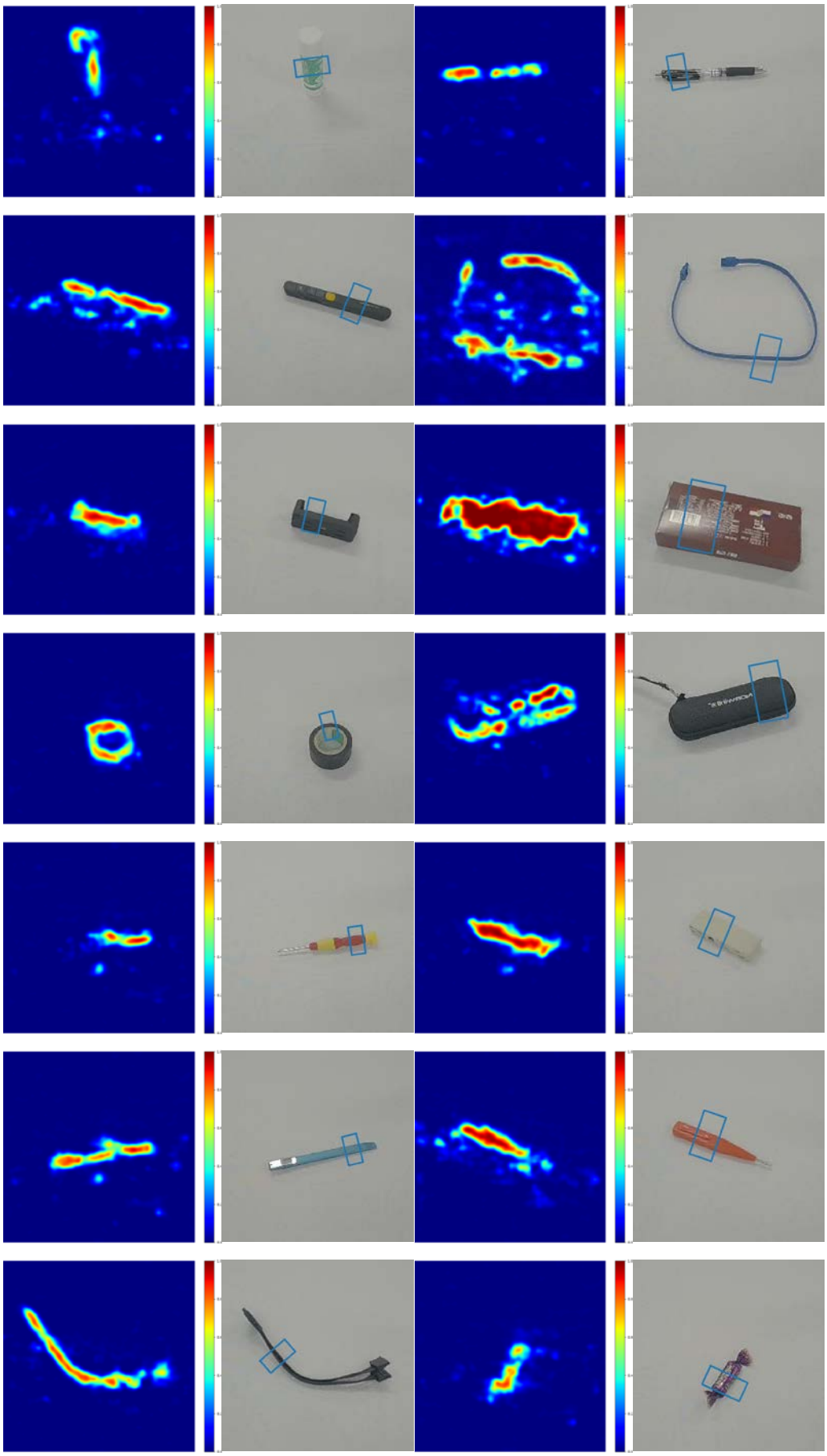
Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Appendix A: Sample of successful grasps

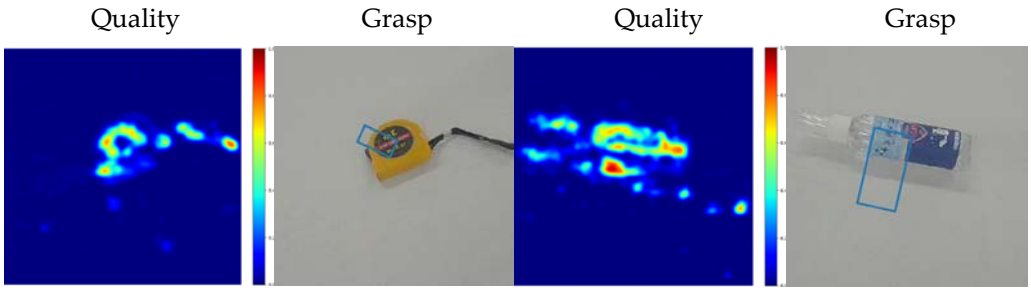


Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 March 2023 doi:10.20944/preprints202302.0382.v2



 doi:10.20944/preprints202302.0382.v2

Appendix B: Sample of unsuccessful grasps



References

1. Tian, Y.; Chen, C.; Sagoe-Crentsil, K.; Zhang, J.; Duan, W. Intelligent Robotic Systems for Structural Health Monitoring: Applications and Future Trends. *Autom. Constr.* **2022**, *139*, 104273, doi:https://doi.org/10.1016/j.autcon.2022.104273. [CrossRef]
2. Torres, R.; Ferreira, N. Robotic Manipulation in the Ceramic Industry. *Electronics* **2022**, *11*, doi:10.3390/electronics11244180. [CrossRef]
3. Zhang, H.; Lan, X.; Bai, S.; Zhou, X.; Tian, Z.; Zheng, N. Roi-Based Robotic Grasp Detection for Object Overlapping Scenes. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); IEEE, 2019; pp. 4768–4775. [CrossRef]
4. Du, G.; Wang, K.; Lian, S.; Zhao, K. Vision-Based Robotic Grasping from Object Localization, Object Pose Estimation to Grasp Estimation for Parallel Grippers: A Review. *Artif. Intell. Rev.* **2021**, *54*, 1677–1734. [CrossRef]
5. Sun, Y.; Falco, J.; Roa, M.A.; Calli, B. Research Challenges and Progress in Robotic Grasping and Manipulation Competitions. *IEEE Robot. Autom. Lett.* **2022**, *7*, 874–881, doi:10.1109/LRA.2021.3129134. [CrossRef]
6. Pinto, L.; Gupta, A. Supersizing Self-Supervision: Learning to Grasp from 50k Tries and 700 Robot Hours. In Proceedings of the 2016 IEEE international conference on robotics and automation (ICRA); IEEE, 2016; pp. 3406–3413. [CrossRef]
7. Wang, Z.; Li, Z.; Wang, B.; Liu, H. Robot Grasp Detection Using Multimodal Deep Convolutional Neural Networks. *Adv. Mech. Eng.* **2016**, *8*, 1687814016668077. [CrossRef]
8. Asif, U.; Tang, J.; Harrer, S. GraspNet: An Efficient Convolutional Neural Network for Real-Time Grasp Detection for Low-Powered Devices. In Proceedings of the IJCAI; 2018; Vol. 7, pp. 4875–4882. [CrossRef]
9. Karaoguz, H.; Jensfelt, P. Object Detection Approach for Robot Grasp Detection. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA); IEEE, 2019; pp. 4953–4959. [CrossRef]
10. Song, J.; Patel, M.; Ghaffari, M. Fusing Convolutional Neural Network and Geometric Constraint for Image-Based Indoor Localization. *IEEE Robot. Autom. Lett.* **2022**, *7*, 1674–1681. [CrossRef]
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. **2021**. [Arxiv]
12. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: Montreal, QC, Canada, October 2021; pp. 548–558. [CrossRef]
13. Jiang, Y.; Moseson, S.; Saxena, A. Efficient Grasping from RGBD Images: Learning Using a New Rectangle Representation. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation; 2011; pp. 3304–3311. [CrossRef]
14. Morrison, D.; Corke, P.; Leitner, J. Learning Robust, Real-Time, Reactive Robotic Grasping. *Int. J. Robot. Res.* **2020**, *39*, 183–201. [CrossRef]
15. Lenz, I.; Lee, H.; Saxena, A. Deep Learning for Detecting Robotic Grasps. *Int. J. Robot. Res.* **2015**, *34*, 705–724. [CrossRef]

16. Zhou, X.; Lan, X.; Zhang, H.; Tian, Z.; Zhang, Y.; Zheng, N. Fully Convolutional Grasp Detection Network with Oriented Anchor Box. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); IEEE, 2018; pp. 7223–7230. [CrossRef]
17. Zhang, H.; Lan, X.; Bai, S.; Zhou, X.; Tian, Z.; Zheng, N. ROI-Based Robotic Grasp Detection for Object Overlapping Scenes 2019. [CrossRef]
18. Laili, Y.; Chen, Z.; Ren, L.; Wang, X.; Deen, M.J. Custom Grasping: A Region-Based Robotic Grasping Detection Method in Industrial Cyber-Physical Systems. *IEEE Trans. Autom. Sci. Eng.* **2023**, *20*, 88–100, doi:10.1109/TASE.2021.3139610. [CrossRef]
19. Redmon, J.; Angelova, A. Real-Time Grasp Detection Using Convolutional Neural Networks. In Proceedings of the 2015 IEEE international conference on robotics and automation (ICRA); IEEE, 2015; pp. 1316–1322. [CrossRef]
20. 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); IEEE, 2017; pp. 769–776. [CrossRef]
21. Kumra, S.; Joshi, S.; Sahin, F. Antipodal Robotic Grasping Using Generative Residual Convolutional Neural Network. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); IEEE, 2020; pp. 9626–9633. [CrossRef]
22. Mahler, J.; Liang, J.; Niyaz, S.; Laskey, M.; Doan, R.; Liu, X.; Aparicio, J.; Goldberg, K. Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics. In Proceedings of the Robotics: Science and Systems XIII; Robotics: Science and Systems Foundation, July 12 2017. [CrossRef]
23. Yu, S.; Zhai, D.-H.; Xia, Y.; Wu, H.; Liao, J. SE-ResUNet: A Novel Robotic Grasp Detection Method. *IEEE Robot. Autom. Lett.* **2022**, *7*, 5238–5245. [CrossRef]
24. Wu, Y.; Zhang, F.; Fu, Y. Real-Time Robotic Multigrasp Detection Using Anchor-Free Fully Convolutional Grasp Detector. *IEEE Trans. Ind. Electron.* **2021**, *69*, 13171–13181. [CrossRef]
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2021; pp. 10012–10022. [CrossRef]
26. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090. [Arxiv]
27. Wang, S.; Zhou, Z.; Kan, Z. When Transformer Meets Robotic Grasping: Exploits Context for Efficient Grasp Detection. *IEEE Robot. Autom. Lett.* **2022**, *7*, 8170–8177. [CrossRef]
28. Guo, M.-H.; Liu, Z.-N.; Mu, T.-J.; Hu, S.-M. Beyond Self-Attention: External Attention Using Two Linear Layers for Visual Tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**. [CrossRef]
29. Chu, F.-J.; Xu, R.; Vela, P.A. Real-World Multiobject, Multigrasp Detection. *IEEE Robot. Autom. Lett.* **2018**, *3*, 3355–3362. [CrossRef]