Article

A Smart Grasp Deep Network Based on One-Way Fusion Strategy

Yuequan Yang *, Wei Li, Xin Cang, Zhiqiang Cao, Jiatong Bao *

Posted Date: 2 October 2024

doi: 10.20944/preprints202410.0076.v1

Keywords: robot grasp; lightweight structure; deep network; fusion strategy; transformer

 doi:10.20944/preprints202410.0076.v1

Article

A Smart Grasp Deep Network Based on One-Way Fusion Strategy

Yuequan Yang [1,*], Wei Li [1], Xin Cang [1], Zhiqiang Cao [2,3], Jiatong Bao [4,*]

[1] College of Information Engineering (College of Artificial Intelligence), Yangzhou University, Yangzhou 225127, China
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China
[4] College of Electrical, Energy and Power Engineering, Yangzhou University, Yangzhou 225127, China
* Correspondence: yang@yzu.edu.cn (Y.Y.); jtbao@yzu.edu.cn (J.B.)

Abstract: Robot grasp modeling and implementation is one of essential abilities for a robot with embodied artificial intelligence. However, most existing deep learning-based grasp methods suffer a large number of parameters and heavy computation overheads. To address this issue, by fully exploiting the complementary capabilities of both CNNs and Transformers, we propose a smart grasp deep network with one-way fusion strategy via context path and spatial path (SGNet), which enjoys a lightweight structure, fast inference speed and easy deployment on devices with limited computation resources. Specifically, the context path employs lightweight depthwise separable convolution to achieve fast down-sampling while a novel DSFormer module mainly by integrating Transformer is to extract global and context-rich features. The spatial path efficiently fuses feature information from the context path in one-way manner and generate high-resolution feature maps via point-by-point convolution operations. Experimental results show the proposed model with only 1M parameters has a significantly overall performance, achieving 99.4% accuracy on Cornell dataset and 93.4% accuracy on Jacquard dataset, as well as within 12.5ms inference time.

Keywords: robot grasp, lightweight structure, deep network, fusion strategy, transformer

1. Introduction

With the rapid evolution of modern technology and industry, robotics has been occupying an important position in a wide range of applications, from manufacturing to healthcare, from rescue missions in hazardous environments to agricultural fields, and so on. Among widely applied fields in robotics, grasp detection tasks certainly play a crucial role. Yet, the complexity and unpredictability of the real-world scenarios lead to weak robustness and slow response by using traditional grasp methods. Thus, there are many limitations existed in practical applications.

Nowadays, deep learning technologies have achieved great successes in such fields as object detection, natural language processing, automated driving, lesions detection just to name a few. Deep learning methods based on convolution neural networks (CNNs)[1–3] are able to identify and localize target region more accurately as well as to learn the geometric features directly from image data. CNNs provide the ability to extract and learn features directly from raw data by sharing convolution kernels. By stacking multiple convolution and pooling layers, they can capture different levels of features. With their superior local feature extraction capability, robot grasping detection can be effectively realized. Lenz et al.[1] adopted five-dimensional grasping configurations for the representation of robot grasping, which was realized by regressing the rectangular boxes. Zhang et al.[3] proposed a multi-task deep network to fulfill autonomous grasping of robots in complex scenes, which could help robots to find targets and make target grasping planning. Although CNNs have obtained good results in the works above, they have limited receptive field which only captures local feature information, and hard to catch global context information. For vision tasks such as object detection, instance segmentation, it is critical to model global and long-range correlations.

 doi:10.20944/preprints202410.0076.v1

Recently, following the success of Transformer[4] in the field of natural language processing, there has been a boom in the application of Transformer in the computer vision field. EAPT[5] utilizes deformable attention to provide learnable offsets for each position in each patch and employs the En-DeC module for global communication. Transformers[6,7] captured global context information via the self-attention mechanism, allowing the models to assign different weights based on information from different locations of the input, creating remote dependencies on the target and extracting more effective. Whereas Transformer is gradually applied to various computer vision tasks, it is still essential to train network weights using large-scale datasets. And small-scale robot grasp image datasets make it difficult to meet the training requirements, constraining the model efficiency. Accordingly, while Transformer can reflect the complex spatial transformations and long-distance elemental dependencies of global representations, Transformer is inferior to CNNs in local feature extraction, and suffers from small amount of training datasets[8].

In view of the above observations, and considering the deployment of robotic grasping tasks on resource-limited mobile terminals, we propose a smart grasp hybrid network by combining the advantages of CNNs and Transformer. Besides, a spatial path and a context path are constructed, respectively, which are employed to compensate for the loss of spatial information and the receptive field problem to further improve the model performance. Our main contributions are summarized as follows:

(1) We present a smart robot grasp deep hybrid network with lightweight structure and fast inference speed, which, with only 1M parameters by efficient fusion in one-way fusion manner between the spatial path and the context path, can greatly exploit complementary advantages between CNNs and Transformer to improve robot grasp performance.

(2) A universal attention module called DSFormer is established, which can efficiently model local and global feature maps with fewer parameters. Specifically, a lightweight depthwise separable convolution block and style-based recalibration block are integrated to complement modeling capability through spatial dimension and channel dimension, while the Transformer block is used to extract global and context-rich features.

(3) We validate the proposed deep hybrid network model on Cornell grasp dataset and Jacquard dataset, which exhibit excellent accuracy performance with 99.4% and 93.4% , respectively, as well as both within 12.5ms inference time.

2. Related work

In this section, we briefly review the related works, including robotic grasping, lightweight neural network and CNN-Transformer hybrid network.

2.1. Robotic Grasping

In the field of robot, the research on robotic grasping work has always attracted much attention, because grasping is the most fundamental and critical ability of robots and has the potential to have a significant positive impact on society. The advantages of robot grasping are not only shown in industrial production recently but also in the steady entry of robots into people's lives, especially with the creation of intelligent service household robots that can assist people with their daily grasping jobs. Although to grasp an object appears simple in real life, it may become highly sophisticated in the digital world, which maybe involve in many complex computations of its form, surface properties, grasping angle, grasping force, etc. In order to obtain stable grasping, early researchers[9] carried out grasping analysis and established models with grasping mechanics and grasper-object contact interaction as the core issues, and also researchers[10] grasped objects by manually designing grasping features. There are also some earlier efforts[11] that used 3D simulation to identify solid grasps. All of these techniques rely on the object's complete 3D model or other pertinent object information to determine appropriate grasping locations. Yet, actual grasping frequently occurs with unknown objects, making it challenging to achieve successful grasping with these techniques. Since 2012, deep learning techniques have made significant advancements in the domain of image detection, which has drawn the interest of specialists in robotic grasping. Lenz et

 doi:10.20944/preprints202410.0076.v1

al.[1] revealed that the five-dimensional grasping representation of 2D images can be projected into 3D space, and proposed an accurate real-time grasp detection method based on convolutional neural networks, which was the first application of deep learning in robotic grasping. Since then, a growing number of deep neural network frameworks have been brilliant in the field of robotic grasping.

2.2. Lightweight Neural Network

Deep neural networks are widely used in computer vision tasks such as image classification and object detection and have achieved great success. Typically, the network is designed as a basic stack of convolutional layers, which introduces a lot of redundant parameters into the network and degrades its real-time performance. For instance, the network parameters of the AlexNet network[12], which only included 5 convolutional layers and 3 FC layers, reached 60 million. In addition, the models are both computationally and memory-intensive due to the abundance of duplicated parameters, making it difficult to implement them in hardware with limited memory or in applications with strict latency requirements. Accordingly, under the premise of not significantly reducing the model performance, model acceleration and compression to create lightweight networks have become a research hotspot in recent years[13]. To reduce the model size, Song Han et al.[14] concentrated on filter pruning and created filter pruning as an optimization issue. By using the knowledge distillation method, a transfer learning technique[15] were able to learn the superior performance of complicated models through basic models and ultimately achieve model compression. Microfactored convolution was proposed by Yunsheng Li et al.[16], which involved breaking the convolution matrix into low-rank matrices, incorporating sparse connectivity into the convolution, and building the MicroNet network, which reduced the computational cost by learning the sparse matrix structure.

2.3. CNN-Transformer Hybrid Network

In recent years, the model of fusing CNNs and Transformer has become a hot topic in the field of deep learning. CNNs excel at processing image data and capturing localized features, while Transformers are good at processing sequential data and capturing long-range dependencies. Combining these two architectures makes it possible to exploit their advantages in order to improve the performance of a variety of tasks, especially in computer vision. CaiT[17] combined the structure of CNN and Transformer, and made a useful exploration of Transformer. It introduced LayerScale and Class-Attention, which significantly improves the accuracy of the deep Transformer. MobileViT[18] can efficiently encode local and global information by unfolding and collapsing feature maps along with using the self-attention mechanism of Transformer.

Also, it enabled efficient image processing and computer vision tasks on mobile while keeping the model complexity and computational resource requirements low. TranxNet[19] proposed a lightweight Dual Dynamic Token Mixer (D-Mixer) that aggregated global information and local details in an input-dependent way and introduced MS-FFN for multi-scale Token aggregation in feed-forward networks.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2024 doi:10.20944/preprints202410.0076.v1

3. Methods

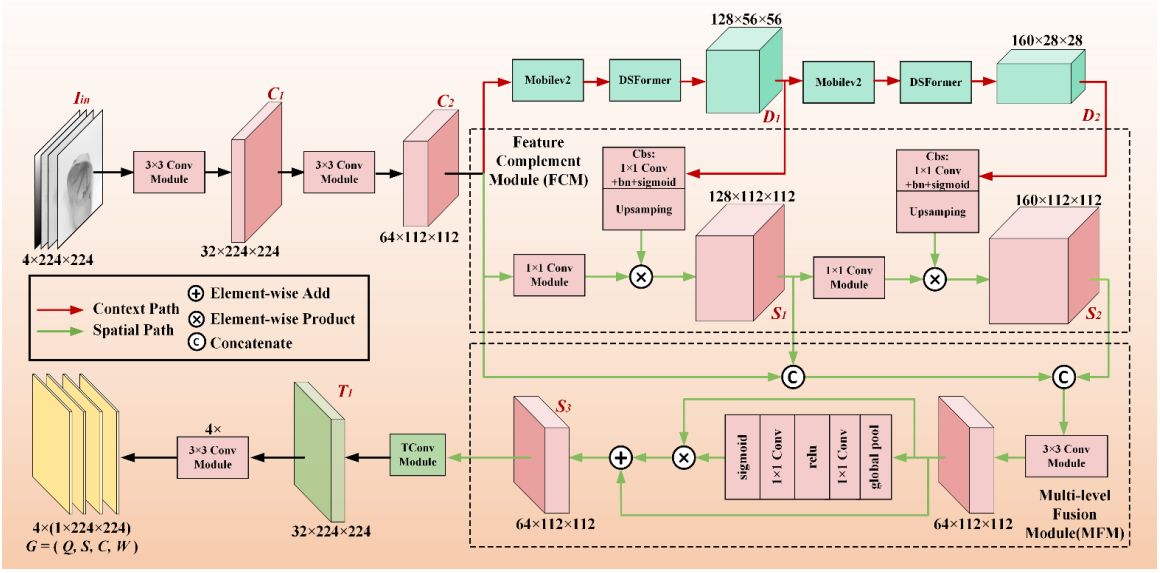

Figure 1. The overall architecture of the proposed grasp detection network (SGNet).

3.1. Network Structure

To realize the robot grasp detection task, we propose a smart grasp deep hybrid network with one-way fusion strategy via context path and spatial path, named SGNet, which is a generative detection network.

As shown in Figure 1, the network is separated into two paths, and the context path uses a lightweight framework integrating CNNs and Transformer to achieve fast down-sampling, obtain a large receptive field, and improve the local and global representation ability. The spatial path maintains half the spatial size of the original input image size and adopts convolution to encode rich spatial information and retain more detailed information.

Given an input RGB-D image I_{in} with the size of 224×224 different size feature maps C_1 and C_2 are obtained by two Conv Modules. Then C_2 is fed into the context path and fast downsampling is realized by Mobilev2 and DSFormer module, to obtain different size feature maps D_1 and D_2 with rich semantic information. For the two paths, the features have different level expressions, low-level via spatial path and high-level via the context path. Thus, the Feature Complement Module (FCM) is added to the spatial path, enabling low-level spatial features to fuse high-level semantic information, which can be represented as follows:

$$S_1 = Cov_{1\times1}(C_2) \otimes Cbs(Up(D_1)) \tag{1}$$

$$S_2 = Cov_{1\times1}(S_1) \otimes Cbs(Up(D_2)) \tag{2}$$

where $Cov_{1\times1}(\cdot)$ denotes a 1×1 convolutional layer, $Cbs(\cdot)$ contains a 1×1 convolution operation, batch normalization, and sigmoid activation function. $Up(\cdot)$ represents an upsampling operation and denotes element-wise multiplication. In the spatial path, S_1 and S_2 are generated via the FCM which complements the features with rich semantic information and the low-level spatial map.

Then the Multi-level Fusion Module (MFM) is used, which recomposes the information of various levels of spatial features. As illustrated in Figure 1, the features are reweighted employing methods like SENet[20]. After C_2, S_1 and S_2 are fused into the MFM module, the feature map S_3 is refined. Followed by a transposed convolution operation, S_3 is restored to T_1 with the same size as I_{in}. The process is represented as follows:

$$S_3 = SE(Cov_{3\times3}(Cat(C_2, S_1, S_2))) \tag{3}$$

 doi:10.20944/preprints202410.0076.v1

where S_1 and S_2 are the outputs of the FCM, $Cat(\cdot)$ indicates a concatenation operation, $Cov_{3\times3}(\cdot)$ is a 3×3 convolutional layer, and $SE(\cdot)$ denotes SE-ResNet Module.

Finally, the output feature of the TConv Module is processed by four parallel Conv Modules to yield the grasp prediction $G = (Q, S, C, W)$, where Q, S, C, and W correspond to feature maps of grasp quality, sine and cosine related to the grasp angle, and grasp width, respectively.

3.2. DSFormer

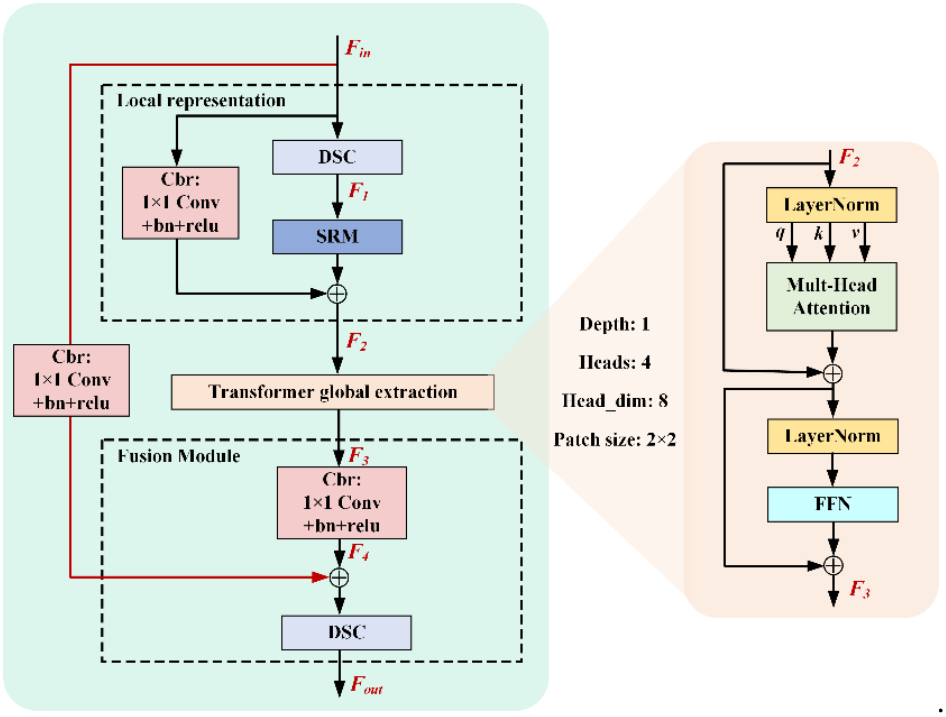

Figure 2. The pipeline of a DSFormer block.

Inspired by depthwise separable convolution (DSC)[21] and style-based recalibration module (SRM)[22], we elaborate a lightweight common attention module, termed DSFormer, to model local and global information via fewer parameters. As shown in Figure 2, unlike traditional vision transformer, DSFormer assembles DSC, SRM and point-by-point convolution in the local representation session to enhance the local feature information with less computation cost. The sizes of F_{in}, F_1, F_2, F_3, F_4 and F_{out} in the first DSFormer of the SGNet are 128×56×56, 160×56×56, 160×56×56, 160×56×56, 160×56×56, and 128×56×56, respectively, while the corresponding ones in the second are 160×56×56, 190×56×56, 190×56×56, 190×56×56, 160×56×56, and 160×56×56, respectively. The local representation session can be represented as follows:

$$F_2 = DSC(SRM(F_{in})) \oplus Cbr(F_{in}) \tag{4}$$

where F_{in} represents the input of the DSFormer, $Cbr(\cdot)$ contains a 1×1 convolution operation, batch normalization, and relu activation function, $\oplus$ represents the element-wise addition. Besides, $DSC(\cdot)$ and $SRM(\cdot)$ is DSC as well as SRM operation mentioned above. It's worth noting that DSC is a lightweight convolutional operation that reduces the parameters while maintaining the performance of the model, helping to improve computational efficiency in resource-limited environments, but this also leads to a relative isolation between spatial and channel information, degrading informational communication across different spatial locations. Thus, SRM is introduced to compensate for missed channel correlation caused by DSC. By extracting the style information from individual channels within the feature map and subsequently estimating recalibration weights for each channel by channel-independent style integration, it effectively enhances the network's representation.

 doi:10.20944/preprints202410.0076.v1

Then, the features flow through the Transformer block for global information modeling as follows:

$$F_3 = FFN(L(M(L(F_2)) \oplus F_{in})) \oplus (M(L(F_2)) \oplus F_{in}) \tag{5}$$

where F_2 is the output of the local representation, $L(\cdot)$, $M(\cdot)$, and $FFN(\cdot)$ represents LayerNorm, Multi-Hhead Attention, and feed-forward neural network, respectively.

After a Transformer block, we add a fusion unit to further merge local and global features. The process is illustrated below:

$$F_{out} = DSC(Cbr(F_3)) \oplus Cbr(F_{in}) \tag{6}$$

where F_{out} represents the output of the DSFormer, and F_3 is the output of the Transformer block.

The pipeline of DSFormer is as follows. The input feature F_{in} is firstly preprocessed by the Local representation to generate the enhanced feature F_2. It is pointed out that we upgrade the dimension of the feature by 1×1 convolution including point-by-point convolution in DSC such as F_1, and the number of channels is raised from 128 to 160, which provided more local features for the subsequent transformer global extraction. Then, the feature flow through the transformer global extraction with depth of 1, heads of 4, head_dim of 8, and patch_size of 2×2 to yield the global feature F_3, whose feature scale is the same. In the last session Fusion, we use 1×1 convolution to reduce the feature dimension into the feature F_4 of the same size as the input F_1, and further fuse the local and global information through the residual structure and DSC operation.

3.3. The Context and Spatial paths

The context path is designed to capture context-rich information and to provide the network with sufficient region of receptive field. For a generative grasp detection network, the receptive field plays an essential role in model representation. Also, considering the receptive field scale and efficient computation, we utilize the lightweight MobilNetV2 block[23] for down-sampling and expanding the channel dimensions, such operations ensure that the model has expressive capability on basis of decreased calculations. Further, DSFormer is inset after each down-sampling to enhance the global presentation.

Shallow convolution layer features have abundant spatial detail information, which is gradually lost as the scale declines during constant down-sampling operations. Hence, we introduce the spatial path to get high resolution information with rich spatial details, which is fused with context path information through FCM to offset the missing detail information. Then, the features of different layers are fused by MFM, and the final output feature map is half of the original image, which encodes plenty of spatial information.

The context path extracts high-level semantic information, but details are lost. The spatial path encodes detail information, but is merely low-level and inexpressive. The complementary strengths of the two paths collectively constitute advantages of the smart network.

3.4.Loss Function

We consider the defined grasping as a regression problem, so the proposed network is trained by Smooth L1 loss function[24] under the supervision of ground truth, which is described as follows:

$$L\left(G^p, G^L\right) = \frac{1}{n}\sum_m S\left(G^{Pm} - G^{Lm}\right), m \in \{Q, A, W\} \tag{7}$$

$$S(x) = \begin{cases} 0.5x^2 & if\ |x| \le 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{8}$$

 doi:10.20944/preprints202410.0076.v1

where n denotes the total number of grasp positions, each prediction consists of grasp quality Q, angle A, and width W. G^P and G^L refer to the prediction result of SGNet and the corresponding ground truth, respectively.

4. Experiments

The proposed method is validated on the Cornell grasp dataset[1] and Jacquard dataset[25]. The Cornell grasp dataset is created by a real-world RGB-D camera and specifically designed for an angle-symmetric robotic grasper, which contains 885 RGB images of 240 different objects. Simultaneously, due to the small size of the grasp dataset, image enhancement operations such as random cropping, scaling, and rotation is carried out to extend the dataset so as to prevent the training process from overfitting. The Jacquard dataset is constructed in the simulated environment which contains 54,000 RGB-D images with 1.1 million grasping instances. We divide the dataset into train and test sets in the ratio of 4:1.

4.1. Grasp Detection Metric

For a grasp prediction result, it is regarded as valid if the following conditions are satisfied[26–34]: (1) Angle Index. The predicted grasp rectangle angle is less than 30° from the actual angle; (2) IoU Index. The intersection of the predicted grasp rectangle and the actual rectangle exceeds 25% of the union set.

4.2. Implementation Details

We conduct network construction on the PyTorch framework, and all experiments are executed on a single NVIDIA RTX 3070 GPU, AMD Ryzen 7 5800H platform. During the training process, each image is scaled before feeding into the network. Moreover, the backpropagation gradient of the network is optimized with Adam optimizer, the learning rate is 0.001 and the Batchsize is set to 16.

Table 1. Comparison Results of Existing Main Methods

Methods	Input size	Input Mode	Accuracy (%)		Inference Time (ms) (hardware env)
			Cornell	Jacquard	
GraspNet[26]	244×244	RGB-D	90.6	-	24 (NVIDIA Jetson TX1 GPU)
GN[27]	227×227	RGB-D	96.1	-	120 (NVIDIA GeForce GTX Titan X GPU)
EFCNet[32]	400×400	RGB-D	91.0	-	8 (NVIDIA GeForce GTX 1060 GPU)
ROI-GD[34]	800×600	RGB/RG-D	93.5	93.6	40 (NVIDIA GeForce GTX 1080Ti GPU)
GR-ConvNet[28]	224×224	RGB-D/ RGB-D	96.6	94.6	20(NVIDIA GeForce GTX 1080Ti GPU)
GPN-GD[30]	227×227	RGB-D	97.2	-	81 (NVIDIA GeForce RTX 2080 Ti GPU)
Q-YNet[29]	112×112	RGB-D/ RGB-D	95.2	92.1	48 (NVIDIA GeForce RTX 2080TI GPU)
TF-Grasp[33]	224×224	RGB-D/ RGB-D	98.0	94.6	41.6 (NVIDIA GeForce RTX 3090 GPU)
DFusion[31]	224×224	RGB-D/ RGB-D	98.9	94.0	15 (NVIDIA GeForce RTX 3080 GPU)
SGNet (ours)	**224×224**	**RGB-D/ RGB-D**	**99.4**	**93.4**	**12.5 (NVIDIA GeForce RTX 3070 GPU)**

Table 2. Comparison of Existing Lightweight Methods

Methods	Input Mode	Parameters	Accuracy (%)	
			Cornell	Jacquard
GG-CNN2[35]	D	66K	64.0	84.0
GraspNet[26]	RGB-D	3.8M	90.2	-
TM-DNet[36]	RGB-D	1.5M	92.5	-
DFusion[31]	RGB-D	7.2M	98.9	94.0
SGNet (ours)	**RGB-D**	**1M**	**99.4**	**93.4**

 doi:10.20944/preprints202410.0076.v1

4.3. Comparisons with the Existing Methods

The proposed SGNet is compared with the existing methods including GraspNet[26], GN[27], EFCNet[32], ROI-GD[34], GR-ConvNet[28], GPN-GD[30], Q-Ynet[29], TF-Grasp[33], DFusion[35]. TF-Grasp[33] is the first to utilize the transformer to fuse local and global features to achieve grasping detection, and DFusion[31] is a network composed entirely of convolutions to achieve generative pixel-level grasp. Table 1 shows the comparison of our method with other methods in terms of accuracy and real-time performance on the Cornell grasp dataset and the Jacquard dataset. The results indicate that our grasping method obtains 99.4% accuracy on the Cornell grasp dataset, outperforming state-of-the-art networks composed of Transformer or CNNs. Besides, the proposed method also achieves good performance on the larger Jacquard dataset where 80% of the data is used for training and 20% for validation. As shown in Table 2, our method has only 1M parameters, much less than most lightweight networks. Despite having more parameters than GG-CNN2[35], the accuracy surpasses it by 18.6% on the Cornell grasp dataset, which meets the accuracy requirement of the grasping application.

We feed the network with RGB-D images and visualize the detection results with heat maps. Figure 3 demonstrates the detection results on the Cornell grasp dataset and the Jacquard dataset. The first column is the grasp image as the output of the network and the second column is the input depth image. The third, fourth, and fifth columns show the predicted results of the network as grasp image, quality score heatmap, angle heatmap and width heatmap, respectively. It apparently demonstrates that our network is capable of yielding effective grasping detection locations for various shapes of objects. Besides, Figure 4 illustrate the predicted grasping rectangles for objects of different shapes, sizes, and colors on different datasets, which also demonstrates that the proposed method is reliable. Furthermore, we visualize the outputs of the main modules in SGNet, as shown in Figure 5. As seen in C2, the shallow features contain rich detail information, while the deep feature D2 loses certain features although the receptive field is larger. Hence, we use multilevel feature fusion, illustrated in S3, to enable the proposed network with better focus on the desired area of the task.

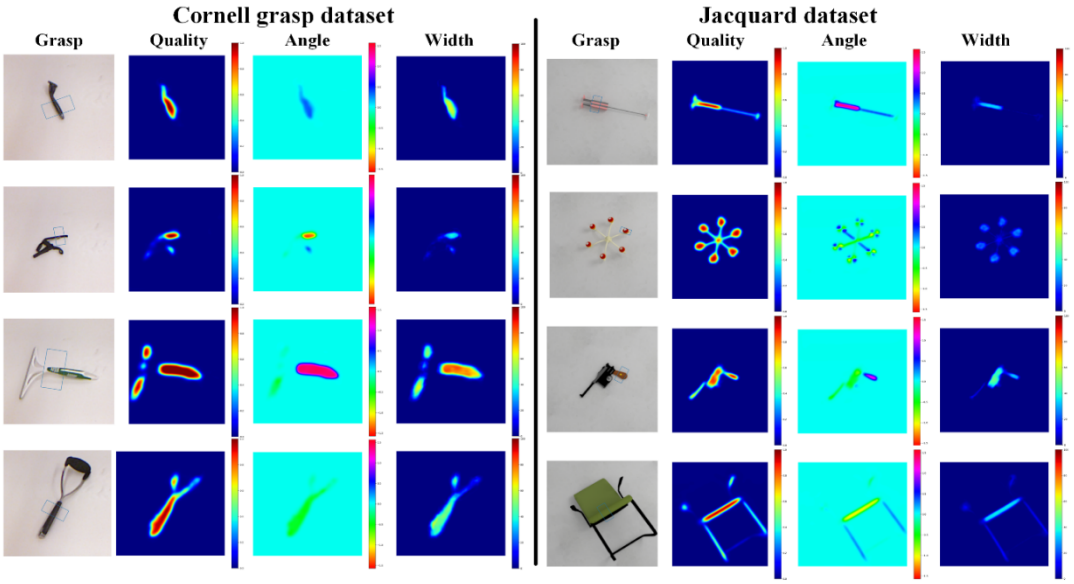

Figure 3. The detection results on Cornell grasp dataset and Jacquard dataset.

 doi:10.20944/preprints202410.0076.v1

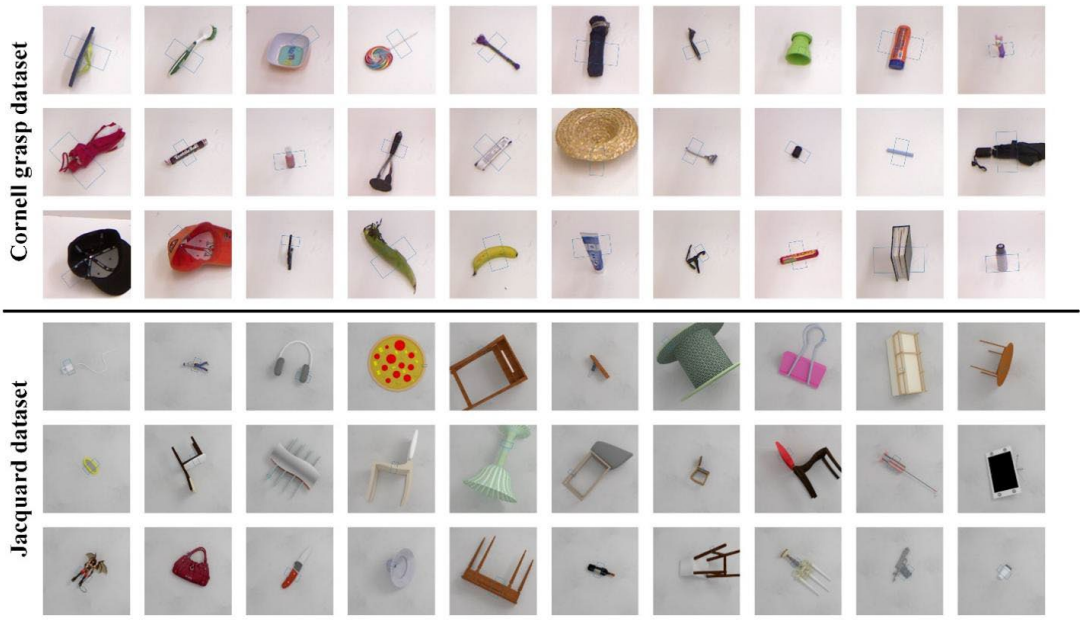

Figure 4. The predicted grasp rectangles of the SGNet on the Cornell grasp dataset and the Jacquard dataset.

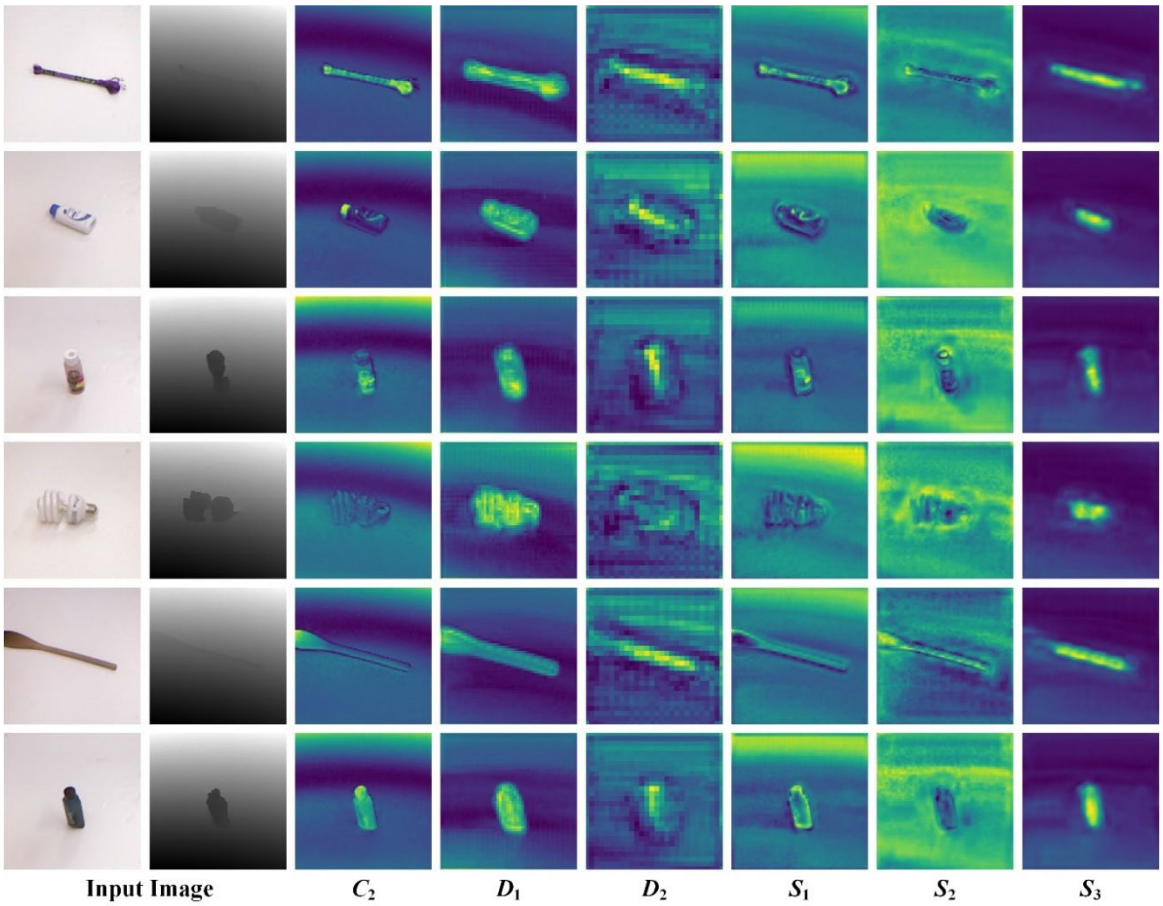

Figure 5. The visualization of the outputs of the main corresponding modules in SGNet. The first column and the second column represent the input RGB and D images, respectively, while the column C_2, D_1, D_2, S_1, S_2, and S_3 represent the output feature maps of the corresponding modules respectively.

 doi:10.20944/preprints202410.0076.v1

Table 3. Comparison of Different Variants on the Cornell Grasp Dataset

Method	MVIT	DSF	BP	Parameters	Accuracy (%)
SGNet-I	√	×	×	2,037,504	97.2
SGNet-II	×	√	×	1,047,220	97.7
SGNet-III	√	×	√	2,062,464	98.3
SGNet (ours)	**×**	**√**	**√**	**1,072,180**	**99.4**

4.4. Ablation Studies

To verify the effectiveness of our SGNet, its three variants SGNet-I, SGNet-II, and SGNet-III are constructed according to whether or not DSFormer (DSF) and Bilateral paths (BP) are involved. All variants adopt the same encoder architecture as that of SGNet. Table 3 presents the comparison results of the different variants on the Cornell grasp dataset. The SGNet-I replaces DSFormer with a MobileViT (MVIT) block in the base encoder structure, after which three up-sampling modules are cascaded. The results of SGNet-I are with the lowest accuracy. The DSF improves the performance with fewer parameters, as seen in the result of SGNet-I vs SGNet-II. Compared to SGNet-I, the BP structure in SGNet-III adds only 3K parameters and improves 1.1% accuracy. By the combination of DSF and BP, SGNet achieves the best.

5. Conclusions

Although many large language models (LLMs) nowadays show excellent capabilities, there is no such thing as a free lunch, and they require an incredible amount of computing resources and big data support. In this work, we propose a novel grasp detection deep network with one-way fusion strategy. The prominent property of SGNet is smart, which excels at the good integration of lightweight, fast inference and convenient deployment. Specifically, the DSFormer is designed to improve the ability to extract of local and global feature informatio. The spatial and context paths fully fuse the shallow and deep information in a one-way manner and achieve the supplementary property of CNN and Transformer. In the future, we will explore information fusion and alignment methods among multiple modalities to further enhance lightweight networks performance.

Author Contributions: Conceptualization, methodology, validation, formal analysis, writing—original draft preparation, visualization, Yuequan Yang, Wei Li and Xin Cang; writing—review and editing, supervision, Yuequan Yang, Jiatong Bao and Zhiqiang Cao; funding acquisition, Zhiqiang Cao. 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 numbers 61973302, 62073322, 61836015, and 62473362.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets that support the findings in this study are openly available at the following URLs : http://pr.cs.cornell.edu/grasping/rect_data/data.php and http://jacquard.liris.cnrs.fr.

Acknowledgments: The authors thank the editor and anonymous reviewers for their helpful comments, which indeed help us to improve our paper.

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

References

1. Lenz I, Lee H and Saxena A. Deep learning for detecting robotic grasps. The International Journal of Robotics Research 2015; 34: 705-724.
2. Zhou X, Lan X, Zhang H, et al. Fully convolutional grasp detection network with oriented anchor box. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2018, pp.7223-7230. IEEE.
3. Zhang H, Lan X, Bai S, et al. A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2019, pp.6435-6442. IEEE.

 doi:10.20944/preprints202410.0076.v1

4. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems 2017; 30.
5. Lin X, Sun S, Huang W, et al. EAPT: efficient attention pyramid transformer for image processing. IEEE Transactions on Multimedia 2021; 25: 50-61.
6. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929 2020.
7. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision 2021, pp.10012-10022.
8. Liu Y, Sangineto E, Bi W, et al. Efficient training of visual transformers with small datasets. Advances in Neural Information Processing Systems 2021; 34: 23818-23830.
9. Bicchi A and Kumar V. Robotic grasping and contact: A review. In: Proceedings 2000 ICRA Millennium conference IEEE international conference on robotics and automation Symposia proceedings (Cat No 00CH37065) 2000, pp.348-353. IEEE.
10. Maitin-Shepard J, Cusumano-Towner M, Lei J, et al. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In: 2010 IEEE International Conference on Robotics and Automation 2010, pp.2308-2315. IEEE.
11. León B, Ulbrich S, Diankov R, et al. Opengrasp: a toolkit for robot grasping simulation. In: Simulation, Modeling, and Programming for Autonomous Robots: Second International Conference, SIMPAR 2010, Darmstadt, Germany, November 15-18, 2010 Proceedings 2 2010, pp.109-120. Springer.
12. Krizhevsky A, Sutskever I and Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 2012; 25.
13. Li Y, Liu J and Wang L. Lightweight network research based on deep learning: a review. In: 2018 37th Chinese control conference (CCC) 2018, pp.9021-9026. IEEE.
14. Luo J-H, Wu J and Lin W. Thinet: A filter level pruning method for deep neural network compression. In: Proceedings of the IEEE international conference on computer vision 2017, pp.5058-5066.
15. You S, Xu C, Xu C, et al. Learning from multiple teacher networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining 2017, pp.1285-1294.
16. Li Y, Chen Y, Dai X, et al. Micronet: Improving image recognition with extremely low flops. In: Proceedings of the IEEE/CVF International conference on computer vision 2021, pp.468-477.
17. Touvron H, Cord M, Sablayrolles A, et al. Going deeper with image transformers. In: Proceedings of the IEEE/CVF international conference on computer vision 2021, pp.32-42.
18. Mehta S and Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:211002178 2021.
19. Lou M, Zhou H-Y, Yang S, et al. TransXNet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition. arXiv preprint arXiv:231019380 2023.
20. Hu J, Shen L and Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2018, pp.7132-7141.
21. Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2017, pp.1251-1258.
22. Lee H, Kim H-E and Nam H. Srm: A style-based recalibration module for convolutional neural networks. In: Proceedings of the IEEE/CVF International conference on computer vision 2019, pp.1854-1862.
23. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2018, pp.4510-4520.
24. Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision 2015, pp.1440-1448.
25. Depierre A, Dellandréa E and Chen L. Jacquard: A large scale dataset for robotic grasp detection. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2018, pp.3511-3516. IEEE.
26. Asif U, Tang J and Harrer S. GraspNet: An Efficient Convolutional Neural Network for Real-time Grasp Detection for Low-powered Devices. In: IJCAI 2018, pp.4875-4882.
27. Chu F-J, Xu R and Vela PA. Real-world multiobject, multigrasp detection. IEEE Robotics and Automation Letters 2018; 3: 3355-3362.
28. Kumra S, Joshi S and Sahin F. Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2020.
29. Liu D, Tao X, Yuan L, et al. Robotic objects detection and grasping in clutter based on cascaded deep convolutional neural network. IEEE Transactions on Instrumentation and Measurement 2021; 71: 1-10.
30. Ouyang W, Huang W and Min H. Robot grasp with multi-object detection based on RGB-D image. In: 2021 China Automation Congress (CAC) 2021, pp.6543-6548. IEEE.
31. Tian H, Song K, Li S, et al. Lightweight pixel-wise generative robot grasping detection based on RGB-D dense fusion. IEEE Transactions on Instrumentation and Measurement 2022; 71: 1-12.

 doi:10.20944/preprints202410.0076.v1

32. Wang S, Jiang X, Zhao J, et al. Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images. In: 2019 IEEE international conference on robotics and biomimetics (ROBIO) 2019, pp.474-480. IEEE.
33. Wang S, Zhou Z and Kan Z. When transformer meets robotic grasping: Exploits context for efficient grasp detection. IEEE robotics and automation letters 2022; 7: 8170-8177.
34. Zhang H, Lan X, Bai S, et al. Roi-based robotic grasp detection for object overlapping scenes. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2019, pp.4768-4775. IEEE.
35. Morrison D, Corke P and Leitner J. Learning robust, real-time, reactive robotic grasping. The International journal of robotics research 2020; 39: 183-201.
36. Le M-T and Lien J-JJ. Lightweight Robotic Grasping Model Based on Template Matching and Depth Image. IEEE Embedded Systems Letters 2022; 14: 199-202.