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

Qiang Zhang , Jianwei Zhu , Xueying Sun * , Mingmin Liu

Posted Date: 22 February 2023

doi: 10.20944/preprints202302.0382.v1

Keywords: Robotic Grasp; Transformer; attentional mechanism

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, China;

[2] Central Research Institute, SIASUN Robot & Automation Co., LTD., NO.16 Jinhui Street, Hunnan District, Shenyang 110168, China;

* Correspondence: sunxueying@ just.edu.cn

Abstract: We introduce a novel hybrid Transformer-CNN architecture for robotic grasp detection, designed to enhance the accuracy of grasping unknown objects. Our proposed architecture has two key designs. Firstly, we develop a hierarchical transformer as the encoder, incorporating the external attention to effectively capture the correlation features across the data. Secondly, the decoder is constructed with cross-layer connections to efficiently fuse multi-scale features. Channel attention is introduced in the decoder to model the correlation between channels and to adaptively recalibrate the channel correlation feature response, thereby increasing the weight of the effective channels. Our method is evaluated on the Cornell and Jacquard public datasets, achieving an image-wise detection accuracy of 98.3% and 95.8% on each dataset, respectively. Additionally, we achieve object-wise detection accuracy of 96.9% and 92.4% on the same datasets. 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-world scenarios. The results of this study show that our proposed method outperforms other state-of-the-art methods.

Keywords: Robotic Grasp; Transformer; attentional mechanism

1. Introduction

In recent years, 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's ability to effectively capture global information through its self-attentive mechanism makes it a more representative model. In this article, we propose a novel robot grasp detection network that combines the Transformer and CNN architectures. The network features an encoder composed of Transformer layers, which provide multi-scale feature information, and a decoder that uses CNN with Res-channel attention blocks for feature aggregation to improve accuracy. The original contributions of this research are outlined below:

 doi:10.20944/preprints202302.0382.v1

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

2. The encoder's Transformer layer is enhanced with efficient external attention to better capture the relationships between different images. The decoder is designed with Res-channel attention blocks to more efficiently learn channel-wise features.

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

2. Related Works

The representation of object grasping is crucial for robot grasp detection. Jiang et al. [13] proposed a 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] proposed a grasp location description method based on a grasp map, 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 types: cascade methods and one-stage methods. Cascade methods 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, and the second stage 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] proposed a robotic grasp detection algorithm named ROI-GD, 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 neural network model to directly output the grasp position. Previous 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] proposed 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] proposed a U-Net based neural network with channel attention modules to better utilize features. Wu et al. [24] introduced an anchor-free grasp detector based on a fully convolutional network that formulates grasp detection as a closest horizontal or vertical rectangle regression task and a grasp angle classification task.

Recently, the transformer has gained traction in the field of computer vision due to its ability to model global information, overcoming the limitations of CNN models in using contextual information. The transformer has been successfully applied to vision tasks [25,26] through its self-attention mechanism and pyramid-like structure. In 2022, Wang et al. [27] used the SWIN Transformer as a backbone for feature extraction with impressive results.

Our proposed model, HTC-Grasp, differs from these efforts in two key ways. Firstly, it employs external attention for the transformer block to enhance the representation of the correlation between different images. Secondly, it uses a Residual connection-based channel attention block for the decoder to efficiently learn discriminative channel-wise features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0382.v1

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 center coordinates, (w,h) represents the width and height of the grasp box, and θ is the angle between the horizontal axis of the grasp box and the horizontal axis of the image.

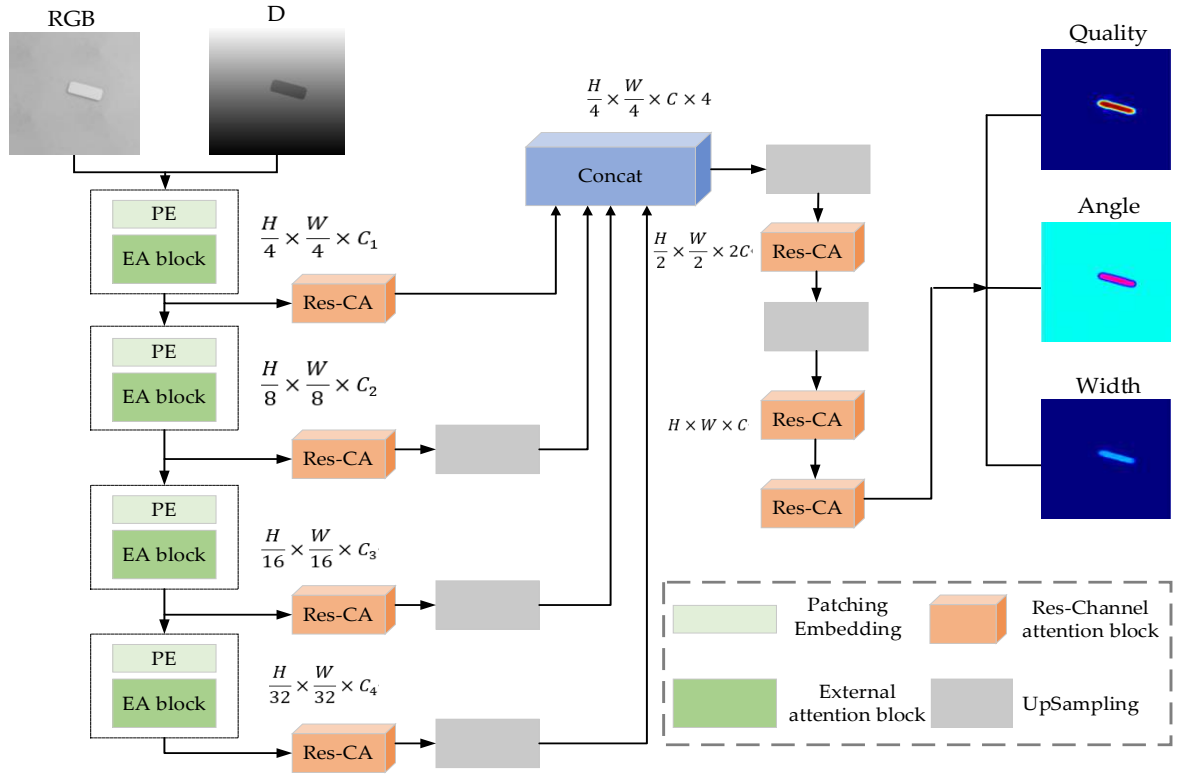
An alternative representation for high-precision, real-time robot grasping was introduced in [14]. In this representation, the grasp is redefined for 2DoF robotic grasping tasks as follows:

$$P = \{\mathrm{Q}, \Theta, \mathrm{W}\} \in \mathbb{R}^{3\times w\times h} \tag{2}$$

where P is a 3-dimensional tensor. The first dimension, Q, represents the grasping quality of each pixel in the image; the second dimension, Θ, denotes the orientation angle of the gripper; and the third dimension, W, represents the width of the gripper. Each pixel, with a specific width $\mathrm{W}_{i,j}$ and angle $\Theta_{i,j}$, corresponds to the width and orientation angle of the gripper at that particular position.

3.2. Grasp Overview

In this section, we present the proposed neural network architecture for grasp detection, which is illustrated in Figure 1. The architecture of the HTC-Grasp network consists of three parts: the encoder, the decoder, and the prediction head. The encoder is built using hierarchical transformers with a pyramid-like structure to extract both high-resolution coarse features and low-resolution fine features. The decoder, made up of 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 quality score map, the angle map in the form of $\sin(2\theta)$ and $\cos(2\theta)$, and the gripper width.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0382.v1

Figure 1. Overall network architecture of HTC-Grasp.

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 passed to the decoder to predict the grasp heatmaps. In the following sections, we will delve into the details of the proposed encoder-decoder design.

3.3. Neural Network Architecture

3.3.1. Hierarchical Transformer Encoder

We introduced a pyramid structure into the Transformer architecture to facilitate the generation of multi-scale feature maps. High-resolution coarse features and low-resolution fine 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 architecture of each stage is similar and consists of a Patch Embedding Layer followed by a Transformer block.

To be more specific, we take an input image with a resolution of $H \times W \times 4$ and feed it into the 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, we intentionally have overlapping parts between each patch in the partitioning. Then the images patches are fed into the encoder to obtain multi-scale features.

The Transformer blocks are used to extract features. Each Transformer block consists of self-attention and feed-forward layer. 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 self-attention is then calculated as follows:

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

The computational complexity of self-attention is $O(N^2)$, which 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, we introduce the multi-head external attention (MEA) [28] mechanism as a replacement for the standard multi-head self-attention (MSA) module in our Transformer blocks.

To improve the efficiency of the transformer layer, we incorporate the use of MEA mechanism. 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, H 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: 22 February 2023 doi:10.20944/preprints202302.0382.v1

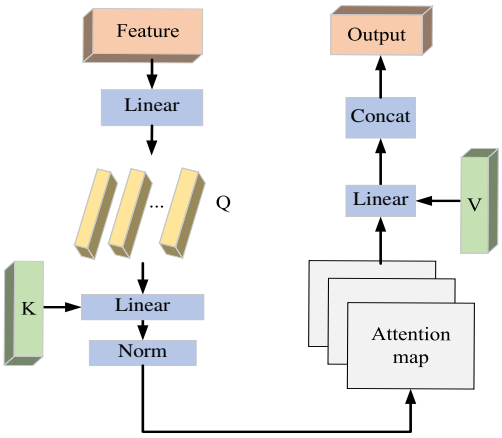


Figure 2. The architecture of external attention block.

3.3.2. Grasp Decoder

Our decoder is designed with a combination of convolutional layers and Res-channel attention blocks. As illustrated 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 three convolutional layers with kernel sizes of 1×1, 3×3 and 1×1. The channel attention block, on the other hand, utilizes global average pooling to reduce the number of parameters contained in the features. This block is then composed of two fully connected layers and one ReLU layer, which utilizes global information to selectively emphasize important features and reduce the emphasis on less relevant features.

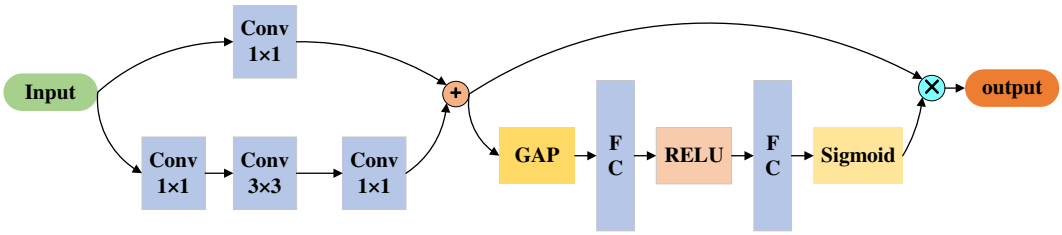


Figure 3. Res-Channel attention block.

Specifically, the decoder we propose involves three key steps. To begin with, the multilevel features F_i from the encoder are fed through the upsample 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 our study, we define the task of robot grasp detection as a regression problem and adopt the smooth $L1$ loss function as our optimization objective. This loss function has the advantage of being robust to outliers and provides 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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0382.v1

In our 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 grasping angle, and w represents the width of the gripper.

4. Experiments and Results

4.1. Dataset

In this chapter, experiments are conducted on two popular datasets, the Cornell and the Jacquard datasets, to fully validate the performance of the proposed method.

(a) Datasets

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

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 our work.

(b) Implementation details

In this ariticle, the model was constructed using the Pytorch framework on the Ub-untu 20.04 platform. For training, we utilized an NVIDIA RTX 3090Ti GPU and an Intel Core i9-12900K CPU. 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, a batch of samples was randomly selected from the training set, with 200 batches of size 32 in each epoch, and 100 epochs are trained in total. AdamW is employed as the optimizer, with an initial learning rate of 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. The number of heads 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 our work, each dataset is divided 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 grasp 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 appear in the training set, 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 discrepancy between the predicted grasping angle and the ground truth must be within 30°

(2) The IOU index, 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 our method against other grasp detection methods, we use the same evaluation metric to compare their results on both the Cornell and Jacquard datasets.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0382.v1

to the best grasp position. Figure 4 presents the grasp detection results of GR-CNN , TF-Grasp [27] and the proposed HTC-Grasp for unseen objects on the Cornell Dataset. The results indicate that HTC-Grasp has a higher grasp quality as compared to the GR-CNN and TF-Grasp methods.

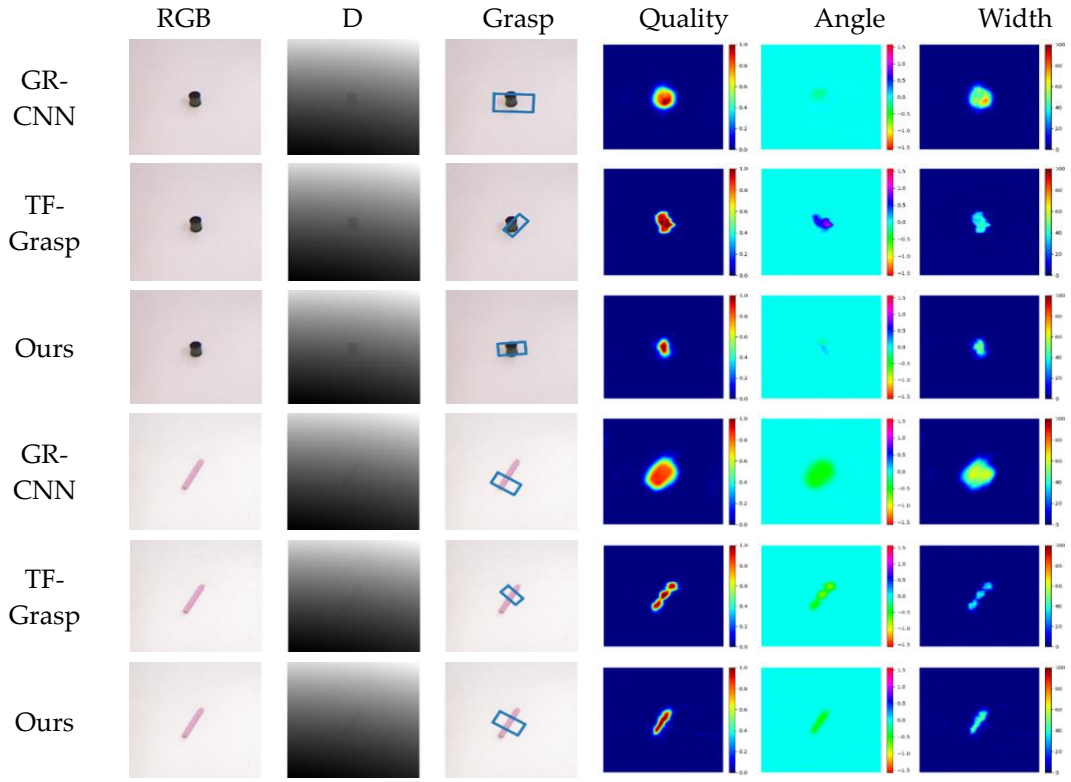


Figure 4. Comparison of predicted heatmaps on Cornell Dataset.

For the classical method experimental results presented in Table 1, we have selected the data reported in their original paper. 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, our model, utilizing the NVIDIA RTX 3090Ti GPU, processes a single frame in approximately 5.4 ms, fulfilling the requirement for real-time processing.

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**

We conducted comparative experiments using the Jacquard dataset. 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 grasping quality compared to GR-CNN and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0382.v1

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% on the Jacquard dataset.

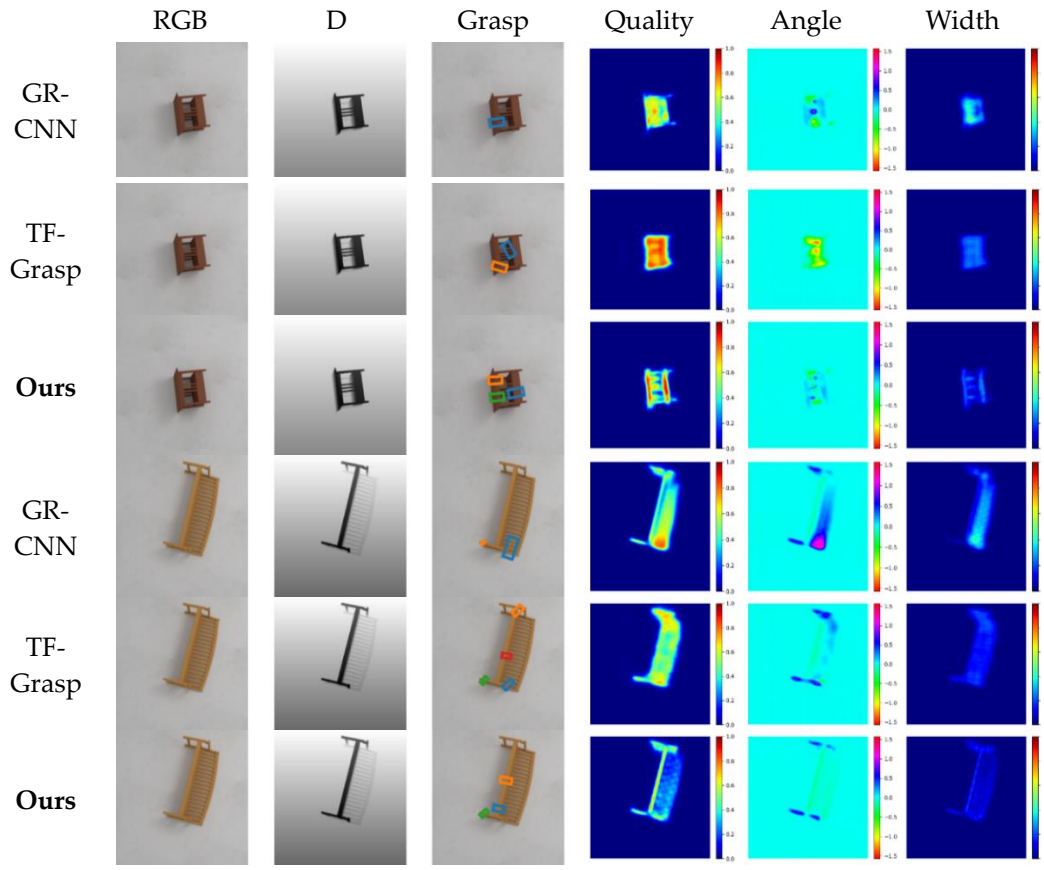


Figure 5. Comparison of predicted heatmaps on Jacquard Dataset.

Table 2. The comparison results on Jacquard Dataset.

Authors	Method	Input	Accuracy (%)
Morrision [14]	GG-CNN2	D	84
Kumra [21]	GR-CNN	RGB-D	94.6
Wang [27]	TF-Grasp	RGB-D	94.6
Ours	HTC-Grasp	RGB-D	**95.8**

We present qualitative comparison results for the Cornell and Jacquard datasets, as 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, our proposed method, 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, we introduced a Residual Channel attention module in 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0382.v1

Experimental results demonstrate that the proposed framework 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 effectively captures both global information and fine-grained features such as the exact lo-cation and shape of the object.

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

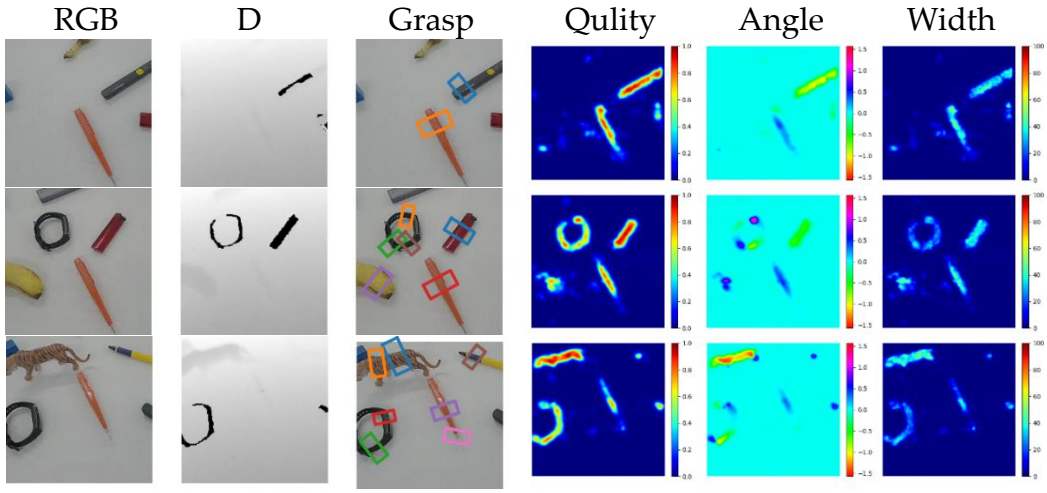


Figure 6. Test result of our method in the real-world multiple objects environment.

4.3. Ablation Studies

To validate the impact of external attention and channel attention on the proposed grasp detection model, we conduct experiments on the Cornell and Jacquard datasets. Our 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

In our grasping experiments, we utilize an Elite EC66 robot and an Orbbec Femto-W RGB-D camera as the experimental setup. The camera is positioned in a fixed location, and the image streams

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0382.v1

are captured by it. The RGB-D images are then fed into our 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 7 illustrates the grasp process. Our method is tested on 180 household objects, and the robot success-fully grasped 168 objects with a 93.3% accuracy rate. The detailed experimental results are presented in Table 4 and demonstrate the effectiveness of our method in real-world robot grasping tasks.

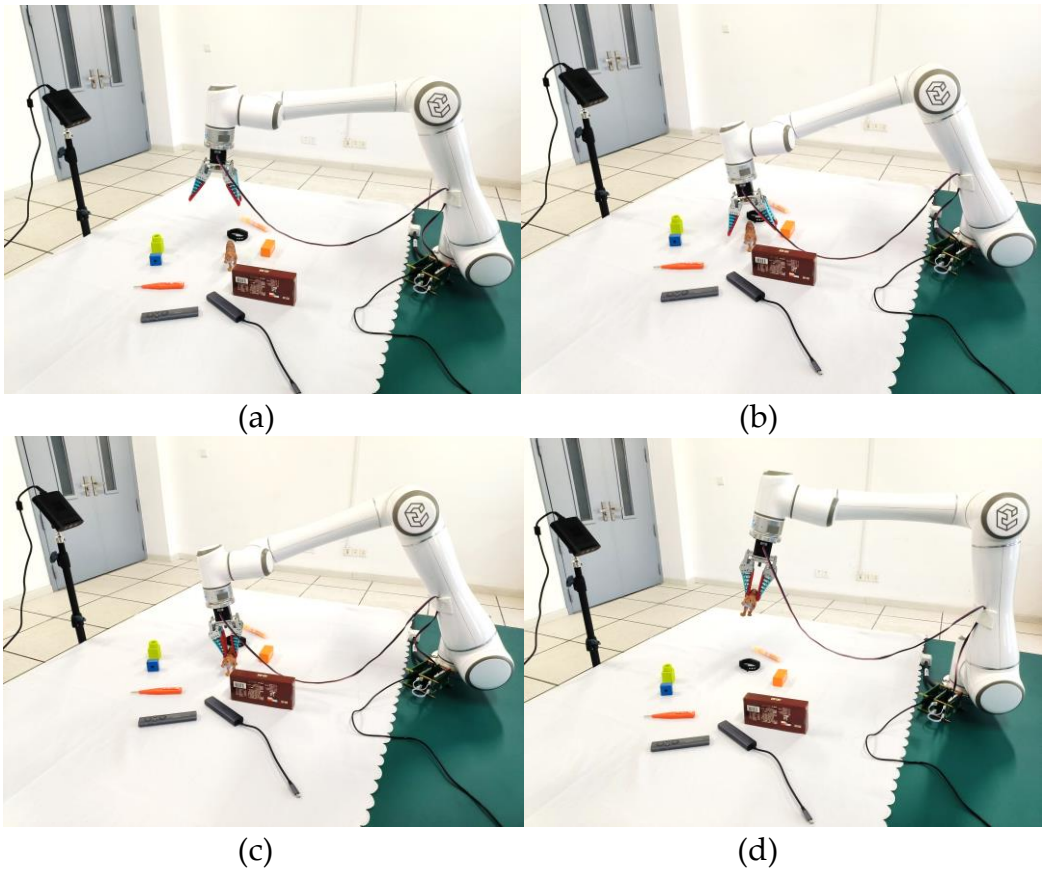


Figure 7. Example of the robotic grasp process. (a) shows the initial state of the robot. (b) illustrates the robot's gripper has moved to the target to be grasped. (c) shows the state of the object being grasped. (d) demonstrates the target being moved to another location.

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

In this article, we present a novel hierarchical hybrid transformer CNN architecture for robotic visual grasping, named HTC-Grasp. The proposed architecture uses hierarchical transformer blocks with external attention as encoders to enhance the ability to capture long-range spatial correlations at multiple scales, and the Res-Channel attention block in the decoder module adaptively recalibrates the channel-wise feature response to achieve precise positioning. We evaluated the performance of

HTC-Grasp on the Cornell and Jacquard datasets and found that it outperformed existing methods, achieving favor-able grasping results.

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.

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.
2. Torres, R.; Ferreira, N. Robotic Manipulation in the Ceramic Industry. *Electronics* **2022**, *11*, doi:10.3390/electronics11244180.
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.
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.
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.
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.
7. Wang, Z.; Li, Z.; Wang, B.; Liu, H. Robot Grasp Detection Using Multimodal Deep Convolutional Neural Networks. *Adv. Mech. Eng.* **2016**, *8*, 1687814016668077.
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.
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.
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.
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**.
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.
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.
14. Morrison, D.; Corke, P.; Leitner, J. Learning Robust, Real-Time, Reactive Robotic Grasping. *Int. J. Robot. Res.* **2020**, *39*, 183–201.
15. Lenz, I.; Lee, H.; Saxena, A. Deep Learning for Detecting Robotic Grasps. *Int. J. Robot. Res.* **2015**, *34*, 705–724.
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.
17. 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. 2019.
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.

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.
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.
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.
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.
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.
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.
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.
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.
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.
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**.
29. Chu, F.-J.; Xu, R.; Vela, P.A. Real-World Multiobject, Multigrasp Detection. *IEEE Robot. Autom. Lett.* **2018**, *3*, 3355–3362.