IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.
Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

A Novel Robotic Grasp Framework for Accurate Grasping under Complex Packaging Factory Environments (september 2024)

Guirong Dong[1], Fuqiang Zhang[1], Xin Li[1], Zonghui Yang[1], and Dianzi Liu[2,3]

[1]Faculty of Printing, Packaging and Digital Media Technology, Xi'an University of Technology, Xian 710058, China.
[2]School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, 350108, P.R. China.
[3]School of Engineering, University of East Anglia, NR4 7TJ Norwich, U.K.

Corresponding author: Guirong Dong (dongguirong2005@xaut.edu.cn).

This work was supported by Collaborative Innovation Center of Shaanxi Provincial Education Department (No. 23JY064), Special Project for Talent Cultivation in Western Region from China Scholarship Council (No. 2208615060).

ABSTRACT As grasping behaviors in real packaging scenarios are apt to be influenced by various disturbances, visual grasping prediction systems have suffered from the poor robustness and low detection accuracy. In this study, an intelligent robotic grasp framework (RTnet) underpinned by a linear global attention mechanism has been proposed to achieve the highly robust robot grasp prediction in real packaging factory scenarios. First, to reduce the computational resources, an optimized linear attention mechanism has been developed in the robotic grasping process. Then, a local window shifting algorithm has been adapted to collect feature information and then integrate global features through the hierarchical design of up and down sampling. To further improve the developed framework with the capability of mitigating noise interference, a self-normalizing feature architecture has been established to empower its robust learning capabilities. Moreover, a grasping dataset in the real operational environment (RealCornell) has been generated to realize a transition to real grasping scenarios. To evaluate the performance of the proposed model, its grasp prediction has been experimentally examined on the Cornell dataset, the RealCornell dataset, and the real scenarios. Results have shown that RTnet has achieved a maximum accuracy of 98.31% on the Cornell dataset and 93.87% on complex RealCornell dataset. Under the consideration of real packaging situations, the proposed model have also demonstrated the high levels of accuracy and robustness in terms of grasping detection. Summarily, RTnet has provided a valuable insight into the advanced deployment and implementation of robotic grasping in the packaging industry.

INDEX TERMS Attention Mechanism, Packaging Factory, Robot Grasping, Stylistic Reconstruction.

I. INTRODUCTION

Packaging factory, as a typical discrete manufacturing industry, has been an indispensable part of industrial development. With the introduction of industrial automation, intelligent grasping robots are widely used in packaging factories due to their high efficiency and ease of management, enabling the grasping robots to replace traditional human hands for daily packaging and handling work [1][1]. However, due to the limitations of perception, robots are not able to recognize objects and understand the spatial layout of the target in the same way as people perform. That is the reason why robots do not work well and precisely when complex scenarios such as packaging factory, must be considered or when the grasping unknown objects is required. Deep learning enables robots to intelligently learn, allowing grasping robots to predict grasping points autonomously without human assistance [2]. The popular method for representing grasping points is achieving the rectangle representation grasp [3]. For grasping tasks, an accurate point cloud segmentation is essential. Techniques to achieve this goal are categorized into deep learning-based or clustering-based approaches [4]. The existing methods for robot grasping include the parallel grasping method and 6D pose estimation [5]. However, in the packaging factory

environment, parallel grasping methods are more efficient. Most existing deep learning-based grasping detection methods rely on convolutional neural networks to extract features and map them between the object being grasped and the predicted grasping rectangle [6]. Recently initially proposed SSDResNet, a composite network for object detection. Subsequently, the identified targets were utilized to utilized four-dimensional grasp predictions [7]. To tackle the issue of the inefficient computation caused by candidate bounding boxes extraction, GGCNN [8], a typical closed-loop single-stage grasp prediction network, was developed to directly generate grasp poses on pixels and achieved a lightweight representation of grasp predictions. Cheng et al. [9] regarded the grasp pose as a rotating enclosing frame in the image plane and proposed a single-stage fully convolutional grasp generation network, which eliminate the intermediate grasp candidate stage and achieves accurate pixel-level grasp directly.

Research on robotic grasp prediction belongs to the field of computer vision, and the above methods have achieved excellent performance. Nevertheless, determining an appropriate grasping posture for a captured object entails a process of extensive information search. CNN, while suitable for target detection and object recognition, lack the ability to achieve remote modeling and extract global features. As a result, these networks cannot effectively utilize the continuity and correlation of grasping posture for feature extraction. The attention mechanism with its global interaction capability provides a solution to this issue.

The attention mechanism was initially introduced in the transformer [10]. The powerful global interaction and remote modeling capabilities have garnered significant attention in the field of natural language processing and image processing. Compared to CNN, transformer captures more spatial and contextual information. The well-known transformer vision networks include VIT [11], and Swin Transformer [12]. Apart from the above research, many scholars have also applied attention mechanisms to robotic grasping tasks. With the development of point cloud segmentation technology, 6D pose estimation method has been widely applied in the field of robot grasping [13]. Zou et al. [14] proposed a transformer-based 6D vision transformer, primarily utilized for estimating target poses on RGB-D images to achieve high-precision robotic grasping. Wang et al. [15] applied the global attention mechanism to enhance robot grasping prediction task and utilized local window attention to extract local information, achieving a remarkable accuracy of 97.99% on the Cornell Grasp Dataset.

In industrial settings, robots require visual grasping capabilities that exhibit high model accuracy and robustness to successfully execute a variety of complex grasping tasks [16]. Jiang et al. [17] converted the six degrees of freedom grasping attitude estimation task into a two-dimensional registration problem and achieved its application in industrial parts grasping with reflective surfaces through information coding and feature alignment. Ge et al. [17] improved the accuracy of grasping prediction by assigning categories to each pixel, and utilizes residual pyramid feature module to achieve accurate grasping prediction of medical devices in unstructured scenes. Wei et al. [19] proposed a robust, two-stage grasping attitude network that fine-tuned the low- quality grasping and reduced local noise to enhance the estimation of object grasping attitudes in complex settings. Niu et al. [20] proposed a visual enhanced grasping detection model (VERGNet) to improve the robustness of robot grasping in low light imaging scenes, in response to poor grasping prediction performance under low light conditions.

Current research is focused on enhancing the performance and robustness of models in complex scenarios by improving algorithmic robustness. However, there has been limited investigation into samples across various scenarios. Since most of grasp detection methods on deep learning are trained and tested in the laboratory settings, their high accuracy during training using simulated data often fails to translate into effective performance in industrial settings. Therefore, the limited scope of the training dataset and the presence of various types of noise significantly compromise data quality, making it challenging to obtain high-resolution, single background images in real factory scenarios as those found in Cornell datasets. In this case, one possible solution is to enhance the diversity of the training dataset by introducing noise from the real industrial capture settings. This approach not only improve the performance of the model in the packaging factory grasping environment but also ensures its adaptability to various situations.

In this paper, focusing on poor generalization and low accuracy of robot grasping under the environment of producing the real packaging production, especially in industrial packaging and grasping applications, a highly robust robot grasping detection model (RTnet) is proposed. The model not only inherits the Swin Transformer mechanism for extracting local features through window sliding, but also linearly reduces attention calculation while ensuring global feature extraction. With weight initialization assumed, scaled exponential linear units (selus) are applied to endow the model with self-normalization property, thereby, enhancing its robustness in complex settings. Subsequently, a U-shaped architecture is employed to endow the model with ability to learn detailed features. Additionally, a practical dataset (RealCornell) is generated through stylistic transfer of the original Cornell dataset, which captures real-world packaging factory grasping scenarios and enhances the robustness and generalization capability of RTnet.

II. THE PROPOSED RTNET FRAMEWORK

A. GRASP REPRESENTATION

Accurate determination of the grasping position is a prerequisite for robotic manipulation. In the case of two-finger grasping, a five-dimensional formulation [3] is defined to

transform robot grasp into representative rectangles as shown in (1).

$$g = (x, y, w, h, \theta) \tag{1}$$

where (x, y) is the center of the grasping rectangle, (w, h) is the grasping width and the parallel grippers width, θ indicates the angle of the grasping rectangle for the horizontal axis.

For a 2D image, the rectangular grasp of each pixel points (x, y) for a known width of the parallel fixture as $G(x, y)$ is formulated in (2).

$$G_{(x,y)} = (Q, w, \phi) \in R^{n \times W \times H} \tag{2}$$

where Q , the grasp quality, indicates the success rate of capturing per pixel, with a value range of [0,1]. The closer the value is to 1, the higher the success rate of grabbing. w is the width of each pixel location at the time of capture. In real scenarios, width is defined within a range of [0,150] pixels. ϕ is the orientation angle within the range of -90 to 90.

B. RTNET FRAMEWORK

To achieve effective grasping in real packaging industrial settings, an efficient and robust robot grasping detection framework (RTnet) is proposed in this section. Based on Swin-transformer, the developed framework comprises four modules including Linear Embedding, Encoder, Decoder, and Linear Project. The Encoder and Decoder are connected by a skip-connection and Robust Transformer Block (RT Block) as shown in Fig. 1. Skip-connections facilitate the direct transfer of multi-scale feature information extracted by the encoder to the decoder. By means of feature graph fusion, this connection approach effectively integrates the original encoder features with those obtained after up-sampling in the channel dimension, thereby compensating for information loss caused by down-sampling and restoring crucial spatial information. The incorporation of skip-connections not only enhances the model's ability to capture features across different scales, leading to the improvement of prediction accuracy, but also expedites both training and reasoning processes by alleviating computational burden on the decoder. A detailed exposition of the model is provided below.

1) IMAGE SEGMENTATION AND LINEAR EMBEDDING

The image information will be partitioned into numerous small blocks by RTnet and each block represents adjacent and non-repeating pixels. These blocks are then expanded based on the channel direction. Subsequently, the segmented small block images are inputted into a linear embedding layer for dimensionality conversion, ensuring efficient feature extraction and speeding up the process of subsequent data.

2) ENCODING AND DECODING STAGE

The encoding-decoding stage is the core component of the entire model and an important phase for feature extraction. Referring to the U-shaped network architecture, there are two parts contained in the encoding and decoding stages of the RTnet, respectively. The encoding stage include a Robust Transformer Block (RT Block) and a Feature Merging Block (Patch Merging). In the decoding stage, there is also a Robust Transformer Block to improve the robustness of model. In addition, Feature Expansion Block (Patch Expanding) is also implemented.

Linear Embedding divides the input image into non-overlapping Patches of size 4×4 . Taking RGB image as an example, each Patch with a feature dimension of $4 \times 4 \times 3$ is linearly mapped to a linear vector of size 96. The encoder section employs RT block to perform attention and shift window mechanism, which are introduced in the following Section C. Additionally, the Patch Merging section gradually expands channel size to achieve downsampling functionality. With the synergy of these two modules, RTnet achieves multi-scale feature extractions. The decoder section is upsampled and comprises a RT block and Patch Expanding, which reshape the feature map into a high-resolution one while halving the feature dimension accordingly.

To generate the feature map extracted from each downsampling with the new feature map obtained from the upsampling in channel dimension, the entire framework employs a skip-connection module to combine the features together by adding a stitching layer. Through integrating the underlying features with the higher-level characters, RTnet can recover spatial information while restoring high image resolution simultaneously. Ultimately, the feature information is delivered throughout Linear Projection to obtain the output information including the grasping point, angle, and grasping success rate.

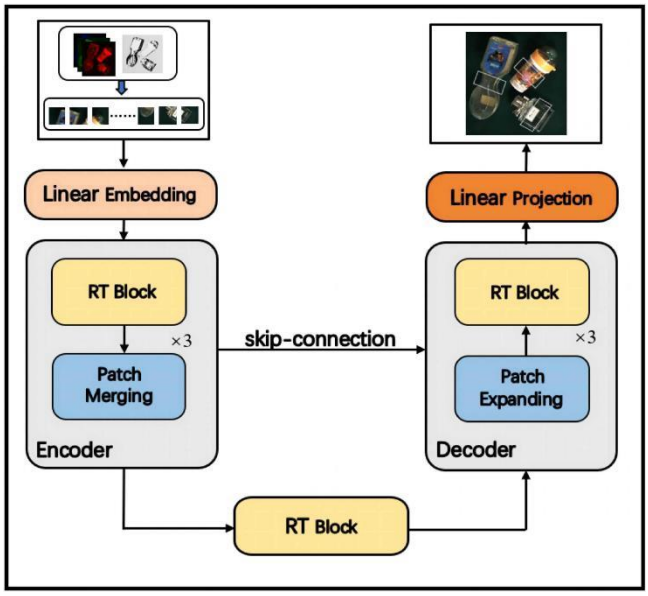


FIGURE 1. **RTnet network architecture.**

C. RT BLOCK

RT block is a crucial component of RTnet, enabling highly robust and accurate grasp prediction. The attention mechanism typically involves complex computation for each patch, leading to significant increases in computational complexity when dealing with numerous patches or large size image. To

enhance the robustness and computational efficiency of the model, a RT block is developed to leverage the powerful shift window modeling and self-normalization capability. Typically, RT block uses a global attention mechanism to facilitate information interaction across all regions. Its working mechanism which optimizes both attention computation and MLP processing is based on the Swin-transformer. This process is formulated by (3). Fig. 2 is the flowchart of RT block, Robust Multilayer Perceptron (R-MLP), and Robust Linear Attention (RLAttention).

$$\begin{aligned} \hat{z}^{l} &= \mathrm{W-MRLA}(\mathrm{LN}(\hat{z}^{l-1})) + z^{l-1} \\ z^{l} &= \mathrm{R-MLP}(\hat{z}^{l}) + \hat{z}^{l} \\ \hat{z}^{l+1} &= \mathrm{SW-MRLA}(\mathrm{LN}(z^{l})) + z^{l} \\ z^{l+1} &= \mathrm{R-MLP}(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{3}$$

where W-MRLA and SW-MRLA denote robust multi-headed attention mechanism modules based on the window and shift window partitioning, respectively. R-MLP denotes the robust multilayer perceptron module. z^{l-1} denotes the input, $\hat{z}^{l}$ and $\hat{z}^{l+1}$ denote the output feature variables of the W-MRLA and SW-MRLA modules, respectively. The z^{l} denotes the output of the R-MLP module.

R-MLP module is one of the key components in established model. It enhances the robustness of the network through the utilization of self-normalizing functions, for example, Selu and Layer Normalization. The Relu activation function is generally utilized in the standard transformer block, however, the negative gradient of Relu causes the feature information masking and ultimately lead to the suboptimal network performance when tackling complex tasks. With the implementation of self-normalization (SN) in the RTnet framework, the output of each layer can converge to zero mean and unit variance during training. This reduces turbulence in the training output and makes the network less susceptible to disorder while highly robust against noise disturbances [21]. Therefore, the RTnet framework incorporates a self-normalization capability to enhance network robustness and generality. The Self-normalization is realized by the Selu activation function defined in (4), which prevent gradient explosion and disappearance by stabilizing variance. As a result, all parameters including weights, biases and activation values have a mean value of 0 and standard deviation of 1.

$$\mathrm{Selu}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^{x} - \alpha & \text{if } x \le 0 \end{cases} \tag{4}$$

where $\alpha \approx 1.67326324235$ and $\lambda = 1.050700987$ are determined by numerical tests.

Furthermore, weight normalization and alpha dropout are crucial factors that impact self-normalization. The Lecun Norm is employed as weight normalization technique to maintain zero mean and unit variance of the weights. Alpha Dropout is utilized to maintain the self-normalizing property by randomly setting certain elements to zero based on Bernoulli distribution, which reduces output variance. During each forward call, the remaining elements are randomly scaled and shifted to preserve the same mean and variance as that of the input. Thus, Alpha Dropout-assisted Selu enables the RTnet with self-normalization capabilities. To further improve the robustness of the network, R-MLP in Fig. 2 is proposed by eliminating the Layer Norm operation of the MLP and introducing the self-normalization operation in the MLP of each RT block. The R-MLP is defined by (5).

$$\hat{\hat{z}} = \mathrm{AD}(\mathrm{Selu}(\mathrm{Linear}(\mathrm{LNorm}(\hat{z})))) \tag{5}$$

where LNorm is an abbreviation for Lecun Norm, indicating the weights normalization. Linear means the fully connected layer and means the Alpha Dropout operation. The input $\hat{z}$ is passed through the first four modules to generate the output $\hat{\hat{z}}$.

RLAttention module is designed to enhance the computational efficiency of the attention mechanism. In contrast to the traditional softmax attention, RLAttention combine L_2 normalization with Relu activation function to reduce computational complexity while improving prediction accuracy effectively. Due to the unique Self-Attention mechanism, the Transformer module that focusing on the global attention mechanism has garnered significant attention in various fields such as natural language processing and computer vision. To enhance the fitting capability of developed framework, input features are transformed using linear matrices to generate three equal-size vectors including the query, key and value. The attention is then calculated by (6).

$$\mathrm{Attention}(Q,K,V) = \mathrm{SoftMax}\left(\frac{QK^{T}}{\sqrt{d}}\right)V \tag{6}$$

where $Q,K,V \in R^{n\times d}$, n is the number of patches and d refers to the dimension of patches.

The vector dimension in the standard transformer attention mechanism is defined as $n \times d$, and the complexity of the calculation step is estimated as $O\left(n^{2}d\right)$. However, due to its quadratic complexity, this attention calculation requires expensive computation resources, making it challenging to apply in real-world scenarios beyond laboratory environment. In this study, a novel linear attention called RLAttention is proposed for computing attention in the RT Block design, as presented in (7-8).

$$\hat{Q} = \frac{Q}{\| Q \|_{2}^{d}}, \quad \hat{K} = \frac{K}{\| K \|_{2}^{d}} \tag{7}$$

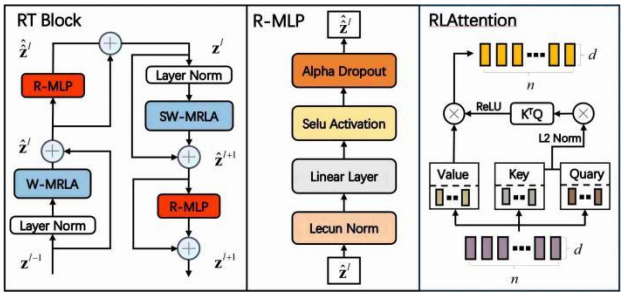


FIGURE 2. RT BLOCK.

$$\mathrm{RLAttention}(Q,K,V)=V(\mathrm{ReLu}\left(\frac{\hat{K}^T\hat{Q}}{\sqrt{d}}\right)) \quad (8)$$

where $\hat{Q}$ is the L_2 normalization of Q along the direction of dimension d, and $\hat{K}$ is the same.

The RLAttention applies L_2 normalization to the Q and K vectors along the patch dimensionality, constraining attention output within a fixed range and demonstrating superior performance compared to standard attention calculation. The rationale behind this modification is to use Relu activation to ensure the relative independence of each mask window in attention calculation through the non-negativity of L_2 normalization, address the redundant and complex SoftMax of the standard attention calculation, achieving a lightweight effect. Additionally, RLAttention reduces the computation order of Q,K,V in the attention mechanism from $O\left(n^2d\right)$ to $O\left(nd^2\right)$, resulting in a significantly reduced computational complexity for RTnet. To prevent overfitting, the sets of W_Q, W_K, Wv are applied to the Multi-Heads RLAttention calculations (MRLA).

D. LOSS FUNCTION

The developed RTnet for robotic grasping prediction achieves the conversion of flat object grasp detection to pixel-level identification. Additionally, a unique orientation angle is determined by $\Phi=(1/2)\arctan(\sin\theta/\cos\theta)$, which is used to represent the unique grasp value of each pixel (Q,w,ϕ). To improve the robustness of RTnet processing discrete points, the loss regression function smooth_{L_1} is designed to combine the L_1 and L_2 loss function synergizing the advantages including the smoothness of function for the small value of x and the stability of function for the large value of x.

For an object represented $X=(x_1,x_2,x_3,\ldots,x_n)$ and its corresponding grasp token $L=(l_1,l_2,l_3,\ldots,l_n)$ in the dataset, Equations (9-11) define four predicted values of the RTnet output for grasp prediction, as well as the difference between predicted and true values through a loss function.

$$G=(q_n,w_n,\sin\theta_n,\cos\theta_n) \quad (9)$$

$$\mathrm{smooth}_{L_1}=\begin{cases}0.5x^2 & \text{if } |x|<1\\ |x|-0.5 & \text{otherwise}\end{cases} \quad (10)$$

$$\mathrm{Loss}(G,L)=\sum_i\sum_{m\in\{q,w,\sin\theta,\cos\theta\}}\mathrm{smooth}_{L_1}\left(G_i^m-L_i^m\right) \quad (11)$$

E. REPRESENTATION OF PACKAGING GRASP SCENARIOS

Robotic grasping prediction in industrial settings requires high robustness to achieve precise grasping detection, given the presence of numerous complex distractions in the real- world environment. However, existing grasping datasets related to industrial packaging environments are generated under the assumption of neat and smooth scenarios, resulting in a significant discrepancy between RGB information of the target object and surrounding environmental data, as shown in Fig. 3. Moreover, the grasping data obtained from the factory scenarios is subject to various interference and noise. Therefore, a remarkable offset between these two representations is indispensable, causing the disagreement between the results based on the current dataset and those derived from real factory scenarios.

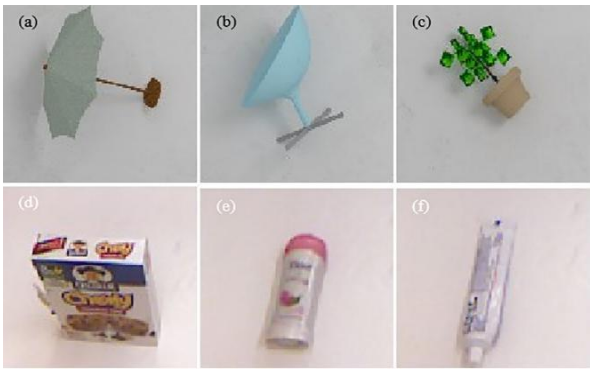


FIGURE 3. Example of an idealized dataset. (a), (b) and (c) from Jacquard dataset; (d), (e), and (f) from Cornell dataset.

As subtle interference or noise in the real-world scenarios affect the accuracy of predictions, high-quality images of inputs are crucial for neural networks. However, low quality and noisy data are often prevalent in such environment. Therefore, training a network on datasets with complex backgrounds will demonstrate higher robustness and better performance. With the incorporation of the dataset through style transfer into the original Cornell dataset, texture noise is successfully integrated, and new captured data information are generated to assess the performance of RTnet for robotic grasping in complex packaging scenarios.

To accurately depict real packaging factory scenarios, the interference that occur in industrial capture scenarios are developed by the integration of different packaging products in the Cornell database with white background shown in (1) of Fig. 4(a-d) and four kinds of realistic factory environments, which include low-resolution grasping environment ((2) in Fig. 4(a)), colored grasping background ((2) in Fig. 4(b)), wooden grasping background ((2) in Fig. 4(c)) and blocky oiled background ((2) in Fig. 4(d)). The primary principle for the above combination is to ensure the representation separation of image style and content, thereby maximizing preservation while transforming styles during style migration.

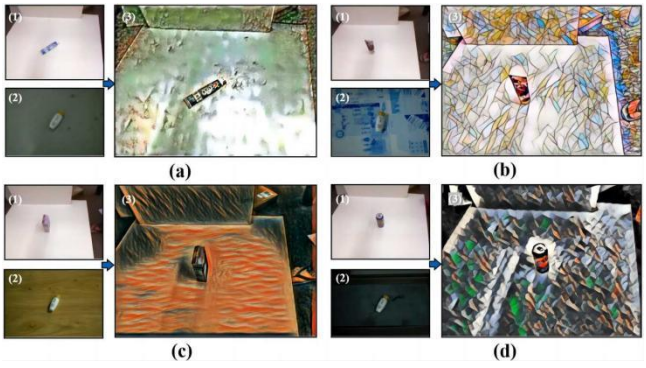


FIGURE 4. Data argumentation by style transfer. In (a-d), Packaging products in Cornell database with white background (1); Four realistic factory environments (2); Four transferred styles (3)

Thus, four transferred styles in (3) of Fig. 4(a-d) are generated and called RealCornell (see Section 3.4). It is noted that data argumentation by style transfer enhances the generality and style transfer ability of the developed RTnet for the realization of various real grasping scenarios.

III. EXPERIMENT AND RESULTS

During the experimental phase, the accuracy and effectiveness of the RTnet model is evaluated through experimental validation on both the Cornell grasping dataset and the developed RealCornell dataset, as well as real grasping experiments conducted on physical robots.

A. DATASETS

The Cornell dataset consists of 885 RGB-D images of 240 objects, while the RealCornell dataset contains 4×885 RGB-D images of 240 objects generated by style transfer in Section 2.5. For each dataset, a random selection of 90% data is used for training the model and the remaining 10% is served as the validation set.

B. EVALUATION INDICATORS AND IMPLEMENTATION DETAIL

To enhance the realism of the grasp, the rectangle metric is accurate provided that certain conditions are satisfied.

- The difference between the predicted grasping angle and the actual grasping scenario is less than 30 degrees.
- The overlap area between the predicted rectangle and the correct rectangular grasp is greater than 0.25 and can be calculated according to (12). G_P is the grasping data of prediction and G_R is the correct grasping value.

$$Jac(G_P, G_R) = \frac{|G_P \cap G_R|}{|G_P \cup G_R|} \tag{12}$$

The model is constructed utilizing Pytorch on Ubuntu 20.04 with an NVIDIA 3060 GPU, employing Adam as the optimizer and a default learning rate of 0.001 that is dynamically adjusted during training.

C. CORNELL DATASET EXPERIMENTS

In the experiments conducted on the Cornell dataset, two methods of partitioning the dataset are employed: image wise (Iw) and object wise (Ow). Table I presents the accuracy of grasping results. The end-to-end grasping prediction network [8] achieved accuracies of 73% and 69% using only Depth as the input. Subsequently, RGB information is incorporated into the prediction model [22]. When the input consists of multi-source information including RGB and Depth, the accuracy can reach nearly 98%, as demonstrated in [15], [23].

The proposed RTnet achieves a superior accuracy of 98.31% on the Cornell dataset when both RGB and Depth data are employed. As compared with the results by other researchers in Table I, RTnet achieves a better accuracy of 96.61% when only RGB is utilized as the input. Overall, RTnet demonstrates a high level of performance in completing the capture task.

TABLE I
CORNELL DATASET TESTING RESULTS

Model	Input	IW accuracy	OW accuracy
GGCNN[8]	D	73.0%	69.0%
GraspNet[22]	RGB-D	90.2%	90.6%
GR-ConvNet[23]	D	93.2%	94.3%
	RGB	96.6%	95.5%
	RGB-D	97.7%	96.6%
TF-Grasp[15]	D	**95.2%**	**94.9%**
	RGB	96.6%	95.0%
	RGB-D	97.99%	96.7%
RRnet	D	94.91%	94.87%
	RGB	**96.61%**	**95.92%**
	RGB-D	**98.31%**	**97.65%**

D. REALCORNELL DATASET EXPERIMENTS

To assess the robustness of the models, GRnet and TF-Grasp are popular tools used to evaluate the accuracy of robot grasping prediction. In this study, only RGB-D as the input of the model is examined on the RealCornell dataset. It is noted in Table II that GRnet and TF-Grasp have the same level of accuracy, which is lower than 92%. The accuracy by RTnet's has remarkably increased by over 2% than the results by GRnet and TF-Grasp, reaching a value of 93.878%.

In order to visualize the accuracy of grasping prediction, it is necessary to generate grasping heat maps. In Fig. 5, typical packaging products from the RealCornell dataset are selected, including regular cylindrical can packaging and irregular packaging, e.g., toothpaste packaging, cosmetics packaging and wine bottle. Through heat map analysis on each pixel of the grasping points, the quality, deflection angle and grasping width are obtained. The blue rectangular boxes labeled in the first row of Fig. 5 illustrate the grasping locations predicted by the RTnet in the four scenarios. The closer to the object the graspable area (red zone in the second row in Fig. 5) is, the higher the grasping accuracy is. The Angle and Width diagrams show the best grasping angle and grasping width predicted by RTnet. Therefore, RTnet has the ability to effectively grasp the objects by identifying the graspable characteristics and its grasping accuracy is evaluated by the quality, angle, and width heat maps. Summarily, the results demonstrate that the proposed grasping network has superior grasp prediction performance in terms of accurate and robust feature extractions under the complex working scenarios.

The robustness of RTnet is further validated on the RealCornell dataset by experimental tests with the Jacc indexes in range of 0.3 to 0.5 in terms of object grasping accuracy. Fig. 6 demonstrates that RTnet outperforms the other two models in terms of the increased accuracy in object grasping. It is noted that when the number of the sample batch reaches 100, the accuracy of grasping is slightly influenced if the batch number is further increased to 200. Therefore, it is recommended to set the batch number to 200 in the experiments from an efficient computation point of view. In summary, RTnet can achieve excellent accuracy and complete the task of grasping prediction.

TABLE II
REALCORNELL DATASET TESTING RESULTS

Model	GR-ConvNet[23]	TF-Grasp[15]	RTnet
Accuracy	91.304%	91.836%	93.878%

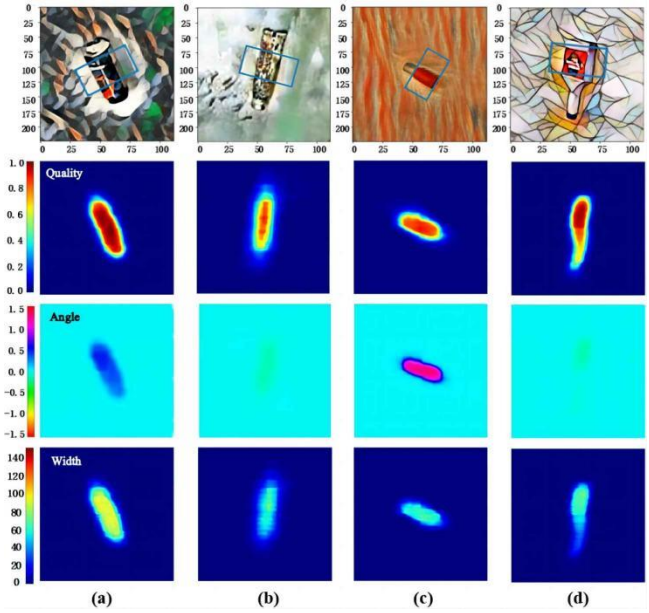


FIGURE 5. Visualization of the results of the RealCornell dataset. Toothpaste packaging (a); Cylindrical cans (b); Cosmetics packaging (c); Wine bottle (d).

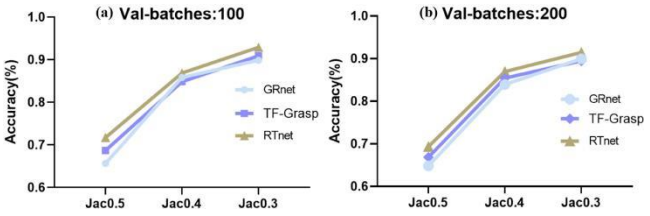


FIGURE 6. Visualization of the results of the RealCornell dataset. Accuracy when the batch is 100 (a); Accuracy when the batch is 200 (b).

E. ABLATION EXPERIMENT

To explore the impact of the developed core modules, e.g., RLAttention and R-MLP, on the overall performance of the RTnet model, ablation experiments are conducted on the Cornell dataset and RealCornell dataset. A model equipped only with Swin transformer block is selected as the benchmark model for the ablation experiment. On this basis, RLAttention and R-MLP are sequentially added for the experiment. The experimental results are assessed by the accuracy index.

Results of Ablation Experiment are presented in Table III. As compared to the baseline model using the Cornell dataset, the model implemented by RLAttention only has a slight improvement in accuracy, which is from 79.78% to 80.48%. Moreover, with the addition of R-MLP into the model, an accuracy of nearly 2% is further increased. The similar conclusion is drawn by the experimental results using RealCornell dataset. It is noted that the RealCornell dataset contains significant noise interference, therefore the implementation of R-MLP has a more significant impact on the model accuracy by the increase of 3.35% from 69.43% to 72.78%.

TABLE III
RESULTS OF ABLATION EXPERIMENT ACCURACY

Base	RLAttention	R-MLP	Cornell	RealCorell
√	-	-	79.78%	68.86%
√	√	-	80.48%	69.43%
√	√	√	82.23%	72.78%

Note: "√" indicates the adopted module in the model , "-" means there is no module adopted in the model .

In summary, the implementations of RLAttention and R-MLP for enhanced self-normalized attributes improves the robustness of the model, enable the suppression of irrelevant features, prioritize target features, and make more accurate predictions of grasping posture in complex packaging factory grasping environment.

F. REAL ROBOT GRASPING EXPERIMENTS

Some advanced robotic grasp models have been developed in the grasping experiments by many researchers and provided useful guidance to accurate prediction for the solution to industrial problems. Nevertheless, these models could only work well under the flat color background, leading to the poor model generalization. Therefore, applications of the grasping models to solve the problems arising from the real production process are severely limited.

To address the above issue, a robot grasping platform based on visual perception is established in this study. A depth camera with a measurement accuracy of 0.1mm and the resolution of 1920×1200 pixels, is installed in the KUKA KR10 robot, enabling the capability of solving the real grasping problems. The view field of the depth specified by the camera provides the measurement of 800mm × 450mm, which aligns with the robot's achievable range. The camera is also equipped with a blue LED light source and positioned at a distance of 1 meter from the desktop in Fig. 7. The repetitive positioning accuracy of the robot is 0.05 mm. The settings of these parameters effectively reflect the grasping environment of the factory.

To validate the model, the prediction results of RTnet are compared with other models in real scenarios. The results are shown in Fig 8(a), where the prediction results for the eyeglass case by different models are provided. Results demonstrate that RTnet accurately predicts the grasping position aligned with the center of gravity of the target object, whereas TF-Grasp exhibits a biased prediction. On the other hand, GRnet erroneously identifies the background as an object, leading to inaccurate predictions. In Fig. 8(b), it is noted that RTnet exhibits the superior capability in accurately recognizing the complete appearance contour of the toothpick box, while TF-Grasp manifests a significant error and GRnet only captures a

partial object contour. Consequently, it can be inferred that as compared to other models, RTnet offers more precise predictions of grasping positions and possesses an enhanced accuracy in discerning object contours.

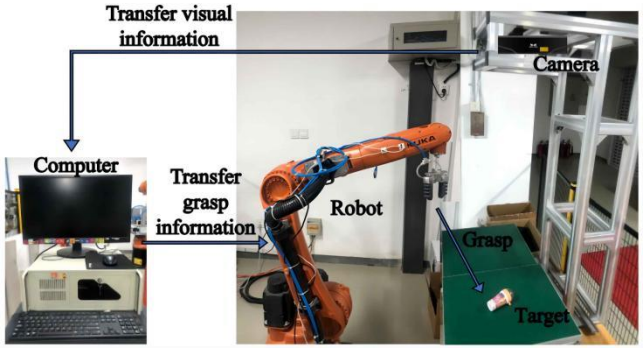


FIGURE 7. Execution process of grasping robot system.

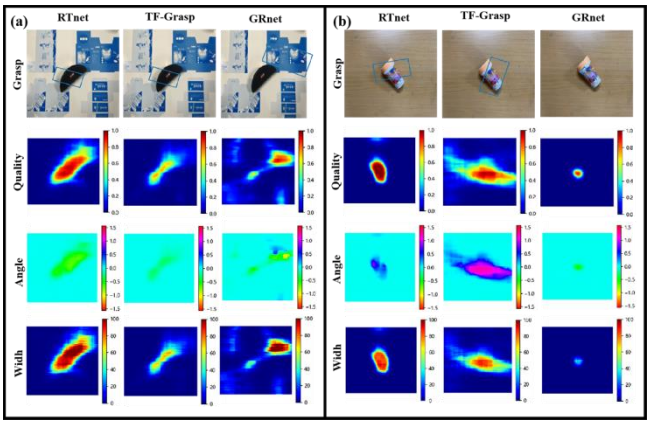
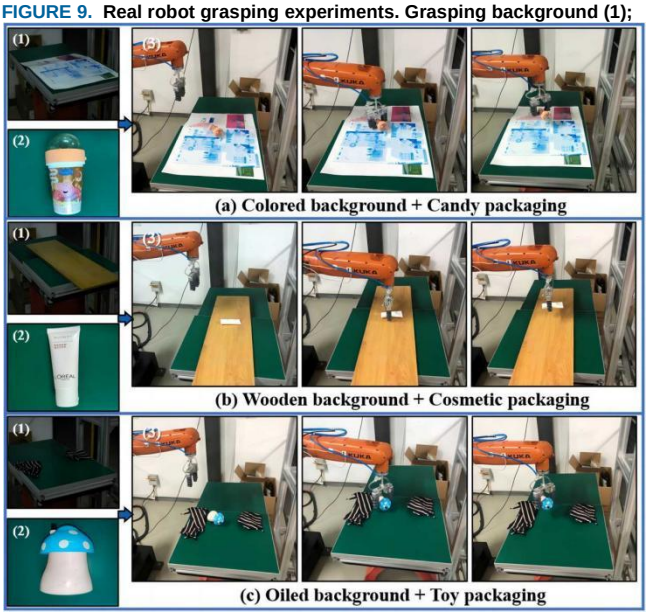


FIGURE 8. Prediction results in real grasp scenarios. Prediction results of eyeglass case(a); Prediction results of toothpick box(b).

The depth camera captures an image of an object in each grasp and transmits the visual information to the computer, providing computers the grasping messages through model processing. The extracted information is then fed to robots end effector, which approaches the optimal target grasping pose by motion planning techniques till the completion of the grasp operation, as depicted in Fig. 9.

To realistically mimic the factory environment, three representative interfaces in (2) of Fig. 4(b-d) (The colored grasping background, The wooden grasping background and the blocky oiled grasping background) are selected for conducting real robot grasping experiments, shown in (1) of Fig. 9(a-c). Three types of irregular shaped objects are used in grasping experiments, including the candy packaging, the cosmetic packaging and the toy packaging, as shown in (2) of Fig. 9(a-c). The entire process of robot grasping experiments are illustrated in (3) of Fig. 9(a-c). In the grasping experiments, each object is placed in the one of three interfaces and then grasped 9 times in one scenario. Therefore, a total number of 81 grasps are determined.

FIGURE 9. Real robot grasping experiments. Grasping background (1); The target object (2); Robot grasping process (3).



The statistical analysis of all data is shown in Table IV, where indicates a successful rate of the grasping experiments in the different scenarios. In terms of the grasp accuracy, the highest success rate of 92.59% in the wooden background and the lowest success rate of 70.37% in the colored working environment are observed. The reason for this lies in that the more remarkable difference between the RGB information in the wooden grasping scenario and that in the colored working environment exists, the more successful grasping rates are distinguished. Thus, the higher success rate with the less the RGB information in the wooden background is achieved. Considering the grasped object, a highest success rate of 88.89% is observed for the candy packaging product due to its regular shape and a successful rate of 77.78% for the toothpick packaging product is also acceptable. In general, the proposed RTnet has ultimately the ability to achieve a high average accuracy of 82.76% and demonstrates a testimony to the model's superior capability of grasping. In summary, the designed experiments in this paper realize the real factory robot grasping and provide an environment setting that reflects the complex scenario of a factory, securing higher levels of model robustness and generalization.

TABLE IV
TEST RESULTS FOR THE REAL ROBOT GRASP

Grasping Back ground	Candy packaging	Toy packaging	Cosmetic packaging	Total result
Colored interface	7/9	6/9	6/9	19/27
Wooden interface	9/9	8/9	8/9	25/27
Oiled interface	8/9	7/9	8/9	23/27
Total	88.89%	77.78%	81.48%	82.76%

IV. CONCLUSION

In this study, a novel robot grasping prediction model based on a linear global attention mechanism (RTnet) is proposed. RTnet linearly optimizes the quadratic complexity of traditional attention mechanisms. To accomplish the task of capturing in complex scenes, RTnet adopts a self- normalized combination of Lecun Norm, Selu and Alpha. Dropout to enhance the model filtering and adaptability to noise interference while improving the robustness of feature learning. An amplified dataset (RealCornell) is generated through style transfer to accurately mimic the packaging factory capture scenes. Experimental tests are evaluated by the Cornell dataset, the RealCornell dataset and real grasping scenarios. Compare with the existing results, the proposed RTnet achieves a better accuracy of grasping predictions (98.31% and 93.88%) on the Cornell and the RealCornell datasets, respectively. In this research, the ablation experiments are also carried out to demonstrate the R-MLP remarkable contributions to the enhancement of the model generalization and the effectiveness in handling complex capture scenarios. In conclusion, RTnet achieves an acceptable accuracy in real robot grasp experiments, demonstrating its generalization ability under variety of packaging scenarios.

Although RTnet achieves precise and robust grasping in complex packaging factory scenarios, practical deployment and widespread applications of the developed model still need to be explored. Firstly, the capture priority can be adjusted to accommodate more complex environments, such as situations involving object overlaps. In scenarios where multiple targets are overlapped, the robot can determine the order of grasping by assessing factors like the difficulty level, importance, and urgency associated with each object. Secondly, in practical applications of robotic grasping, grasping scenarios such as the object inclined positioning is frequently encountered, the robot's capability to execute grasps on slopes is necessary. Therefore, 6D pose estimation technique should be employed in the robotic slope grasping tasks in the future to acquire the precise object positioning and orientation information. Furthermore, to enhance the robustness and intelligence of grasping strategies, future research studies will include the effective integration of information from diverse sensors in the industrial environment, such as amalgamating visual data with force or touch data. Moreover, the RTnet framework has the potential to be extended for the applications beyond packaging factory environments, including medical surgical assistance, disaster relief, home service robots, and different sectors. These domains present distinct requirements and challenges related to grasping accuracy and robustness, which prove to be worth of further investigations.

REFERENCES

[1] M. Farag, A. N. A. Ghafar and M. H. Alsibai, "Real-Time Robotic Grasping and Localization Using Deep Learning-Based Object Detection Technique," in *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, Selangor, Malaysia, 2019, pp. 139-144.

[2] H. Zhang et al., "A Practical Robotic Grasping Method by Using 6D Pose Estimation With Protective Correction," in *IEEE Transactions on Industrial Electronics*, vol. 69, no. 4, pp. 3876-3886, April. 2022.

[3] Yun. Jiang, S. Moseson and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in *2011 IEEE International Conference on Robotics and Automation*, Shanghai, China, 2011, pp. 3304-3311.

[4] L. He, H. Zhang, "Doubly Stochastic Distance Clustering," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6721-6732, Nov. 2023.

[5] Y. Zheng, X. Xu, J. Zhou and J. Lu, "PointRas Uncertainty-Aware Multi-Resolution Learning for Point Cloud Segmentation," in *IEEE Transactions on Image Processing*, vol. 31, pp. 6002-6016, 2022.

[6] Z. Li and S. Liu, "Grasping Detection Based on YOLOv3 Algorithm," in *2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS)*, Suzhou, China, 2021, pp. 824-829.

[7] R. Cruz Villagomez and J. Ordoez, "Robot Grasping Based on RGB Object and Grasp Detection Using Deep Learning," *2022 8th International Conference on Mechatronics and Robotics Engineering (ICMRE)*, Munich, Germany, 2022, pp. 84-90.

[8] D. Morrison, P. Corke and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International jornal of robotics research*, 39(2-3): 183-201, 2020.

[9] H. Cheng, Y. Wang, "A Robot Grasping System With Single-Stage Anchor-Free Deep Grasp Detector," in *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-12, 2022, Art no. 5009712.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł Kaiser, I Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 30, 2017.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal*, QC, Canada, 2021, pp. 9992-10002.

[13] L. He, H. Zhang, "Large-scale Graph Sinkhorn Distance Approximation for Resource-constrained Devices," in *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2960-2969, Feb. 2024.

[14] L. Zou, Z. Huang, N. Gu and G. Wang, "6D-ViT: Category-Level 6D Object Pose Estimation via Transformer-Based Instance Representation Learning," in *IEEE Transactions on Image Processing*, vol. 31, pp. 6907-6921, 2022.

[15] S. Wang, Z. Zhou and Z. Kan, "When Transformer Meets Robotic Grasping: Exploits Context for Efficient Grasp Detection," in *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8170-8177, July 2022.

[16] A. Umam, C. K. Yang, J. H. Chuang and Y. Y. Lin, "Unsupervised Point Cloud Co-Part Segmentation via Co-Attended Superpoint Generation and Aggregation," in *IEEE Transactions on Multimedia*, vol. 26, pp. 7775-7786, 2024.

[17] J. Jiang, Z. He, X. Zhao, S. Zhang, C. Wu and Y. Wang, "REG-Net: Improving 6DoF Object Pose Estimation With 2D Keypoint Long-Short-Range-Aware Registration," in *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 328-338, Jan. 2023.

[18] S. Ge, B. Hou, W. Zhu, Y. Zhu, S. Lu and Y. Zheng, "Pixel-Level Collision-Free Grasp Prediction Network for Medical Test Tube Sorting on Cluttered Trays," in *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 7897-7904, Dec. 2023.

[19] W. Wei, Y. Lou, F. Li, G. Xu, "GPR: Grasp Pose Refinement Network for Cluttered Scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China, 2021, pp. 4295-4302.

[20] M. Niu, Z. Lu, L. Chen, J. Yang and C. Yang, "VERGNet: Visual Enhancement Guided Robotic Grasp Detection Under Low-Light

Condition," in *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 8541-8548, Dec. 2023.
[21] G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, "Selfnormalizing neural networks." *Advances in neural information processing systems*, 30, 2017.
[22] U. Asif, J. Tang and S. Harrer, "GraspNet: An Efficient Convolutional Neural Network for Real-time Grasp Detection for Low-powered Devices." in *IJCAI*, 2018, 7: 4875-4882.
[23] S. Kumra, S. Joshi and F. Sahin, "Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, 2020, pp. 9626-9633.

GUIRONG DONG received the Ph.D. degree from Xi'an University of science and technology, China, in 2018. She is currently an Associate Professor in the School of Printing, Packaging and Digital Media, Xi'an University of Technology, China. Her primary research interests include machine vision, machine learning, and robot technology.

FUQIANG ZHANG received the B.S. degree in Printing Engineering from Henan University of Technology, Luoyang, China, in 2021, the M.S. degree from the Xi'an University of Technology, Xian, China, in 2024. His primary research interests include machine vision, robotic grasping.

XIN LI received the B.S. degree in Printing Engineering from Henan University of Engineering, Zhengzhou, China, in 2023. She is currently pursuing a M.S. degree at Xi'an University of Technology. Her primary research interests include machine vision, robotic grasping, and three-dimensional packaging.

Zonghui Yang received the B.S. degree in Printing Engineering from Xi'an University of Technology, China, in 2022. He is currently pursuing a M.S. degree at Xi'an University of Technology. His primary research interests include machine vision, object detection, and robotic grasping.

DIANZI LIU was received the B.Eng. and M.Sc. degrees from Beihang University, China, in 1999 and 2002, respectively, and the Ph.D. degree from the University of Leeds, U.K., in 2010, sponsored by the Overseas Research Scholarship (ORS) Scheme.

He is currently an Associate Professor with the School of Engineering, University of East Anglia (UEA), U.K. He has published more than 40 technical articles in the past three years. His current research interests include machine learning, composite structures, optimization driven designs, and computational mechanics. He was awarded the runner-up prize in the ISSMO-Springer Prize Competition, in 2009.