Integrating Vision-Language Models for Enhanced Robotic Grasping and Interaction Using RGB Image and Prompt

Nguyen Khac Toan and Nguyen Truong Thinh*

Institute of Intelligent and Interactive Technologies, University of Economics Ho Chi Minh City—UEH, Vietnam
Email: toannk@ueh.edu.vn (T.N.K.); thinhnt@ueh.edu.vn (N.T.T.)
*Corresponding author

***Abstract*—Object detection and grasping is one of the critical challenges in robotic research, particularly when working in complex environments with diverse objects in terms of shape and position. Although methods using RGB images have shown promising results in simpler scenarios, they still face numerous issues in more complex scenes, especially when objects overlap. Furthermore, prior research has primarily focused on object grasping, without focusing on addressing the interaction capabilities between robots and users during the grasping process. Recent advancements in vision-language models have opened up significant potential for the development of human-robot interaction systems based on multimodal data. This paper presents an integrated model combining computer vision and language models to enhance object detection and grasping capabilities in real-world environments. The proposed approach consists of three key steps (1) identifying the locations of objects and generating segmentation masks using a visual-language model; (2) grasp candidates are predicted from the generated masks and bounding boxes via the Grasp Detection Head; and (3) the candidates are optimized and refined using the Grasp Refinement Head. The integration of vision-language models in the proposed approach not only enhances the ability of robot to understand the semantics of language, enabling more accurate grasping decisions, but also strengthens the interaction capabilities of robot with users. Experimental results demonstrate that the proposed model achieves higher grasping accuracy compared to existing methods, particularly in complex scenes with multiple objects. Additionally, the model also shows its ability to understand complex contexts through Interactive Grasp experiments.**

***Keywords*—robot grasping, robot grasping detection, grasp refinement, vision-language integration, image-text integration**

I. Introduction

Robots and their applications have become increasingly prevalent in modern life [1]. Along with this, object grasping based on RGB images has emerged as an important area in robotic research, particularly in complex environments with overlapping objects and diverse shapes [2, 3]. This task is crucial for developing robotic systems capable of performing precise and efficient manipulation in real-world applications. Previous studies have achieved high accuracy in simpler scenarios, such as the work [4]. However, in more complex scenes with multiple objects, this problem still presents significant challenges, especially when it comes to detecting and grasping objects with limited training data or objects that have not appeared in the training data. Moreover, current methods primarily focus on object detection or grasp detection based on a unimodal data, such as RGB or RGB-D images. That less focus on integrating diverse types of information, such as audio or text. This limits the effectiveness of robots in accurately identifying object locations as well as grasping positions. Some modern approaches have attempted to combine deep learning for object detection and segmentation, achieving high performance on benchmark datasets. However, these methods are still limited in their handling of information from RGB images and not yet to fully exploit the potential of diverse and complex data, such as user descriptions of the objects to be grasped.

In recent years, the rapid development of large language models has opened new potentials for the field of language-driven robotic control. Models such as GPT [5], PathwaysLanguage Model (PaLM) [6], and Large Language Model Meta AI (LLaMA) [7] have demonstrated exceptional capabilities in understanding and processing natural language commands. They enable robots to perform complex tasks in real-world environments. This progress has been driven by advancements in the Transformer architecture [8] and the ability to train on large-scale, multimodal datasets. Notably, large language models have been integrated with vision-language models, such as Contrastive Language-Image Pre-training (CLIP) [9] and Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP) [10], allowing computer not only to comprehend language but also to associate semantic information with image data. This integration has unlocked significant potential for developing robotic systems that can accurately recognize and manipulate objects based on language commands to performing complex multi-step

Manuscript received February 10, 2025; revised March 17, 2025; accepted April 23, 2025; published September 19, 2025.

doi: 10.18178/ijmerr.14.5.500-510

tasks. It was done through several steps of identifying and grasping specific objects. The language in text form provides rich context, significantly enhancing performance in complex scenarios. In particular, the combination of language and RGB image data offers a deeper understanding of context and tasks, even when objects have limited or unappeared in training data. This development not only enhances the flexibility and efficiency of robots grasping in real-world situations but also strengthens human-robot interaction, promoting the application of intelligent robots in industries, services, and daily life. However, research related to the integration of language in human-robot interaction, specifically in guiding robots to grasp objects, remains quite limited.

In this research, a deep neural network design is presented that combines grasp detection and semantic segmentation, utilizing RGB images and user-provided text instructions. This approach employs a vision-language model, an advanced system that integrates visual data with natural language to improve the robot capability to understand and navigate complex environments effectively. The research focuses on grasping household items, aiming for applications in home-assistive robots. This approach not only focuses on the detection of grasp candidates but also improves accuracy through refinement steps based on semantic understanding and natural language information. This method enhances performance in complex scenes and opens up the potential for more effective collaboration between robots and humans in real-world environments.

The proposed model was evaluated for its accuracy in object segmentation and grasp location detection on two datasets, Object Clutter Indoor Dataset for Grasp (OCID Grasp) [11] and Jacquard [12]. The results demonstrate that the proposed method improves accuracy in grasp location determination, particularly in complex scenarios. In conclusion, the key contributions of this research can be summarized as follows:

- A method that integrates language models to enhance grasp location detection and improve human-robot interaction.
- Improved accuracy in object detection and segmentation, particularly in cases with complex contexts.
- Enhanced segmentation of objects with limited or no training data, leveraging descriptive text.

II. Related Work

Traditional methods for grasp detection rely on geometric information, physical modeling, and force analysis [13]. While these methods are effective in controlled environments, they often struggle to handle complex real-world scenes with multiple objects and clutter. The development of deep learning has driven the popularity of data-driven approaches in the field of grasp detection [14]. Early methods, such as Ref. [4], employed deep neural networks with supervised learning to predict multiple grasp candidates for each object, achieving significant improvements over traditional methods.

A varied dataset containing grasp detection data is considered an important factor in improving the training and evaluation of neural networks. The Cornell dataset [4] and the Jacquard dataset [12] are commonly used in grasping research, with annotated bounding boxes being included to allow grasping parameters to be predicted from RGB or RGB-D images. These datasets provide an important foundation for the development of modern grasp detection models.

Many researches have focused on improving Convolutional Neural Networks (CNNs) [15] to address the limitations of previous methods. Morrison *et al.* [16] introduced GGCNN, a real-time neural network model designed to predict grasp poses directly from depth images without the need for time-consuming candidate sampling. This appoarch eliminates the need for discretized sampling of grasp candidates, reducing computation time and achieving a grasp success rate of 83%. GGCNN was further enhanced by adopting a multi-view approach, which resulted 94% in the grasping success rate in cluttered environments [17]. Kumra *et al.* [18] developed GR-ConvNet, a convolutional neural network that addresses the vanishing gradient problem and achieves high accuracy on the Cornell and Jacquard datasets. Yu *et al.* [19] proposed Squeeze-and-Excitation ResUNet, demonstrating that this mechanism enhances the generalization ability of model across different datasets.

Two-stage detection methods, including region proposal networks and object detectors, have been effectively applied in grasp detection [20]. In the first stage, the region proposal network identifies candidate regions. In the second stage, features are extracted from these regions to detect the object. Although high accuracy is achieved, this approach often requires substantial computational time. To reduce computation time, one-stage detectors have been developed [21]. This method divides the input image into a grid and performs detection on each cell. However, this approach often results in reduced accuracy compared to two-stage methods.

Moreover, semantic segmentation plays a crucial role in supporting grasp detection. Several researches have applied encoder-decoder network architectures to improve accuracy in object recognition [22, 23]. Fine-grained semantic segmentation at the object level, which helps clearly identify areas related to each object in the image, has been the focus of some studies [24–26]. In robot vision, various methods have been developed to segment unknown objects, helping the manipulation process of robot in complex scenarios. A notable contribution comes from the research of Araki et al. They designed a multitask deep neural network to integrate semantic segmentation and grasp detection [27]. These advancements have highlighted the importance of semantic segmentation in enhancing the effectiveness of grasp detection and object manipulation in modern robotic applications.

Recently, the integration of natural language with grasp detection has opened up a new research direction. It enables robots to better understand and execute human commands with more flexibility. Approaches in [28] leverage large language models and vision-language models to enhance human-robot interaction. However, many current studies focus solely on single-object scenarios or limit grasp detection to 3D space [29, 30]. This reduces the effectiveness of robotic systems in complex real-world environments.

Inspired by the previous works, this study integrates object detection and object recognition models based on user prompts and 2D image into a grasp location detection model to improve accuracy. We validate model on the OCID Grasp and Jacquard datasets, with the goal of accurately predicting the grasp locations of objects based on user prompts. This approach effectively addresses challenges related to extracting multimodal features including both images and text while also enhancing user interaction with the robot.

III. Grasp Detection Model

A. Overview

This section presents the architecture of the proposed model, which integrates a visual-language model for the task of grasp detection. The flowchart in Fig. 1 provides a visual representation of the proposed grasp detection pipeline. The process begins with input data, comprising an RGB image and a descriptive text prompt, which are jointly processed to enhance object localization. This input is fed into the Object Detection module, which leverages a vision-language model to generate bounding boxes. From there, two parallel processes are initiated. Firstly, Image Processing and Feature Extraction prepares visual features for grasp analysis. These outputs are then used in the Grasp Candidates Prediction step to estimate potential grasp poses. Secondly, Segmentation Mask Generation detect object boundaries. Both results are fed into Grasp Refinement stage, where semantic and geometric features are fused to enhance the grasping accuracy, to further correct. The pipeline concludes with the Grasp Output, expressed as a five-dimensional vector, representing the optimal position, size, and angle for object grasping.

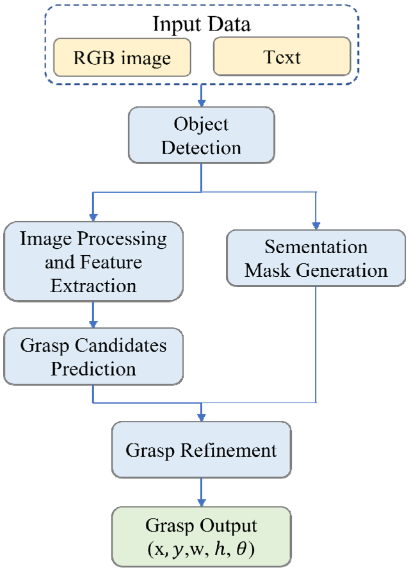
Input Data
RGB image
Text
Object Detection
Image Processing and Feature Extraction
Sementation Mask Generation
Grasp Candidates Prediction
Grasp Refinement
Grasp Output
(x, y,w, h, θ)

Fig. 1. The proposed methodology.

Specifically in this study, we utilize typical visual-language models such as Grounding DINO [31], which stands for Grounding DEtection TRansformer with Improved deNoising anchOr Boxes, and Segment Anything Model (SAM) [25]. These models are used to identify bounding boxes and perform semantic segmentation for the objects within the image, as shown in Fig. 2. Grounding DINO model analyzes both the input image and descriptive text prompt to determine the positions of each object. Then a mask is generated to highlight the area corresponding to each object by SAM.

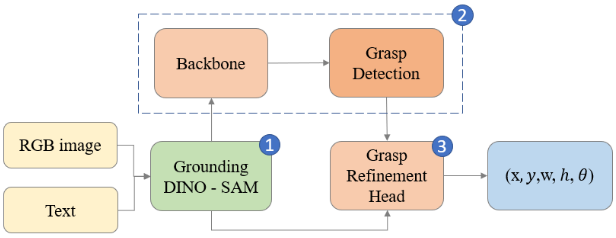

Fig. 2. Architecture of proposed grasp detection model.

Based on the identified bounding boxes, the objects within these regions are cropped and passed through the backbone. It is a deep convolutional neural network designed to extract key features from the images. The backbone processes the cropped images and transforms them into feature maps. They represent the spatial and shape information of the objects, providing the necessary input for subsequent grasp prediction steps. These features are then forwarded to the Grasp Detection Head, where the model predicts potential grasp candidates including parameters of center coordinates (x, y), size (h, w), and rotation angle θ. This process generates a list of grasp candidates for each object, with each candidate assigned its parameters. Finally, the mask information from the Grounded SAM module is combined with the grasp candidates from the Grasp Detection Head. They are then refined by the Grasp Refinement Head. In this stage, parameters such as the center coordinates, width, height and rotation angle are optimized to ensure the highest accuracy and manipulation capability.

The final output consists of optimal grasp points that accurately located on the objects in the input image. With this approach, the model enhances the ability of robot to perform effective manipulation even in complex environments, where multiple objects with diverse shapes and intricate relationships exist. The integration of vision-language models improves the ability of robot to recognize objects that are either rarely seen or not present in the training data. This study focuses on the three main stages highlighted in the blue circle of Fig. 2. The entire architecture can be formulated into three optimized stages as Eq. (1). The targets are determined by the 2D image captured the enviroment I and the description of object T. The image of the object is cropped based on the bounding box returned by $E_{G\text{-}SAM}$ denoted as I_{crop}. The term $\mathrm{M}_{\mathrm{crop}}$ and C represent the mask results of the $\mathrm{E}_{\mathrm{G\text{-}SAM}}$ process cropped according to the bounding box, and the list of grasp candidates, respectively. The three parameters

$\alpha_o,..,\alpha_2$ represent the sequence in which the stages are executed. In each stage, only one term in Eq. (1) is applied.

$$\begin{aligned} E(I,T) &= \alpha_0 E_{G-SAM}(I,T) \\ &+ \alpha_1 E_{Det}(I_{crop}) \\ &+ \alpha_2 E_{Ref}(M_{crop}, C) \end{aligned} \tag{1}$$

Unlike previous methods that typically rely solely on image data without robust user interaction capabilities, proposed approach utilizes vision-language models to understand user commands. Specifically, the visual language model like Grounding DINO is integrated to detect objects within the image based on user prompts. A notable strength of this method is its ability to recognize objects that the robot has never encountered before. In addition, the SAM is integrated to determine the segmentation masks of objects, enhancing precision in adjusting grasping parameters and improving accuracy.

B. Grounding DINO-SAM

The Grounding DINO-SAM architecture is designed similarly to the Grounded SAM [32] architecture. It is an advanced model that addresses the challenges of detection and segmentation in open-vocabulary tasks. Grounding DINO-SAM focuses on identifying and segmenting objects based on prompt input. The model combines the power of Grounding DINO, an open-vocabulary object detection system based on text, and Segment Anything Model (SAM), a robust segmentation model. The process operates in two main stages. In first stage, Grounding DINO detects bounding boxes based on prompt. In second stage, results of previous stage are processed by SAM to generate detailed segmentation masks. However, Grounding DINO-SAM differs from Grounded SAM as it solely integrates Grounding DINO and SAM, without incorporating RAM or other models. This design choice aims to minimize model complexity and enhance processing speed. The architecture of Grounding DINO-SAM is illustrated in Fig. 3.

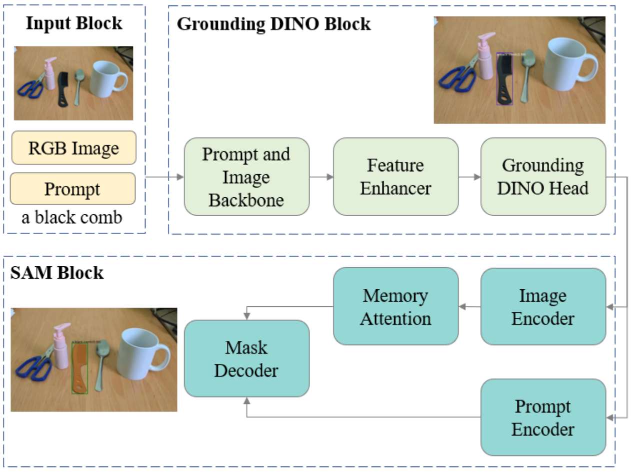

Fig. 3. Architecture of Grounding DINO-SAM.

1) Prompt and image backbone

The Prompt Backbone uses BERT-base model [33], a well-known Transformer network in natural language processing. BERT-base consists of 12 layers, a hidden size of 768 and 12 self-attention heads, resulting in a total of about 110 million parameters. The key distinction is using sub-sentence level text representations [31] to reduce unwanted interference between unrelated phrases in a sentence. The Image Backbone employs the Swin Transformer [34] consisting of 4 stages, where stages 1, 2, and 4 each contain 2 Swin Transformer Blocks, while stage 3 uniquely contains 18 blocks. This is an efficient image-transforming neural network. The image backbone is used to extract image features.

2) Feature enhancer

This is a key part of Grounding DINO, designed to enhance the integration of information between language and image. The architecture of the feature enhancer consists of three main components Self-Attention, Image-to-Text Cross-Attention, and Text-to-Image Cross-Attention. The general mathematical formulation for attention is presented in Eq. (2). Where Q, K, and V represent the query matrix, key matrix, and value matrix, respectively. d_k is the dimension of the embedding vector in the key matrix. The primary objective of the Feature Enhancer is to create feature consistency between the two data domains including images and prompt. It improves contextual understanding and enhances object detection performance in complex scenarios.

$$Attention(Q,K,V) = sofmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

3) Grounding DINO head

This head is responsible for querying and selecting the necessary information from the embeddings of both the image and text. These embeddings are processed after passing through the Feature Enhancer block to produce the results. The key components of this block include the Language-guided Query Selection and Cross-Modality Decoder. The Language-guided Query Selection utilizes language information to guide the selection of relevant queries from the image features. The goal is to identify the top K features with the highest correlation from the image feature set E_I. The text features are denoted as E_T. The $Max^{(-1)}$ function is applied to extract the maximum values along dimension −1. The result I_K is the set of selected indices from the image feature set. Eq. (3) represents this feature selection process. Subsequently, the Cross-Modality Decoder merges the information from the two semantic domains.

$$I_K = Top_K\left(Max^{(-1)}\left(E_I E_T^{\mathrm{T}}\right)\right) \tag{3}$$

4) Prompt and image encoder

This module is similar to the Prompt and Image Backbone but uses different models to extract features due to the different tasks being performed. In the SAM block, the Vision Transformer pre-trained with MAE [35, 36] is used as the image encoder. MAE is an asymmetric encoder-decoder design, where the encoder processes only a small subset of unmasked image patches without incorporating mask tokens, while a lightweight decoder reconstructs the original image from the latent representation combined with mask tokens. This process involves randomly masking a substantial proportion of the input image patches, with the encoder encoding the remaining visible patches and the decoder leveraging this encoded information alongside mask tokens to predict the missing pixel values. The text encoder from the CLIP model basing Transformer architecture with a multi-layer structure and self-attention mechanisms to generate contextualized embeddings.

5) Memory attention

This component in Grounding DINO is designed to integrate information from previous frames and predictions. It enables the model to handle long-term context in videos or a sequence of related images. Memory Attention enhances the model ability to make more accurate predictions in complex scenarios, particularly in long videos with objects that appear and disappear. It also allows the model to maintain high accuracy across tests on diverse datasets.

6) Mask decoder

This component is responsible for performing segmentation based on input features from images and prompt within the SAM block. The inputs consist of features from the image encoder and prompts such as points, bounding boxes, or masks. They are encoded before by the prompt encoder. The mask decoder is designed to be lightweight yet effective in mapping image embeddings and prompt embeddings to mask outputs. The architecture of this component is inspired by Transformer-based segmentation models. The Mask Decoder is customized from the standard Transformer Decoder architecture, incorporating Multi-Layer Perceptron (MLP) blocks and convolutional layers. It aims to ensure high computational efficiency while maintaining high segmentation accuracy. Fig. 4 shows the results of the Grounding DINO-SAM model in identifying and segmenting household objects.

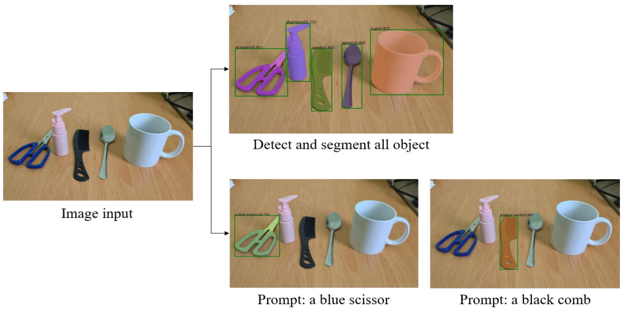

Fig. 4. The object detection and segmentation results of the Grounding DINO-SAM model with different prompts.

C. Backbone and Grasp Detection

This section presents the architecture for robot grasp location detection from RGB images that inspired by the work of Ainetter and Fraundorfer [11]. The model employs ResNet-101 [37] as the backbone with several modifications to suit the tasks of grasp detection. Specifically, ResNet-101 is combined with a Feature Pyramid Network (FPN) [38]. In addition, Synchronized Inplace Activated Batch Normalization [39] and LeakyReLU are used to replace Batch Normalization and ReLU in the original ResNet-101 as shown in Fig. 5. As a result, image features are extracted at multiple resolutions. These features are then passed to the Region Proposal Network [40], which identifies potential regions in the image that may contain grasp points. Then, they are processed by the grasp detection head. A model predicts grasp candidates, including the location and orientation of each grasp point. The input image for this process is a cropped image of the object based on the bounding box determined by the Grounding DINO-SAM model.

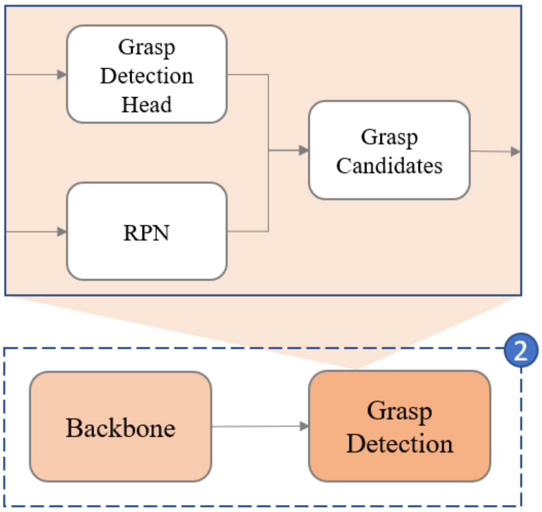

Fig. 5. Detailed architecture of grasp detection.

1) Region proposal network

RPN is a fully convolutional network. The input to the RPN comes from the backbone module, which provides image embeddings. The RPN is responsible for generating rectangular region proposals based on image features. These regions are represented by center coordinates (x, y) and dimensions (w, h). However, due to its architectural characteristics, the output does not include information regarding the rotation angle θ. The RPN plays a critical role in reducing the number of regions to be examined, thereby improving computational efficiency in region detection tasks.

2) Grasp detection head

This component is responsible for predicting the grasp candidates. Each grasp point is determined based on the region proposals that have been computed beforehand. These region proposals are fed into the grasp detection head. The ROIAlign [41] is used to extract feature maps with a spatial resolution of 14×14, corresponding to the region proposals. Next, each feature map undergoes average pooling with a kernel size of 2. The data is then passed through two fully connected layers, each containing 1024 neurons. Both fully connected layers apply Synchronized Inplace Activated Batch Normalization and Leaky ReLU activation. The data is then sent to two subnetworks, which include the prediction of grasp orientation and the prediction of the bounding box. The grasp orientation prediction subnetwork predicts the orientation of the grasp point, classified into 18 classes. Each class corresponds to an even division of orientations. The second subnetwork predicts the bounding box parameters for each grasp point, including position (*x, y*) and size (*h, w*). These parameters are used to determine the grasp candidates.

3) *Loss function cho grasp detection*

The loss function L_{grasp} is defined in Eq. (5), where L_{cls} and L_{reg} represent the loss functions of the Region Proposal Network. L_{box} and L_{rot} correspond to the loss functions for the bounding box location and the rotation angle of the grasp, respectively. L_{cls} and L_{reg} are used that similar to the study in [40]. L_{rot} is defined according to Eq. (5), where K = (K_{-}union K_{+}). K_{+} is the set of valid region proposals, and $K_{-}$is the invalid set generated using the RPN. The terms p_k^{cls} and p_k^{non} are defined as functions determining the likelihood of grasp candidate k being within the true class label and the likelihood of k being an invalid region. The term L_{box} is formulated and presented in Eq. (6). The term i represents the values of each element in [x, y, h, w]. The correction factors t_d and the L1 regularization function $smooth_{L1}$ are detailed in the study [42].

$$L_{grasp} = L_{cls} + L_{reg} + L_{box} + L_{rot} \tag{4}$$

$$L_{rot} = -\left(\frac{1}{|K|} \sum_{k \in K_{+}} \log\left(p_k^{cls}\right) + \frac{1}{|K|} \sum_{k \in K_{-}} \log\left(p_k^{non}\right) \right) \tag{5}$$

$$L_{box} = \sum_{i \in \{x,y,h,w\}} smooth_{L1}\left(i - t_d\right) \tag{6}$$

D. *Grasp Refinement Head*

The Grasp Refinement Head is a crucial component in the grasp detection model aimed at improving the accuracy of the grasp candidates. The input data consists of grasp candidates and semantic segmentation. Grasp candidates refer to the initial predicted grasp locations from the Grasp Detection block. Semantic segmentation is the object segmentation map, cropped according to the size (*H*, *W*) of the bounding box of object. The size of the semantic segmentation equals the input image size for both the backbone and Grasp Detection. Then, the grasp candidate parameters are fused and cropped with the semantic segmentation of object to determine the grasp candidates positions on the segmentation map. This fusion and cropping process allows the network to learn more detailed information about the grasp candidates based on the combination of geometric and semantic segmentation features. Subsequently, they are stacked together. The tensor of semantic segmentation and the tensor containing grasp location information are merged to create a tensor of size (*H*, *W*, 2). This step helps the model better understand the grasp candidates positions on the object. Finally, an MLP block is used to refine the grasp candidates. It improves the accuracy of grasp predictions. The result is a set of refined correction factors, including d_x, d_y, d_w, d_h, d_θ. The architecture was shown in Fig. 6.

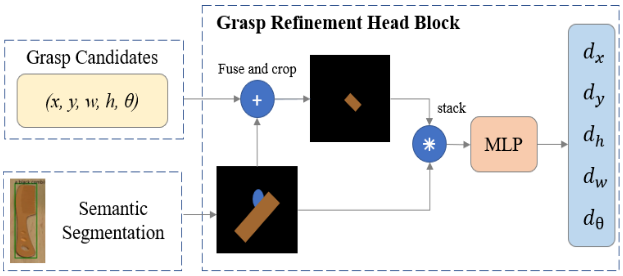

Fig. 6. Architecture of grasp refinement head.

Eq. (7) is used as the loss function for grasp refinement. The output of the grasp refinement head is denoted as d_i. The correction factors are denoted as t_i. The $smooth_{L1}$ function is used similarly to how it is applied in L_{box} in the Loss function cho grasp detection section.

$$L_{refinement} = \sum_{i \in \{x,y,h,w,\theta\}} smooth_{L1}\left(d_i - t_i\right) \tag{7}$$

The proposed model is composed of three components including Grounding DINO-SAM, Grasp Detection, and Grasp Refinement Head. However, this research only integrates and utilizes the pre-trained model for the Grounding DINO-SAM block. Therefore, it is no loss function for first component. The remaining two blocks are trained during the experiments, with the respective loss functions being L_{grasp} and $L_{refinement}$. To perform training for both tasks across these two blocks, a combined loss function is defined as in Eq. (8).
where, α and β are weights that determine the influence of each process on the value of $L_{general}$. In this research, the values of α and β are set to 1.

$$L_{general} = \alpha L_{grasp} + \beta L_{refinement} \tag{8}$$

IV. Experiments and Discussion

A. *Experiment Setup*

To evaluate the effectiveness of the proposed method, experiments were conducted on well-known grasping datasets including the OCID Grasp and Jacquard datasets. They are datasets widely used in the field of grasp detection. Additionally, the model was also tested on our HHI dataset to assess its applicability in human-robot interaction scenarios for grasping objects based on prompt. The proposed method was compared with previous approaches, all of which use RGB data as input. Exceptionally, the Interactive Grasp Evaluation section uses both RGB images and prompts.

In terms of training setup, the backbone is initialized with pre-trained weights from ImageNet. The parameters of the first two modules including conv1 and conv2 are frozen during the entire training process to enhance stability and efficiency. For the Jacquard dataset, the data is split with 95% allocated for training and 5% for testing. In the case of the OCID Grasp and HHI datasets, the data is divided with 80% for training and 20% for testing. All training and evaluation experiments are conducted on an NVIDIA RTX 3070 GPU. For the evaluation method, the Jaccard index is used to evaluate the performance of the methods in the experiments. The evaluation formula is given in Eq. (9). The terms g_C and g_T represent the parameters of the grasp candidates and the grasp ground truth, respectively. A predicted result is considered accurate if the angular deviation between the predicted value and the label is less than or equal to 30º and J(IoU) > 0.25.

$$J(IoU) = J(g_C, g_T) = \frac{|g_C \text{ I } g_T|}{|g_C \text{ U } g_T|} \tag{9}$$

B. Dataset

1) Jacquard

The Jacquard dataset is a dataset specifically designed to evaluate grasp detection and semantic segmentation in robotics. This dataset contains 54,000 RGB-D images generated from 11,000 unique objects. Each image includes ground truth labels, which consist of grasp locations and orientations for the robot and ground truth semantic segmentation. This dataset enables the evaluation of multitask learning approaches, including both grasp detection and semantic segmentation.

2) OCID grasp

The OCID Grasp dataset is an extended version of the OCID dataset. Initially, the OCID dataset contained objects, contexts, varying distances between the camera, viewpoints, and diverse lighting conditions. The primary purpose of this dataset was to evaluate semantic segmentation methods in increasingly complex scenes. Later, the OCID Grasp dataset was expanded by adding manually annotated labels of valid grasp candidates for each graspable object in the images. The OCID Grasp dataset includes 1,763 images. It contains over 11,400 object segmentation masks and more than 75,000 manually annotated grasp candidates. Objects in the dataset is categorized into 31 different classes.

C. Grasp Accuracy Evaluation

1) Evaluation on Jacquard dataset

Images in the dataset is normalized to a size of 512×512 before being input into the model. This resizing helps reduce computational load during training while retaining sufficient information from the images. The proposed model is trained using the Stochastic Gradient Descent method. Specific parameters include a learning rate of α = 0.02, regularization with L2 = 0.0001, and momentum of m = 0.9. The training process is performed solely on two blocks including grasp detection and grasp refinement heads. The Grounding DINO-SAM block is frozen during training. The average Frame Per Second (FPS) achieved was 14 with the RTX 3070 GPU hardware.

The proposed model demonstrates exceptional performance on the Jacquard dataset, achieving a grasp accuracy of 93.12%, as shown in Tables I–III, outperforming competing methods such as Zhang ROI-GD (90.4%), Song ResNet-101 (91.5%), and Kumra GR-ConvNet (91.8%). This high accuracy highlights its effectiveness in detecting precise grasp locations for robotic applications. When evaluated with varying angle thresholds in Table II, the model sustains robust performance at a 30° threshold, showcasing its reliability. Even as the angle threshold decreases, the accuracy experiences only a slight decline, yet it consistently surpasses other models. This performance is attributed to the integration of vision-language models and grasp refinement, enhancing object detection and grasp prediction. The results affirm the model capability to excel in single-object scenarios typical of the Jacquard dataset. In Table III, when applying different IoU criteria, our model continues to lead with 91.37% at a 30% threshold, though it performs slightly worse at a 35% threshold. These results indicate that the proposed model performs well in accurately detecting grasp locations under standard evaluation criteria (30° and 25% J(IoU)). However, when the evaluation conditions become more stringent, proposed model has yet to outperform previous studies. Specifically, when the angle threshold is set to 10°, the grasp accuracy of the proposed model is lower than 0.65% the results of the previous study, Det_Seg_Refine. Furthermore, to enhance the reliability of the results, a 95% confidence interval was computed based on the outcomes of 5-fold cross-validation, with a lower bound of 92.86% and an upper bound of 93.38%.

TABLE I. COMPARISON OF GRASP ACCURACY RESULTS ON THE JACQUARD DATASET

Method	Input	Grasp Accuracy (%)
Zhang, ROI-GD [43]	RGB	90.4
Song, Resnet-101 [44]	RGB	91.5
Kumra, GR-ConvNet [18]	RGB	91.8
Depierre [45]	RGB	85.7
Det_Seg_Refine [11]	RGB	92.95
Ours	RGB	**93.12**

TABLE II. COMPARISON OF GRASP ACCURACY RESULTS ON THE JACQUARD DATASET WITH DIFFERENT ANGLE THRESHOLDS

Method	30º	25º	20º	15º	10º
Zhou [20]	81.95	81.76	81.27	80.23	77.79
Depierre [45]	85.74	85.55	85.01	83.65	80.82
Det_Seg_Refine [11]	92.95	92.88	**92.42**	**91.52**	**88.92**
Ours	**93.12**	**92.90**	92.39	91.46	88.27

TABLE III. COMPARISON OF GRASP ACCURACY RESULTS ON THE JACQUARD DATASET WITH DIFFERENT J(IOU) THRESHOLDS

Method	25%	30%	35%
Zhou [20]	81.95	78.26	74.33
Depierre [45]	85.74	82.58	78.71
Song [44]	91.5	89.7	87.3
Det_Seg_Refine [11]	92.95	91.33	**88.96**
Ours	**93.12**	**91.37**	88.45

2) *Evaluation on OCID grasp dataset*

In the experiment with the OCID Grasp dataset, proposed model was trained with a learning rate of 0.03. The other parameters were similar to those used during training on the Jacquard dataset. The evaluation metric, Jaccard index, was calculated for each object class that could be grasped in the scene, along with the Intersection over Union (IoU) for segmentation in each object class. The accuracy of grasp candidates was evaluated using this metric, which required the center of the grasp candidate to be located within the predicted segmentation mask of the corresponding object class. Experimental results in Table IV demonstrate that the grasp accuracy on the OCID Grasp dataset outperforms the previous Det_Seg_Refine study. Overall, the experimental results evaluating grasp accuracy across different datasets show that proposed method does not perform as effectively on single-object datasets like Jacquard. Instead, in more complex contexts with multiple objects, such as the OCID Grasp dataset, the proposed model achived better results due to its enhanced segmentation capability. Additionally, similar to Jaquard dataset, a 95% confidence interval was also calculated using the results of 5-fold cross-validation on the OCID Grasp dataset, yielding a lower bound of 89.78 and an upper bound of 90.84.

TABLE IV. COMPARISON OF GRASP ACCURACY RESULTS ON THE OCID GRASP DATASET

Method	Grasp Accuracy (%)
Det_Seg_Refine	89.02
Ours	**90.31**

D. *Interactive Grasp Evaluation*

One of main objective in this study is to develop a grasping model capable of interacting with humans through prompts. Therefore, this experiment is implemented and evaluated based on the interaction of model with users during the grasping of household objects. HHI dataset is used to assess the results for two consecutive tasks, including object detection and grasp detection. The HHI contains 427 RGB images of household items, with the number of segmented objects matching the number of prompts, which is 2.6k. Each image in dataset includes multiple instances in same class that differ in color, size, and position. The grasping positions are manually labeled for each image-prompt pair. The labels in the dataset are manually annotated, similar to the OCID Grasp dataset. The model is fine-tuned based on a pre-trained model with the OCID Grasp dataset. This setup helps to evaluate the model ability to understand text prompts and correctly select and grasp the intended object from complex scenes with similar items. The effectiveness of the Interactive Grasp experiments was also assessed using the same method as the Jacquard and OCID Grasp datasets, relying on the Jaccard index and angular deviation to determine grasp accuracy. The evaluation metrics used are consistent with those in the evaluation on OCID Grasp dataset section. The results achieve an accuracy of 82.26% with lower and upper bound of 95% confidence interval are 81.78% and 82.74%, respectively.

The results demonstrate that the proposed model addresses the issue of previous grasp detection about inability of model to understand textual data and interact with users. Fig. 7 shows experimental results, with red rectangular boxes representing the predicted grasping areas by the model. The first pair of images illustrates the model ability to detect object by color. The second and third pairs of images highlight the contextual understanding ability, enabling the robot to correctly identify the object to grasp in scenarios with multiple similar objects. However, since the current model has the Grounding DINO-SAM component frozen, grasp location detection for complex images or prompts remains a challenge. In the other hand, the experimental results also showcase the significant potential of integrating Grounding DINO-SAM into grasp detection to enhance human-robot interaction.

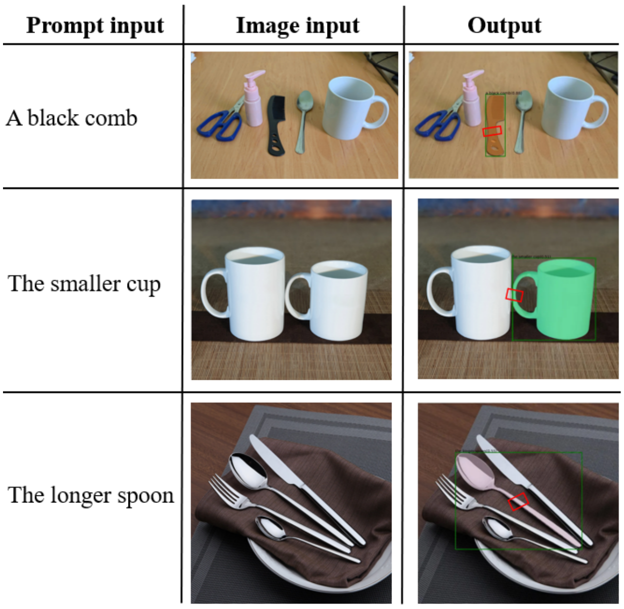

Fig. 7. Experimental results on HHI dataset with image and prompt input.

E. *Ablation Study*

To assess the contributions of individual components in the proposed model, an ablation study was conducted by systematically removing key elements and evaluating the impact on grasp detection performance. Specifically, we removed the segmentation mask generation from the Grounding DINO-SAM module and, consequently, the Grasp Refinement Head, as the latter relies on the segmentation masks for refining grasp candidates. This ablation allows us to isolate the effects of semantic segmentation and grasp refinement on the overall performance of the model. The experiments were conducted on three datasets including Jacquard, OCID Grasp, and HHI, with the same training and evaluation protocols as previous described. The results of the ablation study are summarized in Table V, comparing the grasp accuracy of the full model against the ablated model across the three datasets.

TABLE V. GRASP ACCURACY (%) OF ABLATION STUDY ON JACQUARD, OCID GRASP, AND HHI DATASETS

Method	Jacquard	OCID Grasp	HHI
Full Model (Ours)	93.12	90.31	82.26
Ablated Model	90.74	87.42	79.75

The results showed that the segmentation masks generated by the Grounding DINO-SAM module and the subsequent refinement by the Grasp Refinement Head are integral to the model's performance across all datasets. The consistent drop in grasp accuracy approximately 2.38% to 2.89% underscoring their importance in enhancing both object detection and grasp localization. In simpler datasets like Jacquard, the impact is less significant, as bounding boxes alone provide sufficient information for reasonable grasp predictions. However, in more challenging scenarios, such as those in OCID Grasp and HHI, the absence of segmentation masks leads to a greater loss of contextual understanding, resulting in less accurate grasp candidates. The Grasp Refinement Head, by leveraging semantic segmentation, further fine-tunes these candidates, ensuring higher precision and robustness.

Based on the outcomes of all prior experiments, the integration of vision-language models within the proposed grasp detection framework introduces notable advancements in both grasp accuracy and human-robot interaction. By leveraging the complementary strengths of visual and linguistic modalities, the model is capable of enhancing object recognition in complex scenes that involve overlapping items and unfamiliar objects. The utilization of combining descriptive language prompts with RGB image enables the system to semantically interpret and localize target objects, thereby overcoming limitations commonly associated with vision-only approaches. This capability is particularly beneficial in scenarios where objects have not been encountered during training or multiple items belong to the same class but differ in color, size, or other attributes. Furthermore, the vision-language integration significantly contributes to improved interaction between humans and robots, allowing users to guide robotic actions through intuitive natural language instructions. Evaluations confirm the effectiveness of this approach, with the proposed model achieving a high grasping accuracy on the Jacquard dataset, surpassing previous methods.

V. CONCLUSION

This paper proposes an integrated model combining computer vision and language models to enhance object recognition and grasping capabilities in complex environments, guided by user prompts. The proposed approach leverages the ability of visual-language models to identify and generate masks for objects based on user prompts, thereby improving accuracy. This integration not only enhances the object recognition capabilities of robot but also optimizes the grasping process in complex scenarios. Additionally, the model allows the robot to enhance user interaction through various prompts. The experimental results show that the proposed model achieves a grasping accuracy of 93.12% on the Jacquard dataset. It surpasses previous methods, particularly in contexts with multiple complex objects and diverse shapes. Furthermore, the model demonstrates its ability to understand complex contexts and interact effectively with users in Interactive Grasp experiments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Nguyen Khac Toan is the first author, responsible for the methodology, writing the original draft, programing, and validation of the research. Nguyen Truong Thinh is the corresponding author, contributing to the writing–review and editing, methodology, and project administration. All authors have reviewed and approved the final version of the paper.

FUNDING

This research is funded by University of Economics Ho Chi Minh City—UEH University, Vietnam.

REFERENCES

[1] N. M. Trieu and N. T. Thinh, "A comprehensive review: Interaction of appearance and behavior, artificial skin, and humanoid robot," *Journal of Robotics*, 2023. https://doi.org/10.1155/2023/5589845

[2] R. Qin *et al.*, "RGB-D grasp detection via depth guided learning with cross-modal attention," in *Proc. the 2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, United Kingdom, 2023, pp. 8003–8009. https://doi.org/10.1109/ICRA48891.2023.10161319

[3] S. Zhang and M. Xie, "MIPANet: Optimizing RGB-D semantic segmentation through multi-modal interaction and pooling attention," *Frontiers in Physics*, vol. 12, 2024. https://doi.org/10.3389/fphy.2024.1411559

[4] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, pp. 705–724, 2015. https://doi.org/10.1177/0278364914549607

[5] J. Achiam *et al.*, "GPT-4 technical report," arXiv preprint, arXiv:2303.08774, March 2023.

[6] A. Chowdhery *et al.*, "PaLM: Scaling language modeling with Pathways," *Journal of Machine Learning Research*, vol. 24, pp. 1–113, 2023. https://dl.acm.org/doi/10.5555/3648699.3648939

[7] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," arXiv preprint, arXiv:2302.13971, 2023.

[8] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. https://dl.acm.org/doi/10.5555/3295222.3295349

[9] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[10] J. Li *et al.*, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. the International Conference on Machine Learning (ICML)*, 2022, pp. 12888–12900.

[11] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB," in *Proc. the 2021 IEEE International Conference on Robotics and Automation (ICRA),* Xi'an, China, 2021, pp. 13452–13458.

[12] A. Depierre *et al.*, "Jacquard: A large scale dataset for robotic grasp detection," in *Proc. the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018, pp. 3511–3516. https://doi.org/10.1109/IROS.2018.8593950

[13] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *Proc. the 2000 IEEE International Conference on Robotics and Automation (ICRA)*, San Francisco, CA, USA, 2000, pp. 348–353. https://doi.org/10.1109/ROBOT.2000.844081

[14] F.-J. Chu *et al.*, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, Oct. 2018. http://dx.doi.org/10.1109/LRA.2018.2852777

[15] L. O. Chua, *CNN: A Paradigm for Complexity*, Singapore: World Scientific, 1998. https://doi.org/10.1142/3801

[16] D. Morrison *et al.*, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," arXiv preprint*,* arXiv:1804.05172, 2018.

[17] D. Morrison *et al.*, "Learning robust, real-time, reactive robotic grasping," *The International Journal of Robotics Research*, vol. 39, no. 2–3, pp. 183–201, Feb. 2020. https://doi.org/10.1177/0278364919859066

[18] S. Kumra *et al.*, "Antipodal robotic grasping using generative residual convolutional neural network," in *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, 2020, pp. 9626–9633. https://doi.org/10.1109/IROS45743.2020.9340777

[19] S. Yu *et al.*, "SE-ResUNet: A novel robotic grasp detection method," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5238–5245, Apr. 2022. https://doi.org/10.1109/LRA.2022.3145064

[20] X. Zhou *et al.*, "Fully convolutional grasp detection network with oriented anchor box," in *Proc. the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018, pp. 7223–7230. http://dx.doi.org/10.1109/IROS.2018.8594116

[21] D. Park, Y. Seo, and S. Y. Chun, "Real-time, highly accurate robotic grasp detection using fully convolutional neural network with rotation ensemble module," in *Proc. the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France, 2020, pp. 9397–9403. https://doi.org/10.1109/ICRA40945.2020.9197002

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th International Conference on Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, Munich, Germany, Oct. 5–9, 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[23] U. Asif, J. Tang, and S. Harrer, "GraspNet: An efficient convolutional neural network for real-time grasp detection for low-powered devices," in *Proc. the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, 2018, pp. 4875–4882. https://doi.org/10.24963/ijcai.2018/677

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017. https://doi.org/10.1109/TPAMI.2016.2577031

[25] A. Kirillov *et al.*, "Segment anything," in *Proc. the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 4015–4026. https://doi.org/10.1109/ICCV51070.2023.00371

[26] N. Ravi *et al.*, "SAM 2: Segment anything in images and videos," arXiv Preprint*,* arXiv:2408.00714, 2024.

[27] R. Araki, T. Onishi, and K. Yamamoto, "MT-DSSD: Deconvolutional single shot detector using multi-task learning for object detection, segmentation, and grasping detection," in *Proc. the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France, 2020, pp. 10487–10493. https://doi.org/10.1109/ICRA40945.2020.9197251

[28] D. Garg, S. Vaidyanath, K. Kim, J. Song, and S. Ermon, "LISA: Learning interpretable skill abstractions from language," in *Proc. 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, vol. 35, 2022, pp. 21711–21724.

[29] T. Nguyen, M. N. Vu, B. Huang, and T. V. Vo, "Language-conditioned affordance-pose detection in 3D point clouds," in *Proc. the 2024 IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, 2024, pp. 3071–3078. https://doi.org/10.1109/ICRA57147.2024.10610008

[30] C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, "GraspGPT: Leveraging semantic knowledge from a large language model for task-oriented grasping," *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 7553–7560, Jul. 2023. https://doi.org/10.1109/LRA.2023.3320012

[31] S. Liu *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," arXiv preprint*,* arXiv:2303.05499, 2023.

[32] T. Ren *et al.*, "Grounded SAM: Assembling open-world models for diverse visual tasks," arXiv preprint*,* arXiv:2401.14159, 2024.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[34] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

[35] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 16000–16009. https://doi.org/10.1109/CVPR52688.2022.01553

[36] X. Kong and X. Zhang, "Understanding masked image modeling via learning occlusion invariant feature," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 6241–6251. http://dx.doi.org/10.1109/CVPR52729.2023.00604

[37] B. Koonce, "ResNet 50," in *Convolutional Neural Networks with Swift for TensorFlow: Image Recognition and Dataset Categorization*, 1st ed., Apress, 2021, pp. 63–72. http://dx.doi.org/10.1007/978-1-4842-6168-2_6

[38] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2117–2125. https://doi.org/10.1109/CVPR.2017.106

[39] S. R. Bulò, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of DNNs," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 5639–5647. https://doi.org/10.1109/CVPR.2018.00591

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2961–2969. https://doi.org/10.1109/ICCV.2017.322

[41] T. Gong *et al.*, "Temporal ROI align for video object recognition," in *Proc. the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1442–1450. https://doi.org/10.1609/aaai.v35i2.16234

[42] C.-Y. Fu, M. Shvets, and A. C. Berg, "RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free," arXiv preprint*,* arXiv:1901.03353, 2019.

[43] H. Zhang *et al.*, "ROI-based robotic grasp detection for object overlapping scenes," in *Proc. the 2019 IEEE/RSJ International*

Conference on Intelligent Robots and Systems (IROS), Macau, China, 2019, pp. 4768–4775. https://doi.org/10.1109/IROS40897.2019.8967869

[44] Y. Song *et al.*, "A novel robotic grasp detection method based on region proposal networks," *Robotics and Computer-Integrated Manufacturing*, vol. 65, 101963, 2020. https://doi.org/10.1016/j.rcim.2020.101963

[45] A. Depierre *et al.*, "Scoring graspability based on grasp regression for better grasp prediction," in *Proc. the 2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China, 2021, pp. 4370–4376. https://doi.org/10.1109/ICRA48506.2021.9561198