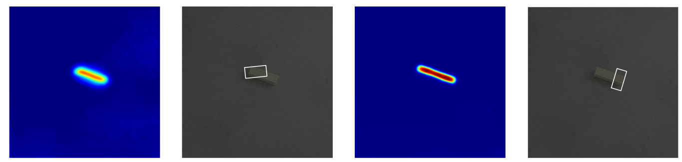
VERGNet: Visual Enhancement Guided Robotic Grasp Detection under Low-light Condition

Mingdi Niu[1], Zhenyu Lu[2], *Member, IEEE,* Lu Chen[1], *Member, IEEE,* Jing Yang[3], Chenguang Yang[2], *Senior Member, IEEE*

***Abstract*—Although existing grasp detection methods have achieved encouraging performance under well-light conditions, repetitive experiments have found that the detection performance would deteriorate drastically under low-light conditions. Although supplementary information can be provided by additional sensors, such as depth camera, the sparse and weak visual features still hinder the improvement of detection accuracy. In order to address these, we propose a visual enhancement guided grasp detection model (VERGNet) to improve the robustness of robotic grasping in low-light conditions. Firstly, a simultaneous grasp detection and low-light feature enhancement framework is designed, which integrates residual blocks with coordinate attention to re-optimize grasping features. Then, the unsupervised low-light feature enhancement strategy is adopted to reduce the dependence on paired data as well as improve the algorithmic robustness to low-light conditions. Extensive experiments are finally conducted on two newly-constructed low-light grasp datasets and the proposed method achieves 98.9% and 91.2% detection accuracy respectively, which are superior to comparative methods. Besides, the effectiveness in our method has also been validated in real-world low-light imaging scenarios.**

***Index Terms*—Robotic grasping, grasp detection, image enhancement, data-driven model**



(a) GR-ConvNet (b) VERGNet

Fig. 1. Comparison of detection results between GR-ConvNet and VERGNet under low-light conditions. Left: grasp quality maps, right: grasp detection results.

I. Introduction

As information technology and artificial intelligence develop, the role of robots is becoming increasingly important in the fields of industrial manufacturing [1], household services [2], agricultural harvesting and space exploration [3]. Robotic grasping, as the most basic skill of robots, is one of the most challenging techniques in robot operation. The robot grasping process includes object localisation, grasp detection, path planning and grasp execution, where grasp detection aims to find the graspable part on object and is one of the key steps. However, due to the unknown target morphology, complex environmental interference, and mutual occlusion of multiple objects, robotic grasp detection in real-world scenarios still faces many serious challenges.

Currently, most existing grasp detection works [4], [5] are usually conducted under well-light conditions, where the target structure is distinguishable and the detail contrast is sharp. When it comes to low-light conditions, the target's visual features are weak and tend to be easily confused with background, making the effective extraction of grasping-specific features difficult.

In order to increase the adaptability of grasp detection under low-light condition, the existing methods can be roughly divided into two aspects: 1) using supplementary data from external sensors, e.g., infrared camera, depth camera and laser radar. However, this strategy generally requires higher energy consumption and more importantly, the provided imaging data are deficient in revealing abundant texture information of object, restricting the further improvement of grasp detection performance. 2) using more powerfully deep neural networks for image enhancement or domain adaptation methods for knowledge transfer. However, some works [6] have shown that simply boosting the visual quality of image does not always benefit other vision tasks, and the performance of transferring learned features between different domains is still restricted [7].

In order to solve the problem of grasping detection under low-light conditions, we propose a novel visual enhancement guided robotic grasping detection network under low-light conditions, which contains residual modules based on coordinate-attention and an unsupervised visual grasping feature enhancement branch. The coordinate attention cap-

Manuscript received: September 3, 2023; Accepted: October 20, 2023.

This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments.

This work was jointly supported by the National Natural Science Foundation of China (Nos. 62003200, 62373233), the H2020 Marie Skodowska-Curie Actions Individual Fellowship (No. 101030691), the Fundamental Research Program of Shanxi Province (No. 202203021222010), the Science and Technology Major Project of Shanxi Province (No. 202201020101006), and the Central Guidance for Local Scientific and Technological Development Funds (No. YDZJSX20231B001). (*Mingdi Niu and Zhenyu Lu contributed equally to this work.) (Corresponding author: Lu Chen)*

[1]Mingdi Niu, Lu Chen are with the Institute of Big Data Science and Industry and the School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China. niumingdi@163.com, chenlu@sxu.edu.cn

[2]Zhenyu Lu, Chenguang Yang are with the Faculty of Environment and Technology and Bristol Robotics Lab at the University of the West of England, Bristol, BS16 1QY, UK. luzhenyurobot@gmail.com, cyang@ieee.org

[3]Jing Yang is with the School of Automation and Software Engineering, Shanxi University, Taiyuan 030031, China. yangjing199002@sxu.edu.cn

Digital Object Identifier (DOI): see top of this page.

tures cross-channel information, direction-aware and position-sensitive information, resulting in the more accurate extraction of target's position information, while the skip connections are used to fuse multilevel features. In addition, the unsupervised visual grasping feature enhancement method not only reduces the model's dependence on paired data, but also constrains the extraction of grasping features, enabling the model to learn more generic features. Extensive experiments show that compared with direct extraction of grasping features on RGB-D input, our method is able to achieve better performance in terms of detection accuracy by introducing the low-light grasp features enhancement sub-task. As illustrated in Fig. 1, the quality map predicted by GR-ConvNet is scattered with low confidence, leading to incorrect detection result. But for our VERGNet, it could generate more concentrated and confident quality map. To summarize, our main contributions are listed below:

1) A simultaneous grasp detection and low-light enhancement framework is proposed to guide the enhancement and detection of grasping-specific features from single visual perspective.
2) The features to be enhanced are learned with semantic level constraints using an unsupervised manner, excluding the requirement of paired "normal-low" light images.
3) Two low-light grasp detection datasets are constructed from Cornell and Jacquard datasets, and the effectiveness of the proposed method is verified both on built datasets and real-world robotic grasping.

The rest of this work is summarized as follows. Section II investigates the related works and Section III describes our proposed method. We conduct the experiments and conclude our work in Section IV and V.

II. Related Work

A. Deep learning based grasp detection approaches

As deep learning has proved its effectiveness in diverse vision tasks, it is also being introduced into the field of grasp detection. In order to avoid the design of complex artificial features, Lenz *et al.* [8] used deep learning for the first time to solve the grasp detection problem. Redmon *et al.* [9] used AlexNet [10] to predict grasping region parameters. However, the accuracies of these methods were still relatively limited. To further improve the accuracy of grasp detection, Chen *et al.* [11] used grasp paths rather than orientated rectangles to represent grasp poses, which allows for a fairer assessment of the graspability of predictive grasps. Kumra *et al.* [12] proposed a pixel-level grasp detection model with the output of three heat maps representing the width, angle and quality of the grasp. Following the new pixel-wise grasp representation, a subset of grasp detection methods [13], [14], [15], [16] were proposed to focus on useful grasping features by adding various attention modules. These methods have largely improved the accuracy and speed of grasp detection.

It observes that the current approaches generally consider the case of sufficient light, and when the imaging condition becomes darker, their grasp detection performances tend to decrease dramatically.

B. Low-light image enhancement

Low-light image enhancement is a significant research direction of computer vision. In the past decades, various methods had been introduced. Specifically, Chen *et al.* [17] designed a Retinex-based low-light image enhancement model (RetinexNet), which could estimate light and reflection simultaneously. In addition, Chen *et al.* also created a brand new dataset (LOL dataset) with synthetic noise obtained by changing the exposure time. Jiang *et al.* [18] proposed a GAN-based low-light image enhancement method with the advantage of eliminating the dependence of the model on paired data. Wang *et al.* [19] presented a new progressive Retinex framework based on the Retinex approach considering decoupling, using mutual enhancement to perceive light and noise in low-light images. Guo *et al.* [20] used a neural network to fit a luminance mapping curve, and then generated an enhanced image based on the curve and the original image, while constraining the optimisation process using some new loss functions during training. Besides, Liu *et al.* [21] proposed a model consisting of a decomposition network and an adjustment network based on Retinex theory, as well as a self-supervised fine-tuning strategy to improve the visual performance.

III. Proposed Method

Instead of the rectangular representation proposed by Jiang *et al.* [22], we use the pixel-based grasping representation proposed in [23], and one grasp p is defined as:

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

where (x, y) denotes the spatial coordinates of grasping point, θ denotes the rotation angle of the grasping rectangle, w is the opening width, and q represents the quality of the grasping pose. Specifically, we denote the pixel-level grasping configuration as $\boldsymbol{P}$, which is defined as follows:

$$\boldsymbol{P} = \{\boldsymbol{W}, \boldsymbol{\Phi}, \boldsymbol{Q}\} \in \mathbb{R}^{H \times W \times 3} \tag{2}$$

where $\boldsymbol{W}$, $\boldsymbol{\Phi}$, and $\boldsymbol{Q}$ denote the three feature maps outputted by the model, and each pixel in these images can be regarded as the width, rotation angle, and grasp quality score of a grasp rectangle candidate.

A. Grasp detection network architecture

We propose a Visual Enhancement guided Robotic Grasp detection Network (VERGNet), aiming at improving the accuracy of grasp detection models under low-light conditions. The fundamental framework is shown in Fig. 2, which includes a feature extraction module, a grasp detection head and a low-light grasp features enhancement head. The inputs to the model are RGB image $I_r \in \mathbb{R}^{H \times W \times 3}$ and depth image $I_d \in \mathbb{R}^{H \times W \times 1}$, while the outputs are three images $\{\boldsymbol{\Phi}, \boldsymbol{W}, \boldsymbol{Q}\} \in \mathbb{R}^{H \times W \times 3}$ respectively.

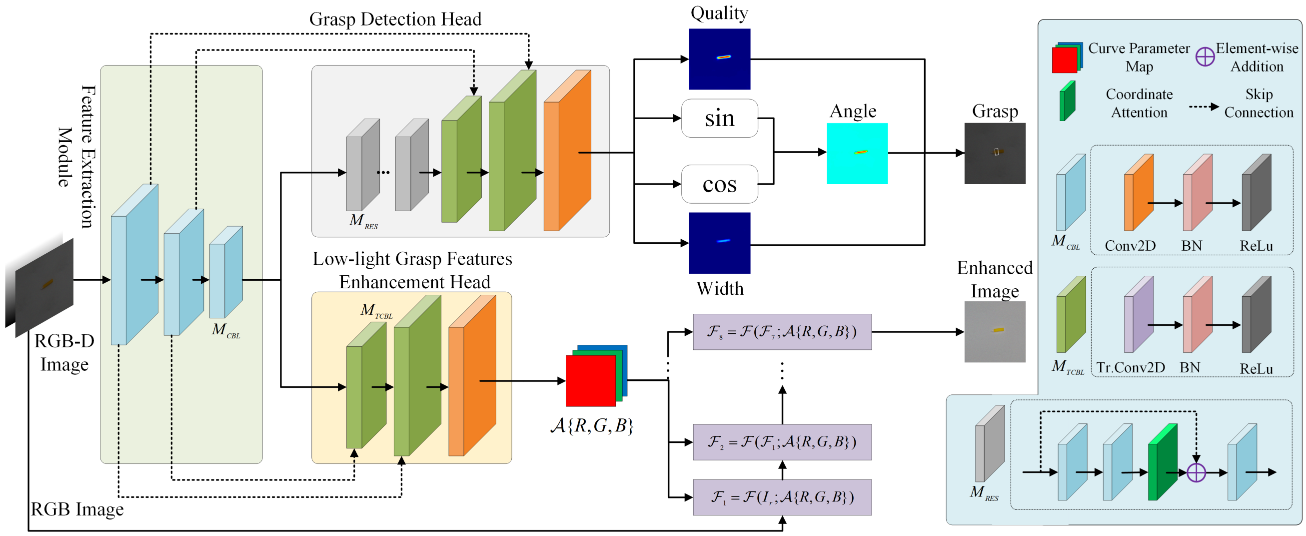


Fig. 2. Framework of the visual enhancement guided robotic proposed grasp detection network.

1) Feature extraction module: Firstly, the RGB image I_r and depth image I_d are spliced into the input model in channel dimension and then the underlying features will be sequentially extracted through three CBL modules M_{CBL} [24]. The CBL modules are defined as follows:

$$M_{CBL} = \{Conv2D, BatchNorm, ReLu\} \tag{3}$$

The process of extracting features is shown below:

$$F_{low} = M_{CBL}\left(M_{CBL}\left(M_{CBL}\left[I_r, I_d\right]\right)\right) \tag{4}$$

where F_{low} denotes the extracted underlying features.

2) Grasp detection head: The underlying features extracted by the encoder are firstly passed into the residual module M_{RES}, and then sequentially into the two transposed convolution modules M_{TCBL}. Finally, the number of channels is changed to 4 by the convolution operation, and the final outputs of the three images are the grasping quality map $\boldsymbol{Q}$, the grasping angle map $\boldsymbol{\Phi}$ and the grasping width map $\boldsymbol{W}$. The specific operations are shown below:

$$\{\boldsymbol{\Phi}, \boldsymbol{W}, \boldsymbol{Q}\} = Conv2D\left(M_{TCBL}^2\left(M_{RES}^5\left(F_{low}\right)\right)\right) \tag{5}$$

Regarding the residual module M_{RES}, it contains three CBL modules and one coordinate attention module M_{CAM} [25], whose fundamental structure is illustrated in Fig. 3. Specifically, M_{CAM} first extracts the spatial information by averaging pooling along W and H directions. Then, feature transformations are deployed to encode the spatial information therein, followed by weighting them across channels in order to achieve the final information fusion. The input features F_{in} firstly enter two CBL modules M_{CBL} sequentially and then flow into the coordinate attention module M_{CAM}. The re-optimized features are summed up with skip connection, followed by one additional M_{CBL} to get the final output features F_{out}. The specific process is shown below:

$$F_{out} = M_{CBL}\left(M_{CAM}\left(M_{CBL}^2\left(F_{in}\right)\right) + F_{in}\right) \tag{6}$$

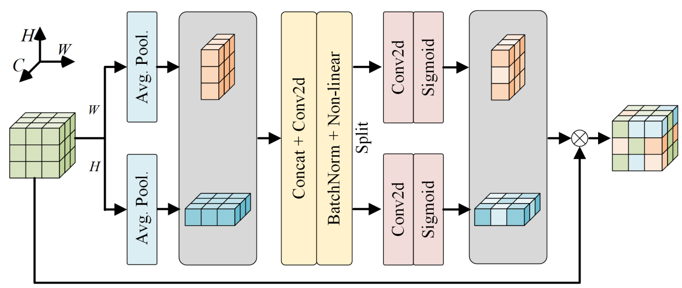


Fig. 3. Fundamental structure of the coordinate attention module M_{CAM}.

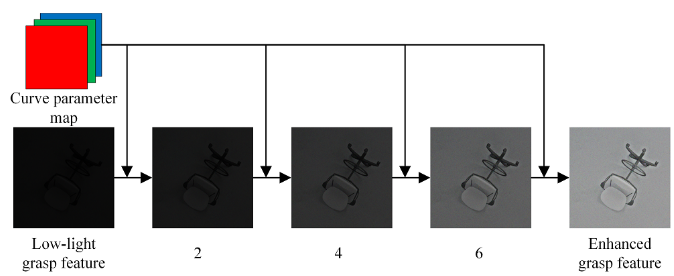


Fig. 4. Grasping feature enhancement iteration results.

3) Low-light grasp features enhancement head: The underlying features F_{low} extracted by the encoder are sequentially fed into the two transposed convolution modules M_{TCBL}, and then a convolution operation is performed to obtain the final curve parameter map. The curve parameter map iteratively enhances the given grasp features and the enhancement result is used as the input for the next iteration, incrementally enhancing the input grasp features. The process for each iteration is expressed as follows:

$$\mathcal{F}_n\left(x\right) = \mathcal{F}_{n-1}\left(x\right) + \mathcal{A}\left(x\right)\mathcal{F}_{n-1}\left(x\right)\left(1 - \mathcal{F}_{n-1}\left(x\right)\right) \tag{7}$$

where $\mathcal{A}$ represents curve parameter map, $\mathcal{F}_i$ denotes the features of ith iteration. The first iteration grasp feature is the input low-light grasp feature and the number of iterations is 8.

The visualization results for each iteration are shown in Fig. 4. The detailed derivation is described in [20]. The specific processes are shown below:

$$I_{enhanced} = \mathcal{T}^8 \left(Conv2D \left(M^2_{TCBL} \left(F_{low}\right)\right)\right) \tag{8}$$

where $I_{enhaced}$ denotes the enhanced grasp feature and $\mathcal{T}$ denotes the iterative process.

B. Loss function

1) Grasp detection loss: Regarding the selection of the loss function for grasp detection, we use the smooth L_1 loss to constrain the optimisation process, as defined below:

$$L_{grasp}(\hat{P}, P) = \sum_{i}^{N} \sum_{k} l_1(\hat{P}^k_i, P^k_i), k \in \{\boldsymbol{\Phi}, \boldsymbol{W}, \boldsymbol{Q}\} \tag{9}$$

where $\hat{P}$ denotes the grasping prediction, P denotes the corresponding label and N represents the number of grasp candidates. L_1 loss is defined as:

$$l_1(\hat{P}^k_i, P^k_i) = \begin{cases} 0.5 \cdot (\hat{P}^k_i - P^k_i)^2, & \left|\hat{P}^k_i - P^k_i\right| < 1 \\ \left|\hat{P}^k_i - P^k_i\right| - 0.5, & otherwise \end{cases} \tag{10}$$

2) Unsupervised low-light grasp feature enhancement loss: Specifically, multiple losses [26], including loss L_{sc} for enhancing the spatial consistency of the image, loss L_{ce} for controlling the exposure, loss L_{ccd} for correcting colour deviations and loss L_{als} for adjusting light smoothness, are used to achieve the structural and perceptual evaluation of enhanced grasp features. The definition of each loss is sequentially listed.

$$L_{sc} = \frac{1}{F} \sum_{p=1}^{F} \sum_{q \in \omega(p)} \left(\left|(H_p - H_q)\right| - \left|(Z_p - Z_q)\right|\right)^2 \tag{11}$$

where F denotes the number of square regions and $\omega(p)$ is the four adjacent square regions centred on region p, as shown in Fig. 5. We indicate H and Z as the average intensity values of the square regions in the enhanced and input grasp features.

$$L_{ce} = \frac{1}{C} \sum_{e=1}^{C} |H_e - E| \tag{12}$$

where C denotes the number of non-overlapping square regions, E defines the average gray value of normal-light image.

$$L_{ccd} = \sum_{\forall(i,j) \in \mathrm{II}} \left(A^i - A^j\right)^2, \mathrm{II} = \{(R, G), (R, B), (G, B)\} \tag{13}$$

where A^i and A^j are the average intensity value of channel i and j in the enhanced grasp feature.

$$L_{als} = \frac{1}{M} \sum_{m=1}^{M} \sum_{c \in \eta} \left(|\bigtriangledown_x \mathcal{A}^c| + |\bigtriangledown_y \mathcal{A}^c|\right)^2, \eta = \{R, G, B\} \tag{14}$$

where M is the number of iterations and $\bigtriangledown_x$ and $\bigtriangledown_y$ denote gradient operations. The loss of the visual enhancement branch is defined as follows:

$$L_{enhance} = L_{sc} + \alpha L_{ce} + \beta L_{ccd} + \gamma L_{als} \tag{15}$$

where α, β and γ are set to 10, 5 and 1600, respectively.

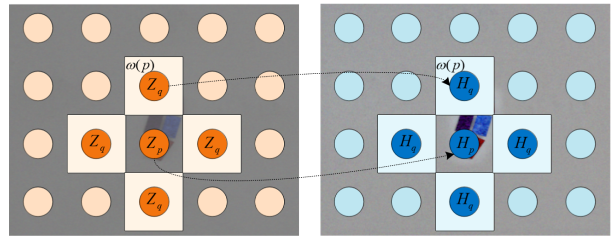


Fig. 5. Schematic diagram about the loss of spatial consistency. Left: low-light image, right: enhanced image.

3) Total loss: The total loss of the model consists of two parts, the grasping detection loss L_{grasp} and the low-light grasp feature enhancement loss $L_{enhance}$, which is expressed as:

$$L_{total} = \lambda L_{grasp} + \mu L_{enhance} \tag{16}$$

where λ and μ are set to 1 and 0.9, respectively.

IV. Experiments and Results

A. Low-light grasping datasets construction

To the best of our knowledge, there exists no grasp detection dataset specifically sampled under low-light condition. In order to simulate the low-light condition, we generate low-light Cornell dataset and low-light Jacquard dataset (URL: https://github.com/Sxudig/Low-light-grasp-dataset) based on the existing Cornell [22] and Jacquard dataset [27] by consecutively adjusting the brightness of the image and adding Gaussian noise. The complete procedure is demonstrated in Fig. 6.

1) Adjust the brightness: The normal light image is defined as I and the output low-light image is defined as I_{low}. The specific processing is shown below:

$$I_{low} = I^g \tag{17}$$

when $g > 1$, the resulting image is darker than original image, and when $g < 1$, the resulting image is brighter. In our experiments we set the values of g as 1.2, 1.5 and 1.8 to simulate different light conditions.

2) Add Gaussian noise: Gaussian noise is a class of noises whose probability density functions follow Gaussian distribution, which commonly appears under low-light and non-uniform illuminations. Hence, we add Gaussian noise to the brightness-adjusted image in order to further simulate the low-light imaging environment. The Gaussian probability distribution is shown below:

$$p(z) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(z-\mu)^2/2\sigma^2} \tag{18}$$

where z denotes the gray value, μ and σ represent the expected value and standard deviation of z, respectively. The level of added noise can be controlled by adjusting μ and σ.

B. Implementation details

A single NVIDIA RTX 3090 GPU with 24G of memory is used for model training and testing, and the entire model implementation is based on PyTorch. In addition, the operating system is Ubuntu 20.04.

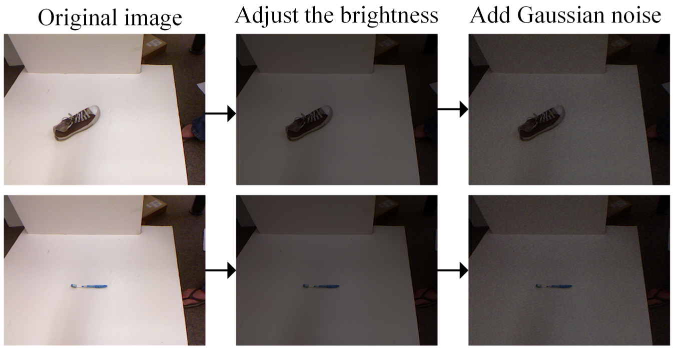


Fig. 6. Procedure of constructing low-light grasp detection datasets.

1) Evaluation metric: The common rectangular metric proposed by [22] is used. The predicted grasping rectangle is considered correct when it satisfies both of the following Eq.(19) and Eq.(20):

$$\left|\hat{Angle} - Angle\right| < 30^\circ \tag{19}$$

where $\hat{Angle}$ denotes the predicted grasp angle and $Angle$ denotes the ground truth.

$$|\hat{P} \cap P| / |\hat{P} \cup P| > 0.25 \tag{20}$$

where $\hat{P}$ denotes the predicted grasp rectangle and the P denotes the ground truth. Besides, $|\star|$ is the area of $\star$.

2) Training details: For low-light Cornell and Jacquard dataset, 90% of the low-light images are used for training while the remaining 10% are used for testing, respectively. Besides, we set the initial learning rate to 0.001 and the parameters are optimised using the adaptive moment estimation (Adam) method. The learning rate is sequentially adjusted during training according to the cosine annealing strategy.

C. Quantitative and qualitative results

Quantitative and qualitative experiments are conducted to compare comparative methods with our VERGNet on low-light Cornell and Jacquard dataset, respectively. From the results in Table I and II, our method achieves 98.9%, 98.3%, and 97.7% accuracy on low-light Cornell datasets under different luminance, which is 1.2%, 1.0%, and 0.7% higher than GR-ConvNetV2, respectively. In addition, we achieve 91.2%, 90.6% and 90.2% accuracy on the low-light Jacquard dataset, which is also a substantial improvement relative to the other methods, further demonstrating the effectiveness of our method. Regarding the inference speeds of different approaches, VERGNet takes about 53ms per image, which is relatively slower compared to other methods, but can still basically satisfy the requirement of real-world robotic grasping.

In addition to the quantitative results, we also visualize the grasping poses predicted by different methods, as well as the output maps of $\boldsymbol{\Phi}$, $\boldsymbol{W}$, $\boldsymbol{Q}$ and the low-light grasp feature enhancement results. As illustrated in Fig. 7 and Fig. 8, when the parameter g is taken as 1.5 and 1.8, VERGNet predicts a higher confidence of the grasping quality relative to GR-ConvNet. Meanwhile, it observes that the predicted grasping quality maps by VERGNet are more complete and concentrated. In addition, according to Fig. 8, when g is taken as 1.5, the edges of the quality maps predicted by VERGNet are clear, while that by GR-ConvNet are fuzzy, which suggests that our model is more capable of distinguishing the objects from backgrounds. Notice than since GR-ConvNet lacks the specialized feature enhancement part, it can only predict graspable rectangles, leading to missing images in the corresponding positions of Fig. 7 and Fig. 8. Finally, we also provide multi-grasp results to verify its generality to different grasp locations of objects, as shown in Fig. 9.

TABLE I
DETECTION ACCURACY COMPARISON OF DIFFERENT METHODS ON LOW-LIGHT CORNELL DATASET.

Author	Algorithm	Accuracy (%)			Speed
		1.2	1.5	1.8	(ms)
Redmon *et al.* [9]	AlexNet	72.3	68.3	63.1	76
Morrison *et al.* [23]	GG-CNN2	87.6	83.1	80.9	20
Kumra *et al.* [12]	GR-ConvNet	97.2	96.1	95.5	20
Kumra *et al.* [28]	GR-ConvNetV2	97.7	97.3	96.9	20
Ours	VERGNet	**98.9**	**98.3**	**97.7**	53

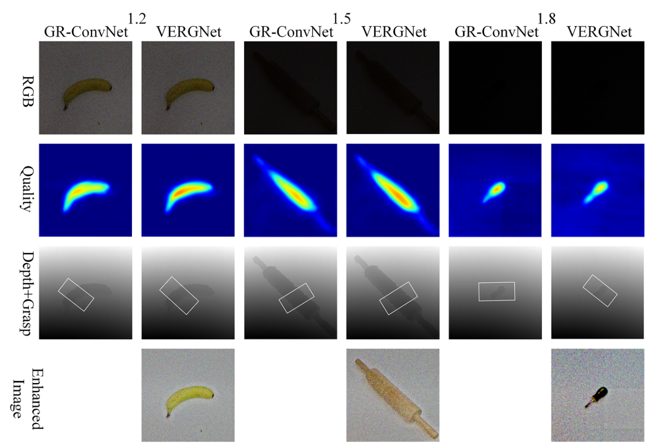


Fig. 7. Detection results comparison of different methods on low-light Cornell dataset. Since there exists no feature enhancement function, the outputs by GR-ConvNet are blank.

TABLE II
DETECTION ACCURACY COMPARISON OF DIFFERENT METHODS ON LOW-LIGHT JACQUARD DATASET.

Author	Algorithm	Accuracy (%)		
		1.2	1.5	1.8
Morrison *et al.* [23]	GG-CNN2	78.1	77.5	76.8
Kumra *et al.* [12]	GR-ConvNet	88.9	87.8	84.6
Kumra *et al.* [28]	GR-ConvNetV2	89.6	87.4	85.7
Ours	VERGNet	**91.2**	**90.6**	**90.2**

To verify whether our method can suppress false-positive grasping, we further conduct experiments by setting Jaccard

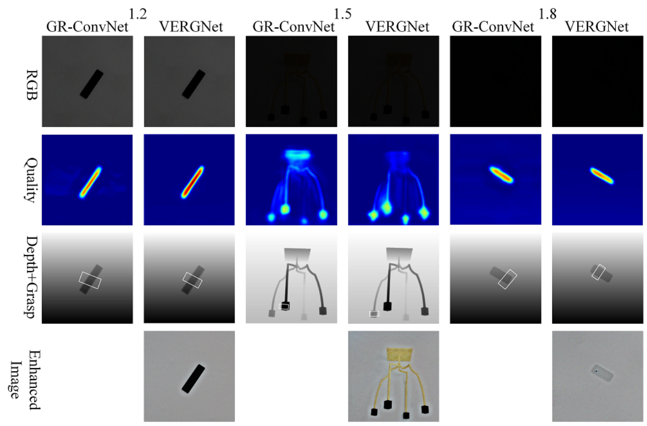


Fig. 8. Detection results comparison of different methods on low-light Jacquard dataset. Since there exists no feature enhancement function, the outputs by GR-ConvNet are blank.

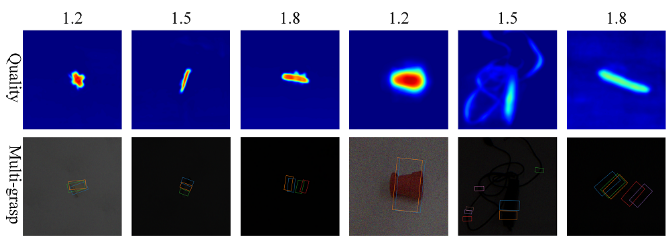


Fig. 9. Illustration of multiple grasp detection results. Left: images from low-light Jacquard dataset, right: images from low-light Cornell dataset.

index to be 0.25, 0.3, 0.35 and angle different to be 30°, 25°, 20°, respectively. The experimental results are shown in Fig. 10, which indicates that the accuracy can reach 88.1%, 87.9% and 87.5% at different brightness when the Jaccard index is 0.35, and 89.7%, 89.0% and 89.2% at different brightness when the angle difference is 25°. As Jaccard index increases and angle difference decreases, the detection accuracy decreases slightly, which to some extent proves that our method can suppress false positive grasping. In addition, the experimental results of GR-ConvNet for low-light Jacquard dataset at different Jaccard indexes and angle differences are shown in Fig. 11. It reveals that under identical angle difference and Jaccard index, the detection accuracy of GR-ConvNet is inferior to that of VERGNet. Besides, as the light condition becomes severer, VERGNet indicates a slighter decrease in detection accuracy relative to GR-ConvNet, verifying its robustness and adaptability to light changes.

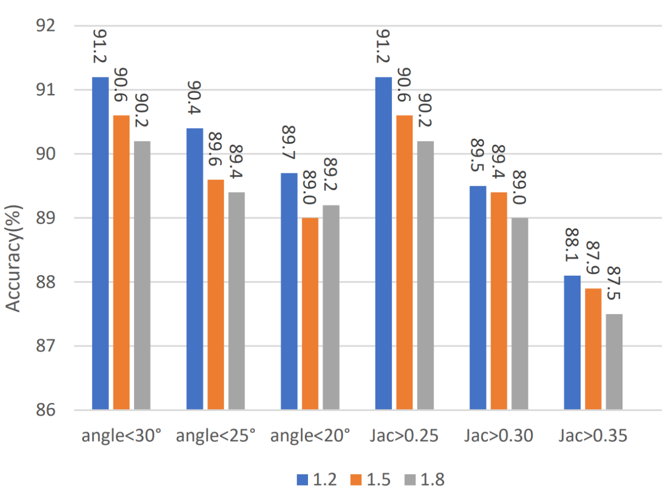


Fig. 10. The experimental results of VERGNet for low-light Jacquard dataset at different Jaccard indexes and angle differences.

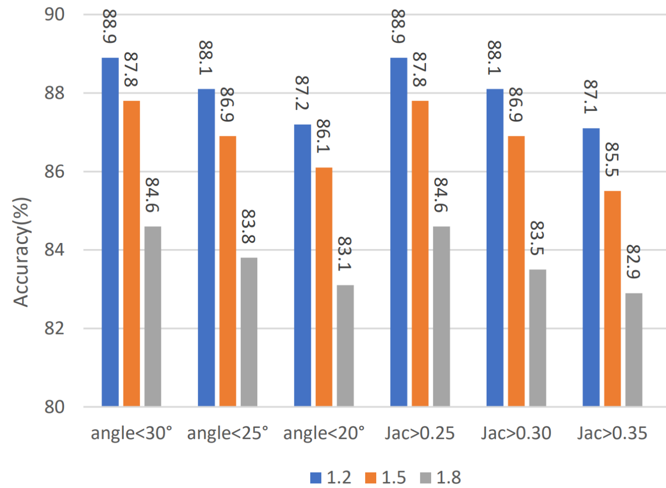


Fig. 11. The experimental results of GR-ConvNet for low-light Jacquard dataset at different Jaccard indexes and angle differences.

D. Ablation study

To verify the effectiveness of the low-light grasping feature enhancement branch B_{FE}, an ablation study is conducted in this section. Specifically, the grasp detection models removing B_{FE} and RGB input are respectively trained, and the corresponding detection accuracies are reported in Table III. When the input is depth image, the accuracies of detection model without B_{FE} under Cornell and Jacquard dataset are 94.4% and 88.9% respectively. When it comes to RGB-D inputs, the accuracies increases to 97.7% and 89.2%, showing that although the RGB image captured under low-light condition is weak and indistinct in colour and texture, they are still beneficial for the accuracy improvement of detection model. After introducing B_{FE}, our VERGNet achieves the best performances of 98.3% and 90.6% detection accuracy. It proves that the low-light grasping feature enhancement branch is able to further highlight and extract grasping-specific features, compensating the side effect of weak lighting environment.

TABLE III
RESULTS OF ABLATION EXPERIMENTS ON THE LOW-LIGHT CORNELL AND JACQUARD DATASET.

Modality	Baseline	B_{FE}	Accuracy (%)	
			Cornell	Jacquard
Depth	✓		94.4	88.9
RGB-D	✓		97.7	89.2
RGB-D	✓	✓	**98.3**	**90.6**

E. Real-world robotic grasping

To verify the feasibility of our method in real scenarios, we construct a low-light robotic grasping system, which consists of the UR5 robotic arm, the Backyard E140 gripper, the RealSense camera, the objects to be grasped, and the host computer, see Fig. 12 for details. In addition, the shade cloth is used to cover the external support frame of the system in order to simulate the low-light imaging condition.

Fig. 12. The built low-light robotic grasping system and objects to be grasped.

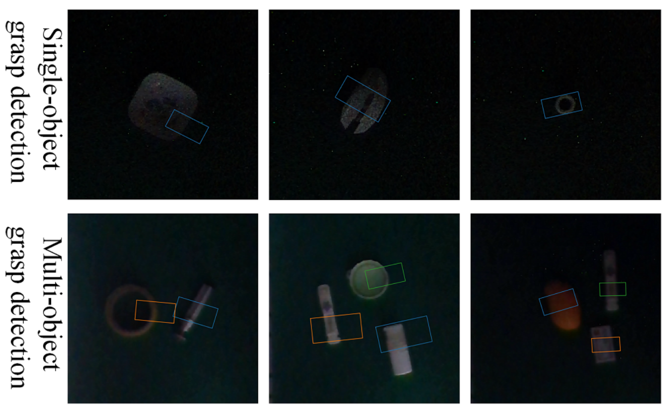


Fig. 13. Single-object and multi-object grasp detection results in real-world low-light environment.

Specifically, we select 20 common objects and sample 111 images in low-light scene. The sampled data are used to fine-tune our VERGNet trained on low-light Cornell and Jacquard datasets. Each object is grasped 10 times respectively, and the average grasping success rate is finally calculated. Note that a successful grasping of object means that the object does not fall during the whole grasping process. Repetitive experiments show that the average grasping success rate of our method reaches 94.6%, while that of [12] is 87.4%. In Fig. 13, some of the single-object and multi-object grasp detection results under real-world low-light imaging environment are given, and in Fig. 14, we show the typical procedures during low-light robotic grasping.

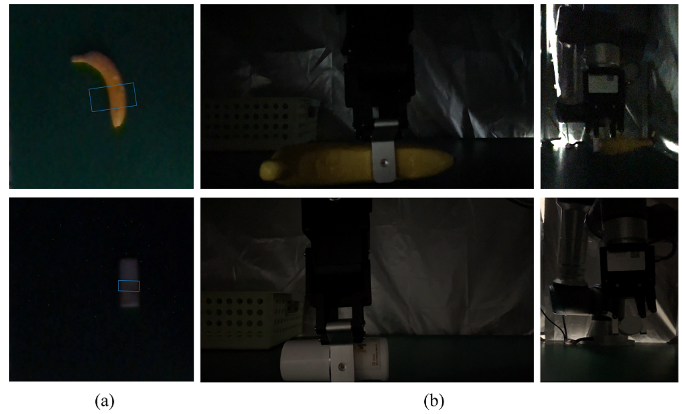


Fig. 14. Visualization of real-world robotic grasping procedure. (a) Grasping rectangle prediction results. (b) Grasping process.

V. Conclusion

In this work, we propose a grasping detection network for low-light conditions. Specifically, a residual module that fuses coordinate attention is first added to the network to make the model more accurate in capturing the location information of target. Then, we impose semantic-level constraints on the extraction of grasping features by using unsupervised visual enhancement methods, reducing the dependence on paired data. Meanwhile, under the newly-constructed low-light Cornell dataset and low-light Jacquard dataset, the proposed VERGNet outperforms the comparative methods in terms of detection accuracy, verifying the effectiveness of enhancing visual features in low-light robotic grasping. Finally, we construct a robotic grasping platform for low-light environments to prove the effectiveness of our method.

References

[1] Jinshan Wang, Bo Tao, Zeyu Gong, Wenfu Yu, and Zhouping Yin. A mobile robotic 3-d measurement method based on point clouds alignment for large-scale complex surfaces. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021.

[2] Matei Ciocarlie, Kaijen Hsiao, Edward Gil Jones, Sachin Chitta, Radu Bogdan Rusu, and Ioan A Şucan. Towards reliable grasping and manipulation in household environments. In *Experimental Robotics: The 12th International Symposium on Experimental Robotics*, pages 241–252. Springer, 2014.

[3] Zhengxiong Liu, Zhenyu Lu, Yang Yang, and Panfeng Huang. Teleoperation for space manipulator based on complex virtual fixtures. *Robotics and Autonomous Systems*, 121:103268, 2019.

[4] Zhiyun Yin and Yujie Li. Overview of robotic grasp detection from 2d to 3d. *Cognitive Robotics*, 2:73–82, 2022.

[5] Hongkun Tian, Kechen Song, Song Li, Shuai Ma, Jing Xu, and Yunhui Yan. Data-driven robotic visual grasping detection for unknown objects: A problem-oriented review. *Expert Systems with Applications*, 211:118624, 2023.

[6] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1792–1800, 2022.

[7] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark-domain adaptation method for merging multiple models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 345–359. Springer, 2020.

[8] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[9] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1316–1322. IEEE, 2015.

[10] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[11] Lu Chen, Panfeng Huang, and Zhongjie Meng. Convolutional multi-grasp detection using grasp path for rgbd images. *Robotics and Autonomous Systems*, 113:94–103, 2019.

[12] Sulabh Kumra, Shirin Joshi, and Ferat Sahin. Antipodal robotic grasping using generative residual convolutional neural network. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9626–9633, 2020.

[13] Yuanhao Li, Panfeng Huang, Zhiqiang Ma, and Lu Chen. A context-free method for robust grasp detection: Learning to overcome contextual bias. *IEEE Transactions on Industrial Electronics*, 69(12):13121–13130, 2022.

[14] Zhenyu Lu, Lu Chen, Hengtai Dai, Haoran Li, Zhou Zhao, Bofang Zheng, Nathan F. Lepora, and Chenguang Yang. Visual-tactile robot grasping based on human skill learning from demonstrations using a wearable parallel hand exoskeleton. *IEEE Robotics and Automation Letters*, 8(9):5384–5391, 2023.

[15] Sheng Yu, Di-Hua Zhai, Yuanqing Xia, Haoran Wu, and Jun Liao. Se-resunet: A novel robotic grasp detection method. *IEEE Robotics and Automation Letters*, 7(2):5238–5245, 2022.

[16] Sheng Yu, Di-Hua Zhai, and Yuanqing Xia. Skgnet: Robotic grasp detection with selective kernel convolution. *IEEE Transactions on Automation Science and Engineering*, pages 1–12, 2022.

[17] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *ArXiv*, abs/1808.04560, 2018.

[18] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

[19] Yang Wang, Yang Cao, Zheng-Jun Zha, Jing Zhang, Zhiwei Xiong, Wei Zhang, and Feng Wu. Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2015–2023, 2019.

[20] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.

[21] Xinyi Liu, Qi Xie, Qian Zhao, Hong Wang, and Deyu Meng. Low-light image enhancement by retinex-based algorithm unrolling and adjustment. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2023.

[22] Yun Jiang, Stephen Moseson, and Ashutosh Saxena. Efficient grasping from rgbd images: Learning using a new rectangle representation. In *2011 IEEE International Conference on Robotics and Automation*, pages 3304–3311, 2011.

[23] Douglas Morrison, Peter Corke, and Jürgen Leitner. Learning robust, real-time, reactive robotic grasping. *The International Journal of Robotics Research*, 39(2-3):183–201, 2020.

[24] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[25] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13708–13717, 2021.

[26] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4225–4238, 2021.

[27] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018.

[28] Sulabh Kumra, Shirin Joshi, and Ferat Sahin. Gr-convnet v2: A real-time multi-grasp detection network for robotic grasping. *Sensors*, 22(16):6208, 2022.