ICRA

IEEE INTERNATIONAL CONFERENCE
ON ROBOTICS AND AUTOMATION

Smaller and Faster Robotic Grasp Detection Model via Knowledge Distillation and Unequal Feature Encoding

Hong Nie, Zhou Zhao, Lu Chen*, Zhenyu Lu, Zhuomao Li, and Jing Yang

山西大学
SHANXI UNIVERSITY

山西大学大数据科学与产业研究院
Shanxi University Institute of Big Data Science and Industry

PRESENTER: Hong Nie

IEEE Robotics & Automation Society

IEEE Advancing Technology for Humanity

Background / Applications.

The application scenarios of robot arms are transforming.

Mobile Robotic Arms

Household Robotic Arms

Industrial Robotic Arms

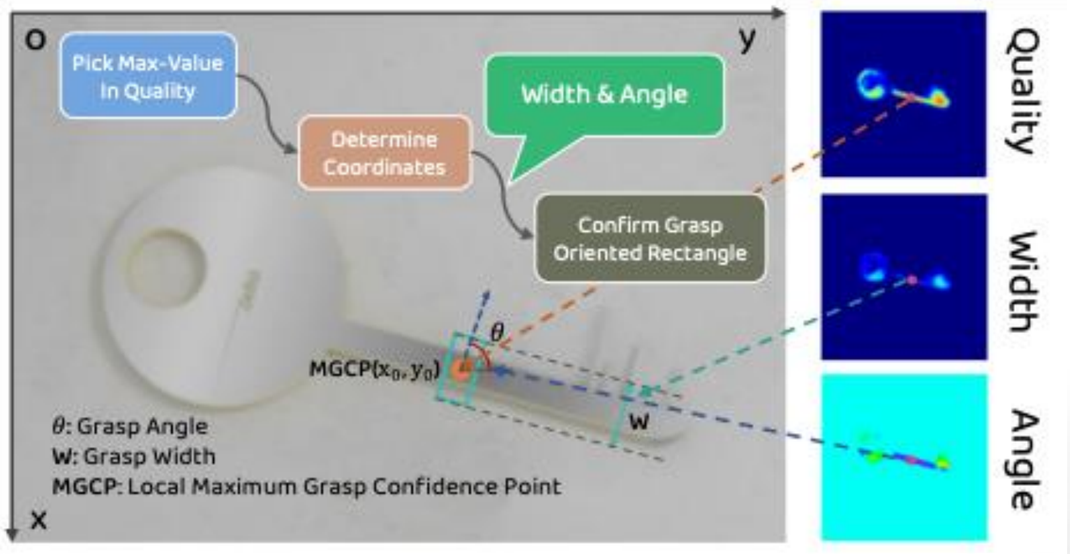
Background / Grasping Configurations.

A good grasp representation achieves twice the result with half the effort.

Grasping Rectangle Synthesis

<u>For</u>

Parallel Gripper.



Alternative Grasping Configuration.

The proposed approach of **contact point grasp** by Le *et al.* [25] may have more potential for **multi-finger grasping**. Jiang *et al.* [26] proposed a **five-dimensional rectangle representation** at the grid level, **requiring adaptive grid size adjustment** based on object dimensions to ensure grasping accuracy.

Adopted Grasping Configuration.

The enhanced **pixel-level grasp configuration** based on the five-dimensional rectangle representation proposed by Morrison *et al.* [27] is adopted.

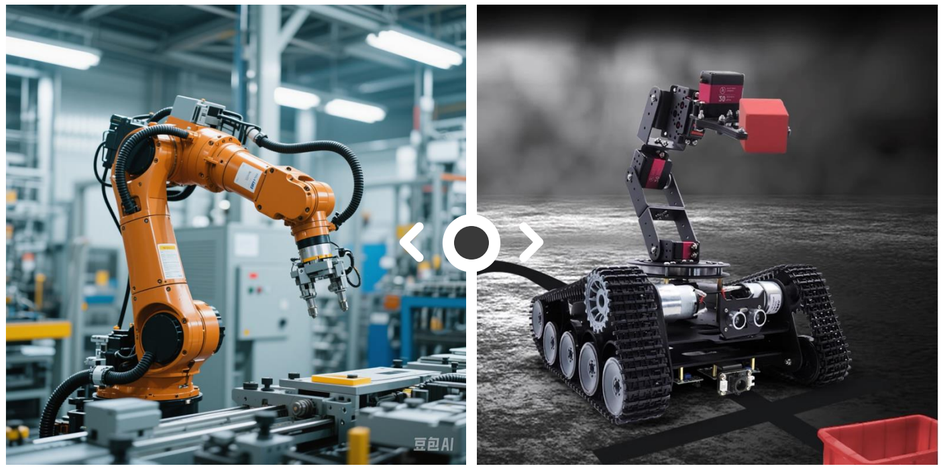
$$g = (x, y, \theta, w, q)$$

(x, y): the grasp rectangle's center, ***Angle↦ θ:*** the angle of the parallel gripper relative to the horizontal axis, ***Width↦ w:*** the width of the rectangle, and ***Quality↦ q:*** the corresponding grasp confidence.

[25] Le et al. Learning to grasp objects with multiple contact points. ICRA.
[26] Jiang et al. Efficient grasping from RGBD images: Learning using a new rectangle representation. ICRA.
[27] Morrison et al. Learning robust, realtime, reactive robotic grasping. IJRR.

Background / Task.

The terminal grasping scenarios necessitates the demand of low-consumption grasping models.

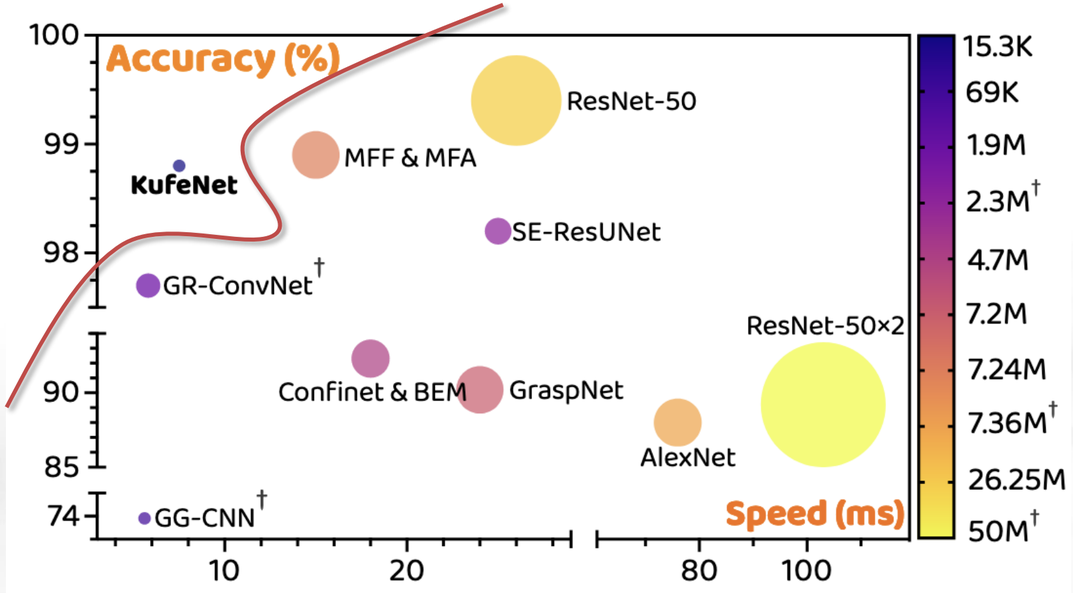


Industrial Robotic Arms ⟼ Mobile Terminal Robotic Arms

Robotic arms are undergoing a transformation from large, fixed configurations to smaller, mobile ones, **reflecting a shift in both size and flexibility**.

In mobile robotic arm systems, the **computational resources are inherently limited**, and the simultaneous operation of **multiple task subsystems**, such as visual detection and path planning, significantly constrains the computational power allocated to the grasping system. Consequently, the development of a **lightweight yet high-precision grasping detection system** has become a critical research focus.

Current grasping detection models face significant challenges in achieving an effective balance among **accuracy, lightweight design, and detection speed**.



Methods / Model Architecture.

Develop computationally efficient models optimized for edge device deployment.

> *Inspired by the widely-investigated model compression strategies in machine learning, exploring a fast, low consumed, and light-weight grasp model by fusing light-weight model design and model compression is feasible.*

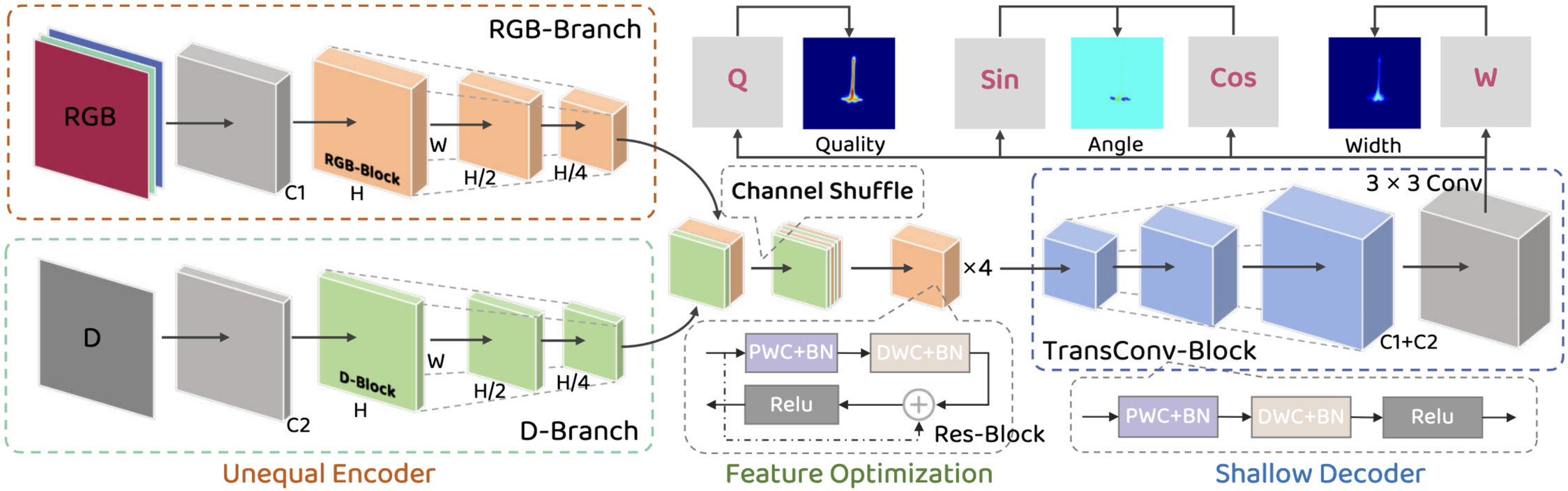


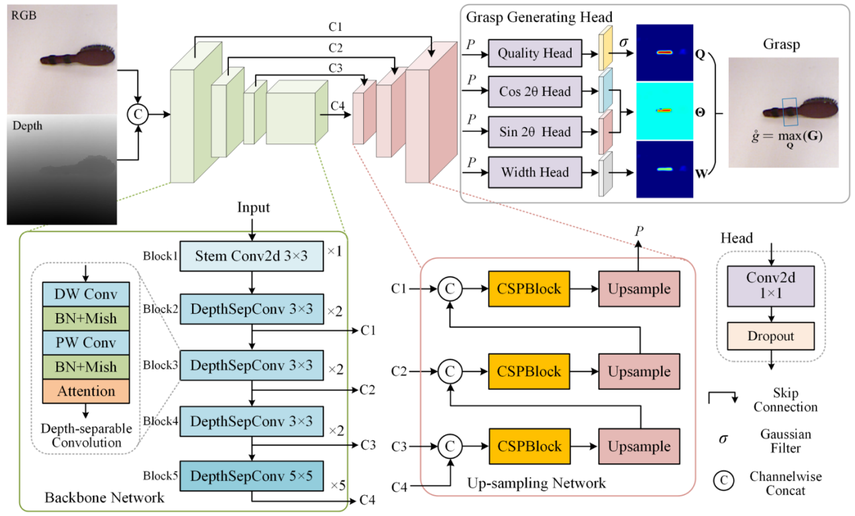
Fig. 2. Fundamental architecture of the proposed KufeNet, including unequal encoder, feature optimization and shallow decoder. In unequal encoder, RGB and D images are inputted into the corresponding blocks for different degrees of grasping feature extraction. The output feature maps are then shuffled along channel and optimized to highlight grasping-specific features in feature optimization stage. The shallow decoder finally decodes grasp configurations and generates the feasible grasping rectangles. Similar to the block in last two stages, the first stage is also based on Depth-wise Separable Convolutions (DSC).

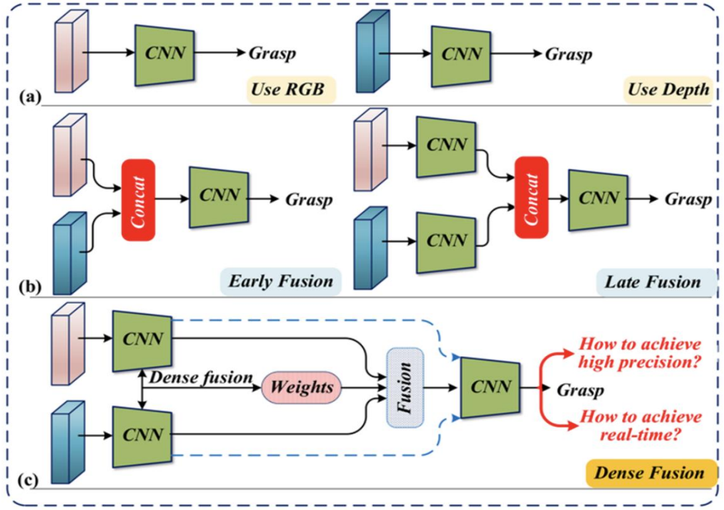
Methods / Unequal Feature Encoding.

To maximize feature extraction efficacy while minimizing parameter dimensionality.

TABLE 3
ACCURACY OF GRASP DETECTION ON CORNELL AND JACQUARD DATASETS UNDER DIFFERENT MODALITIES IN WORKS [14, 16, 28].

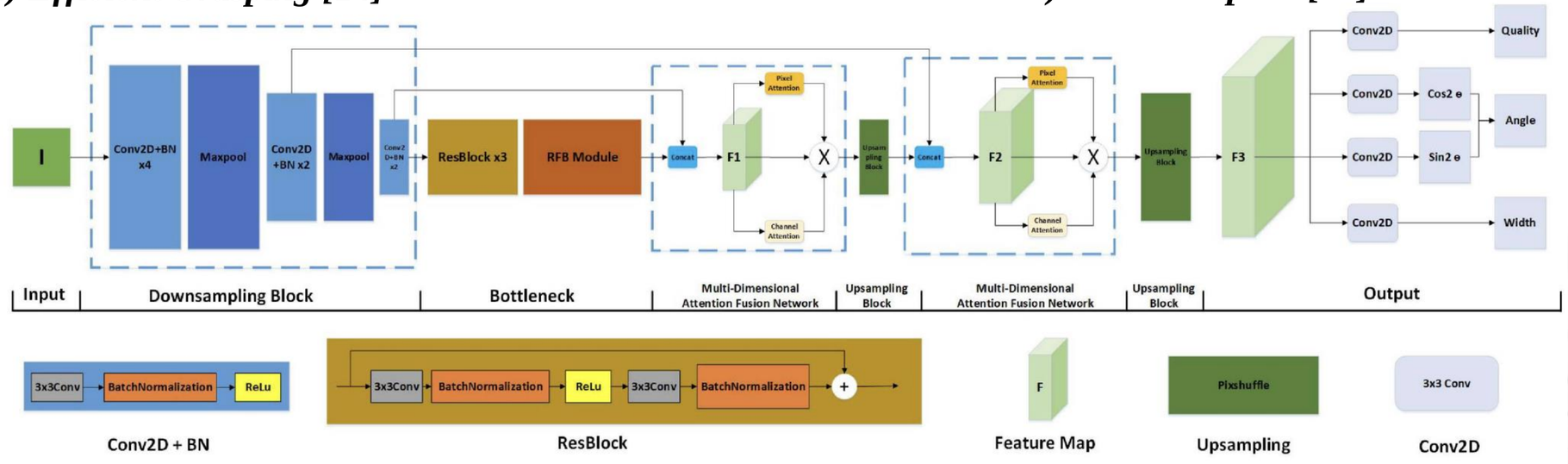
Authors	Modality	Size	Cornell (IW/OW)	Jacquard
Cao et al. [14]	Efficient Grasping-RGB	4.76M	96.6% / 91.0%	91.6%
	Efficient Grasping-D		**98.9% / 95.5%**	**95.6%**
	Efficient Grasping-RGB-D		98.9% / 97.8%	93.6%
Zhou et al. [16]	DSC-GraspNet-RGB	0.64M	97.7% / 98.3%	91.8%
	DSC-GraspNet-D		96.0% / 96.6%	**94.7%**
	DSC-GraspNet-RGB-D		98.3% / 97.7%	93.8%
Tian et al. [28]	RGB	7.24M	97.8% (IW)	92.8%
	D		**98.9%** (IW)	**94.0%**
	RGB-D		98.9% (IW)	94.0%





1) Efficient Grasping [14]

2) DSC-GraspNet [16]



3) "Dense Fusion" [28]

[14] Cao et al. Efficient grasp detection network with gaussian-based grasp representation for robotic manipulation.

[16] Zhou et al. Dsc-graspnet: A lightweight convolutional neural network for robotic grasp detection.

[28] Tian et al. Lightweight pixel-wise generative robot grasping detection based on rgbd dense fusion.

Methods / Unequal Feature Encoding.

To maximize feature extraction efficacy while minimizing parameter dimensionality.

A simple strategy to improve grasping accuracy !

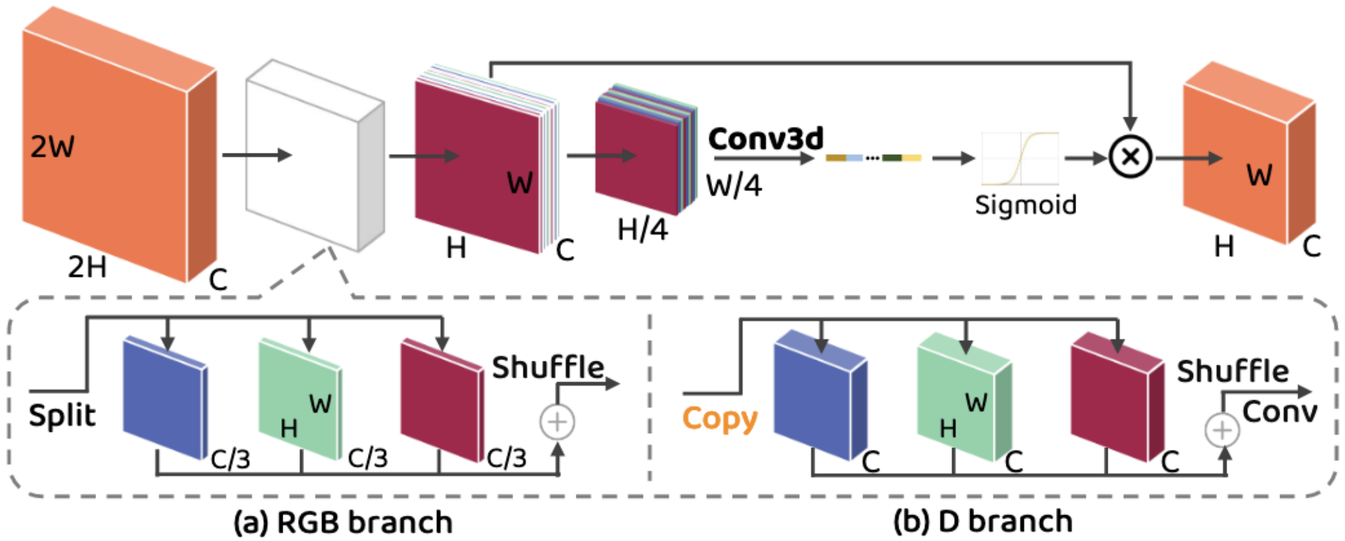
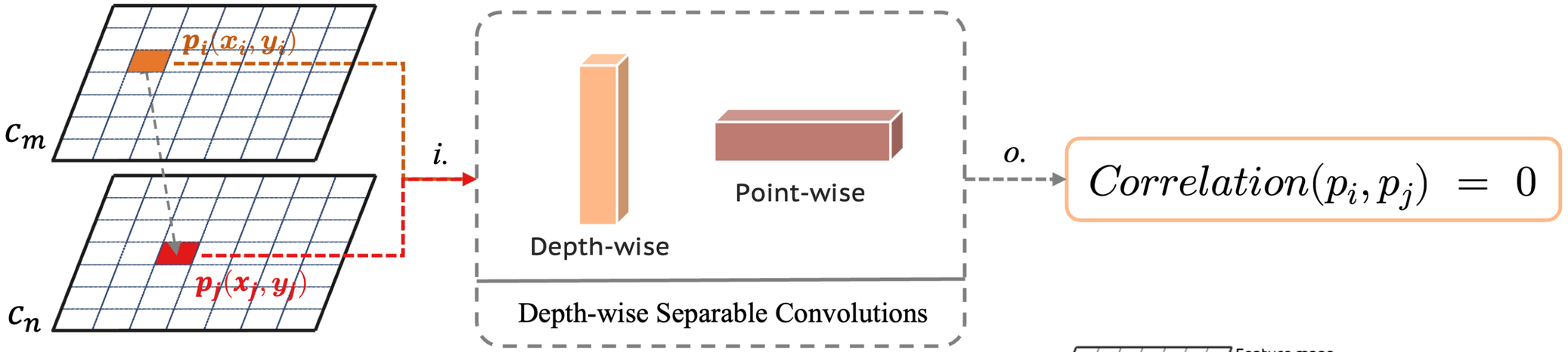


Fig. 3. Composition of the proposed RGB(D)-Block. For RGB branch, the channels are split and then convolved with multi-scale convolution kernels, while for D branch, the copy operation is used instead.

Unequal Feature Encoding Module

Methods / Information Compensation.

Compensation for cost of lightweight architecture.



Due to the extensive use of DSC (Depth-wise Separable Convolution) throughout the network, the correlation between **non-adjacent features** in adjacent feature maps is difficult to be exploited.

The traditional SE-Attention module [29] is rather lightweight, but it makes limited use of spatial features.

We propose the **3D convolution** to generate weighted coefficients, which are based on the corresponding whole feature maps.

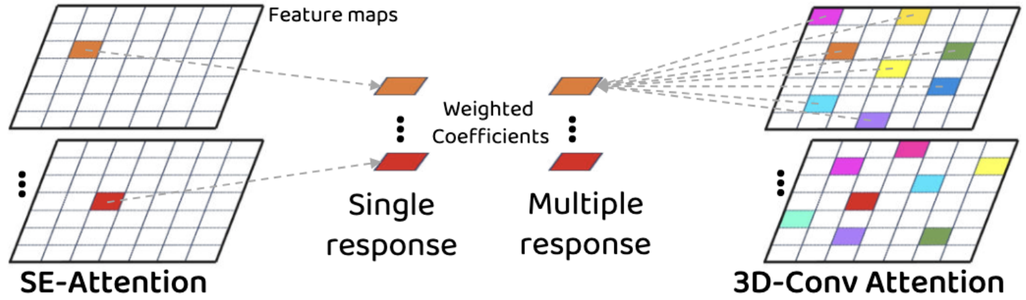


Fig. 4. Difference between 3D-Conv Attention and SE Attention. The former performs convolutions on feature maps to generate weighted coefficients, whereas the latter simply picks the maximum activation value (or the average) from feature maps. 3D Convolution utilizes spatial information more adequately and compensates for the correlation loss caused by DSC.

[29] Hu et al. Squeeze-and-excitation networks. CVPR.

Methods / Grasping Feature Distillation.

Knowledge distillation acquires additional grasping features.

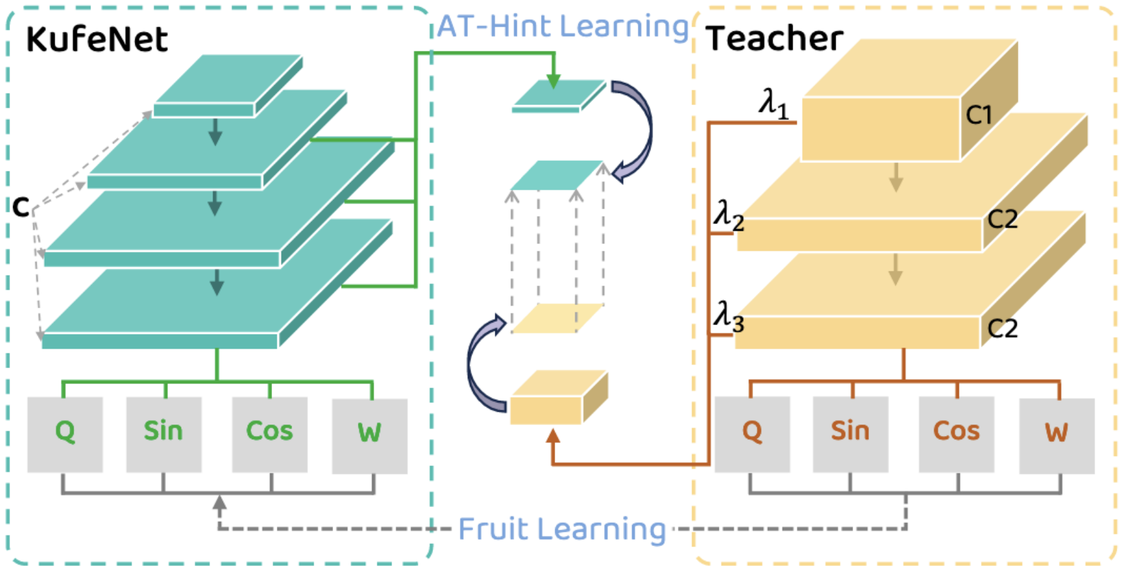


Fig. 5. Illustration of knowledge distillation strategies in KufeNet. For fruit learning, Teacher guides KufeNet at the grasp configuration level, while for hint learning, each layer of the Teacher's decoder guides the corresponding layer of KufeNet's decoder. λ_1, λ_2, λ_3 represent guidance weights and set to $Softmax(0.5, 1, 1.5)$ respectively.

Basic Loss.

We define the loss between the grasp configurations predicted by T/S and the ground truths as $H(y_t, p.)$:

$$\mathcal{H}(y_t, p.) = \mathcal{L}_{Quality} + \mathcal{L}_{Angle} + \mathcal{L}_{Width}$$

Fruit Learning.

$$\mathcal{L}_{fruit} = \mathcal{L}_2(p_T, p_S)$$

where p_T and p_S are the grasp configurations predicted by T and S respectively. $\mathcal{L}_2(\cdot)$ is defined as $\mathcal{L}_2$ loss. $\mathcal{L}_{fruit}$ is used to monitor the progress of knowledge transfer.

Hint Learning.

Transforming each layer output activation tensor $X \in R^{C\times H\times W}$ to a spatial attention map $A \in R^{H\times W}$ using a mapping function $\mathcal{F}$:

$$\mathcal{F}: X \in R^{C\times H\times W} \rightarrow A \in R^{H\times W}$$

To restrain T from imposing excessive constraints on S, we substitute the pixel-based loss function $\mathcal{L}_2$ loss with the distributionbased maximum mean discrepancy loss $\mathcal{H}_{MMD}$:

$$\mathcal{L}_{hint} = \mathcal{H}_{MMD}(A_T, A_S)$$

Grasping Feature Distillation Regularization Term.

$$\mathcal{L}_{kafe} = \begin{cases} \mathcal{H}(y_t, p_S), & \text{if } \mathcal{H}(y_t, p_S) \le \mathcal{H}(y_t, p_T) \\ \mathcal{H}(y_t, p_S) + \alpha\,\mathcal{L}_{fruit} + \beta\,\mathcal{L}_{hint}, & \text{else.} \end{cases}$$

Experiments / Quantitative Results.

Performance comparison with existing models in single object, cluttered and stacked scenarios.

TABLE I
PERFORMANCE COMPARISON RESULTS OF DIFFERENT APPROACHES ON CORNELL DATASET.

Authors	Algorithm	Speed (ms)	Params	Accuracy (%)	
				IW	OW
GraspNet [31]	RGB-D	24	7.2M	90.2	90.6
Morrison [8]	GG-CNN	5.6†	69K	73.9	69.0
Kumra [9]	GR-ConvNet	5.8†	1.9M	97.7	96.6
Wu [32]	ResNet-50	26	26.25M	99.4	98.9
Yu [12]	SE-ResUNet	25	2.3M	98.2	97.1
Cao [14]	Efficient Grasp	6	1.2M	97.8	-
Ours	Teacher	27.1	8.7M	98.9	97.8
	Vanilla KufeNet	**7.5**	**15.3K**	97.7	95.5
	KufeNet			98.9	96.6

† indicates the results tested by us. IW: image-wise; OW: object-wise.

TABLE II
PERFORMANCE COMPARISON RESULTS OF DIFFERENT APPROACHES ON JACQUARD DATASET.

Authors	Algorithm	FLOPs	Params	Accuracy (%)
Morrison [8]	GG-CNN	1.0G	69K	84.0
Kumra [9]	GR-ConvNet	10.9G	1.9M	91.9†
Yu [12]	SE-ResUnet	24.78G	2.3M	92.9†
Cao [14]	Efficient Grasp	5.7G	1.2M	93.6
Ours	Teacher	272.7G	8.7M	93.2
	Vanilla KufeNet	**3.8G**	**80.0K**	92.2
	KufeNet			93.1

TABLE III
PERFORMANCE COMPARISON RESULTS OF DIFFERENT APPROACHES ON GRASPNET AND MULTIOBJ DATASETS.

Authors	Algorithm	Params	Accuracy (%)	
			GraspNet	MultiObj
Kumra [9]	GR-ConvNet†	1.9M	79.6	90.0
Yu [12]	SE-ResUNet†	2.3M	81.8	90.0
Ours	Teacher	8.7M	84.2	90.0
	KufeNet	**263K / 15.3K**	**82.3**	**90.0**

Experiments / Quantitative Results.

Robustness Experiments, Ablation Experiments, and Terminal Deployment Comparisons.

TABLE IV
COMPARISON OF DIFFERENT APPROACHES ON CORNELL/JACQUARD DATASETS UNDER DIFFERENT ROTATION ANGLE THRESHOLDS.

Algorithm	$\Delta\theta = 25°$	$\Delta\theta = 20°$	$\Delta\theta = 15°$	$\Delta\theta = 10°$
GR-ConvNet†	96.6 / 86.1	92.1 / 79.3	87.6 / 74.3	78.6 / 66.2
SE-ResUNet†	96.6 / 91.2	95.5 / 82.0	87.6 / 77.4	80.8 / 70.9
KufeNet	**97.7 / 91.8**	**96.6 / 86.3**	**88.7 / 83.0**	**82.0 / 76.9**

TABLE V
PERFORMANCE COMPARISON RESULTS OF DIFFERENT DISTILLATION STRATEGIES FOR OUR KUFENET.

KD Strategy	Mapping	Accuracy (%)
-	-	**92.18**
Fruit Learning	-	92.38
AT-Hint Learning	$mean$	92.04
	max	92.29
	max^2	**92.46**
	max^4	92.15
FT-Hint Learning	PWC	92.31
Fruit, AT-Hint Learning	max^2	**93.05**

TABLE VI
ABLATION EXPERIMENTS ON JACQUARD DATASET.

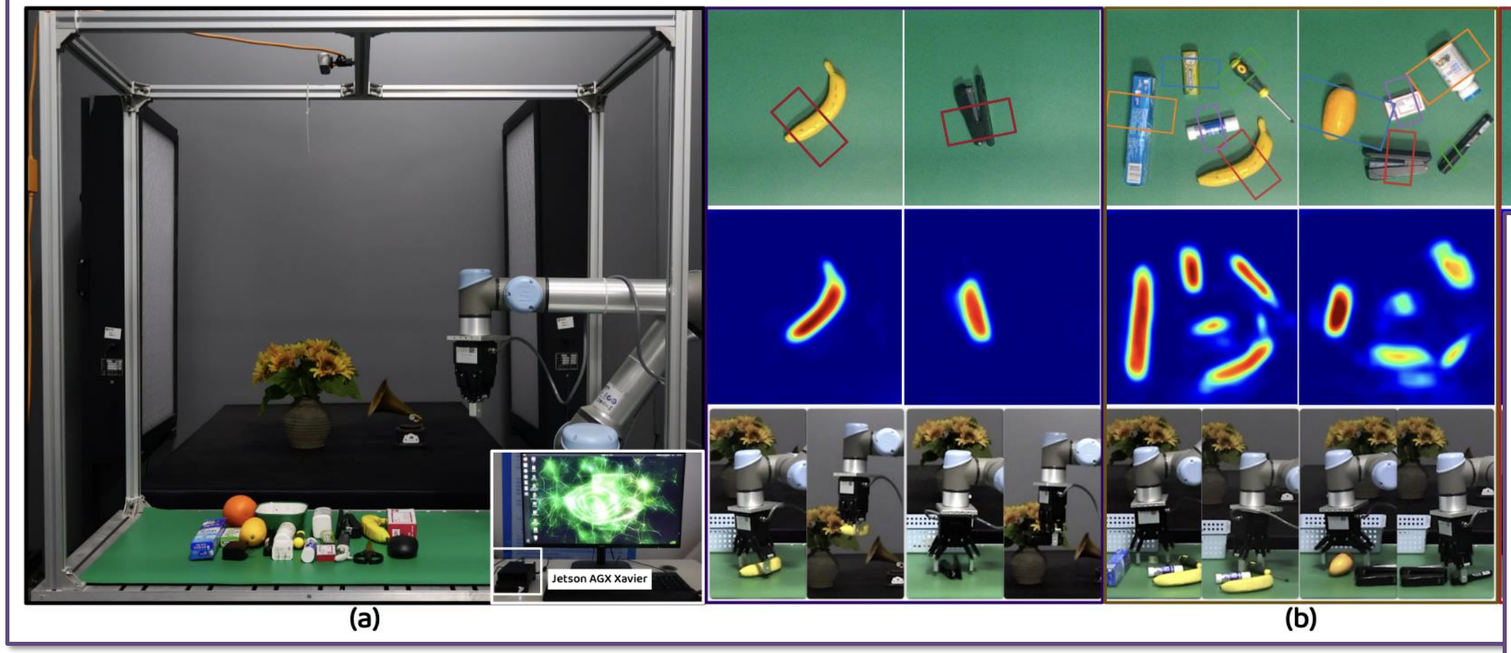
Modality	UE	IC		Params	Speed (ms)	Accuracy (%)
		CS	CA			
RGB	-	✓	✓	81.7K	11.2	86.23
D	-	✓	✓	78.9K	10.4	90.83
RGB-D	×	×	×	79.4K	13.9	88.77
RGB-D	×	✓	✓	85.6K	14.1	90.76
RGB-D	✓	×	×	74.9K	11.6	90.38
RGB-D	✓	✓	×	74.9K	11.7	90.77
RGB-D	✓	×	✓	80.0K	13.3	92.09
RGB-D	✓	✓	✓	80.0K	13.3	92.18

TABLE VII
TRAINING TIME, SPEED AND ACCURACY COMPARISON OF DIFFERENT GRASP DETECTION METHODS ON EMBEDDED AI DEVICE.

Algorithm	Training (h)	Speed (ms)	Accuracy (%)	AI Device
GraspNet	-	133	90.2	Jetson TX1
GG-CNN	18.4	41.65†	73.9	Jetson AGX Xavier
GR-ConvNet	24.0	54.58†	97.7	Jetson AGX Xavier
KufeNet	**16.1**	48.65	**98.9**	Jetson AGX Xavier

Experiments / Qualitative Analysis.

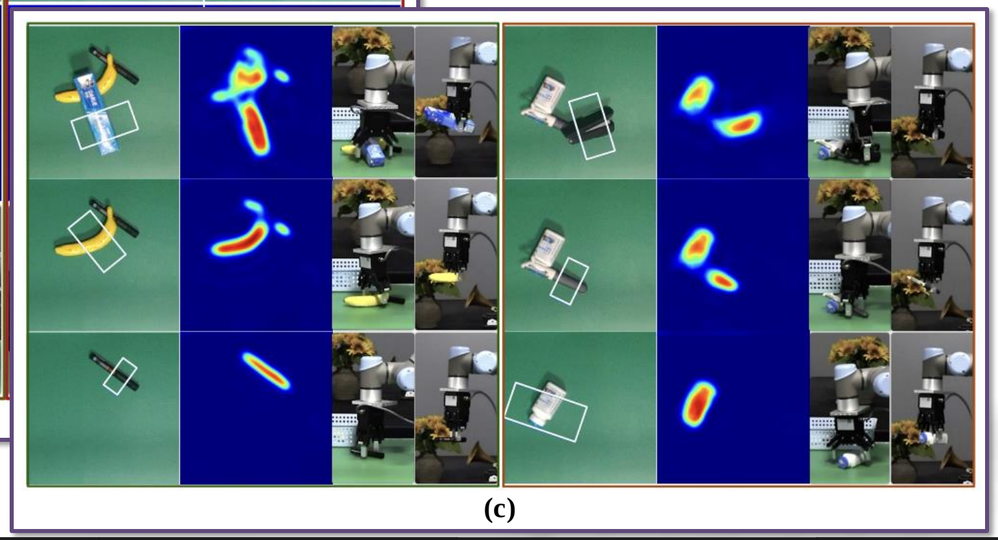
Visualization of grasping experiments in single object, cluttered and stacked scenarios.

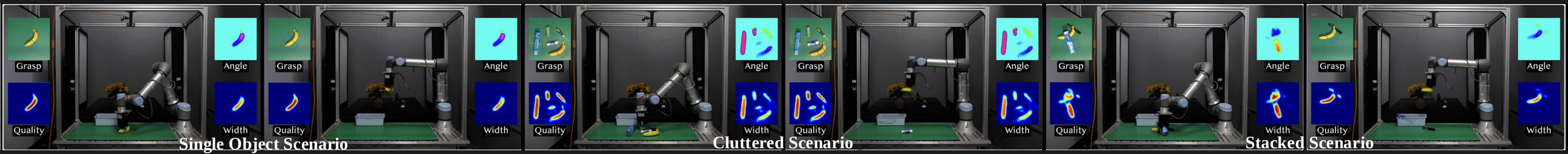


(a) Robotic grasping system based on an embedded AI computing device.

(b) The grasping demonstrations in single object and cluttered scenarios..

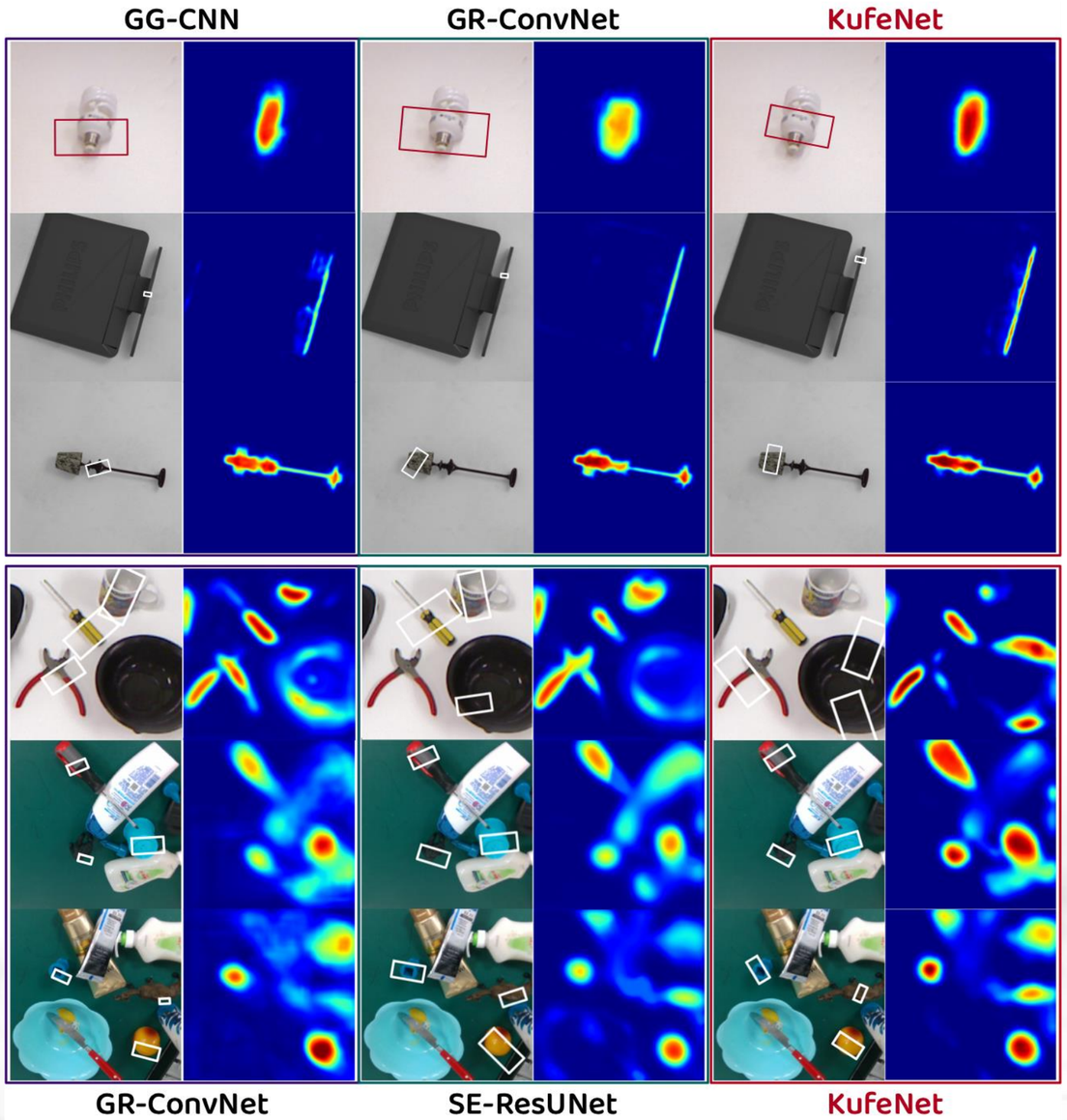
(c) The grasping demos in stacked scenarios.

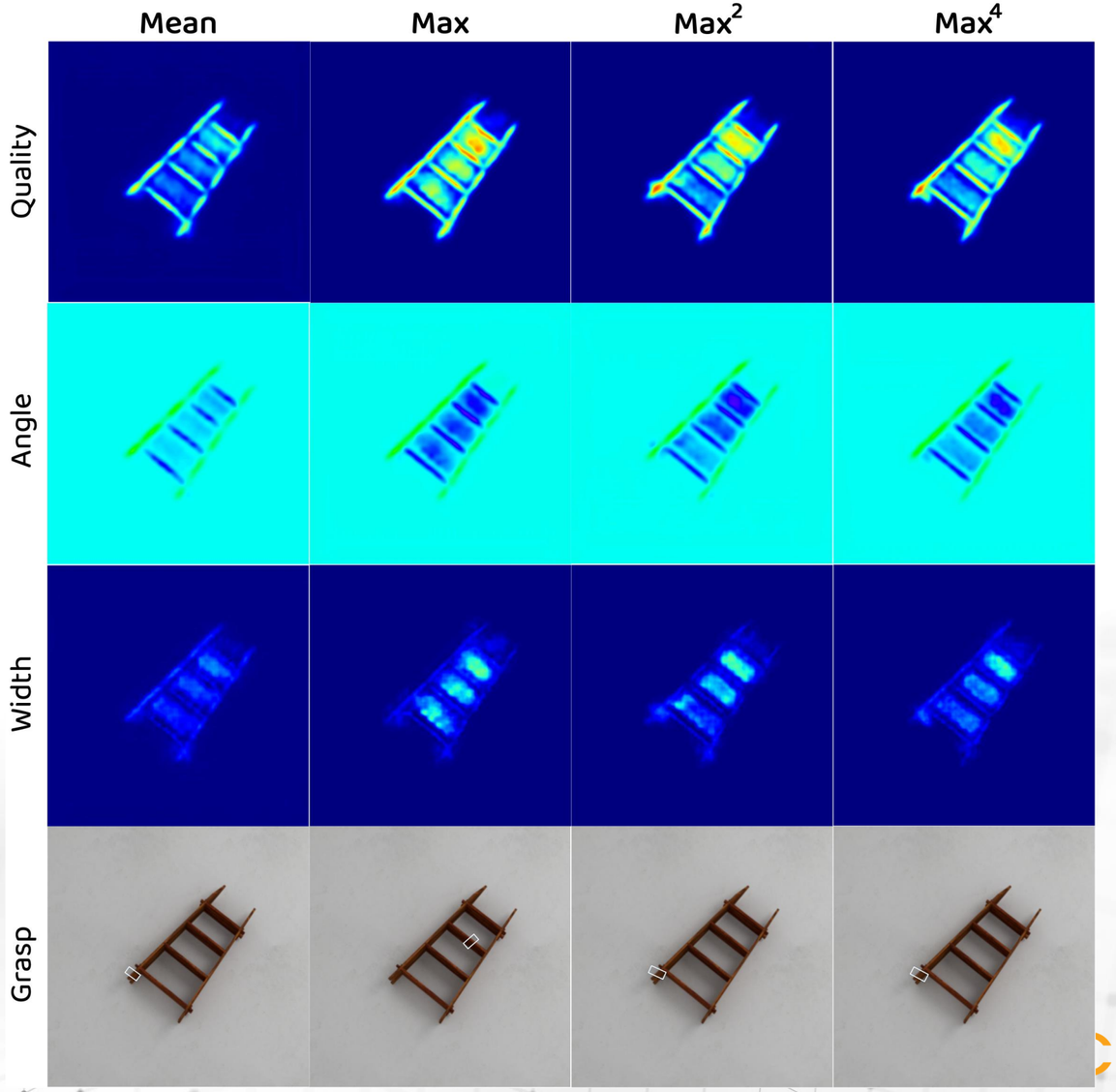




Experiments / Qualitative Analysis.

Visualization comparisons and the impact of various mapping strategies in Attention Transfer.





Experiments / Real-world Grasping.

Real-world grasping demonstration.

Smaller and Faster Robotic Grasp Detection Model via Knowledge Distillation and Unequal Feature Encoding

Hong Nie, Zhou Zhao, Lu Chen*, Zhenyu Lu, Zhuomao Li, and Jing Yang

An Interesting Experimental Setup.

Elucidate the efficacy of lightweight grasping models in some specific contexts.

The experimental configurations:

1. A quarter of the Cornell dataset (old domain) and the entire MultiObj dataset (new domain). These two datasets belong to different domains.
2. Three models to be tested, including GG-CNN, GR-ConvNet and KufeNet.

The experimental details:

1. The models are first trained using a quarter of Cornell dataset and the best performing one for each model is selected as the pre-trained model for later tasks.
2. We assume that it takes 30s for the mobile robot to perform a grasp, including grasp detection, path planning, grasping and placement, etc., and the size of the first small training batch of new object is set as 20, namely, the pre-trained model needs to be re-trained after inferring 20 new objects. The size of the small training batch of seen object is set as 200 (nearly a quarter of the Cornell dataset), which are randomly selected from the seen samples.
3. The size of the next small training batch of new object is determined by the training time. If the training of 220 (200 old images and 20 new images) old and new samples is completed successfully within 10 minutes (30s $\times$ 20), the next small training batch of new object is still 20. The converse is the number of all inferred images during that time.
4. The shorter the time it takes to train on the mixed domain, the better the chances of achieving higher accuracy during later inference on samples from the new domain.

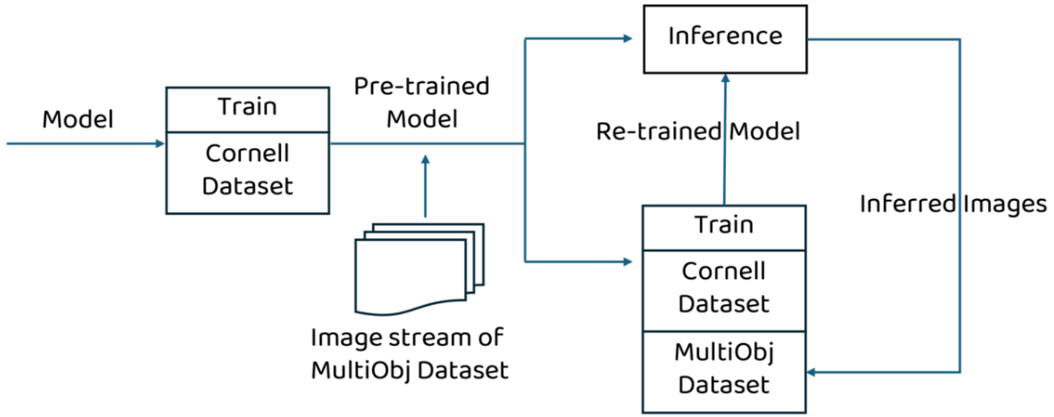


TABLE I
COMPARISON OF MODEL ACCURACY AND TOTAL TIME IN A CONTINUAL LEARNING SCENARIO.

Mode	Accuracy / Time (min)		
	KufeNet	GR-ConvNet	GG-CNN
Inference (0-20 samples, 1)	4/20	2/20	1/20
Training (1)	7.6	12.8	9.2
Inference (2)	8/20	3/26	4/20
Training (2)	7.2	13.1	8.6
Inference (3)	10/20	5/27	6/20
Training (3)	7.7	12.9	8.7
Inference (4)	13/20	6/24	7/20
Training (4)	7.7	12.6	8.4
Inference (5)	12/17	-	6/17
Training (5)	7.1	-	8.2
Accuarcy	**47/97**	16/97	24/97
Total Time	**37.3**	51.4	43.1

Contributions.

1. Considering the discrepancy of information contained in RGB and D images, we propose an unequal architecture to process RGB and D images in parallel, where more parameters are used to learn the features in D image. To the best of our knowledge, it is the first work to explore how to unequally encode RGB and D features with limited number of model parameters in grasp detection task. In addition, the light-weight grasp detection framework is constructed by integrating multiscale feature extraction, channel shuffle, and context channel weighting, leading to fast speed and accurate performance.

2. Apart from constructing the model with light-weight designs, we use knowledge distillation to ensure that the proposed network achieves performance similar to a complex network in terms of grasping feature extraction. Specifically, we adopt fruit learning to achieve the explicit configuration-level transfer of grasp configurations, and utilize hint learning to achieve the implicit feature-level transfer of grasping features within model layers. Our work further demonstrates the applicability of KD for such a dense estimation task of 2D planar grasp detection.

3. The proposed **KufeNet** achieves the accuracy of 98.9% and 93.1% on Cornell and Jacquard datasets with much fewer parameters (15.3K and 80.0K respectively). In more challenging grasping scenarios, KufeNet also performs well to achieve 82.3% and 90.0% accuracy on GraspNet and MultiObj datasets with 263K and 15.3K parameters. In addition, extensive comparisons on embedded AI computing device and realworld robotic grasping scenario are also conducted to prove the effectiveness of the proposed KufeNet.

Thanks for your Attention!

Smaller and Faster Robotic Grasp Detection Model via Knowledge Distillation and Unequal Feature Encoding

Hong Nie, Zhou Zhao, Lu Chen*, Zhenyu Lu, Zhuomao Li, and Jing Yang