Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Natural Language Visual Grounding via Multimodal Learning

Dissertation

with the aim of achieving the degree of

Doctor rerum naturalium (Dr. rer. nat.) at the

Faculty of Mathematics, Informatics and Natural Sciences,

Department of Informatics,

Universität Hamburg

Jinpeng Mi

Hamburg 2020

Submitted on:

November 11, 2019

Date of oral defence:

January 20, 2020

The following evaluators recommend the admission of the dissertation:

Prof. Dr. Jianwei Zhang (advisor)

Department of Informatics,

Universität Hamburg, Germany

Prof. Dr. Stefan Wermter (reviewer)

Department of Informatics,

Universität Hamburg, Germany

Prof. Dr. Chris Biemann (chair)

Department of Informatics,

Universität Hamburg, Germany

To my loving grandparents, parents, and my other family members.

Abstract

Natural language provides an intuitive and effective interaction interface between human beings and intelligent agents. Currently, multiple approaches have been proposed to address natural language visual grounding. However, most of the existing approaches alleviate the ambiguity of natural language queries and achieve target objects grounding by drawing support from auxiliary information, such as dialogues between human users, and gestures. While the auxiliary information-based systems usually make the natural language grounding cumbersome and time-consuming.

This thesis aims to study and exploit multimodal learning approaches for natural language visual grounding. Inspired by the pattern of human beings understanding and grounding target objects according to given natural language queries, we propose different architectures to address natural language visual grounding.

First, we propose a semantic-aware network for referring expression comprehension which aims to locate the most relevant objects in images given natural referring expressions. The proposed referring expression comprehension network excavates the visual semantics in images via a visual semantic-aware network, exploits the rich linguistic contexts in referring expressions by a language attention network, and locates target objects by integrating the outputs of the visual semantic-aware network and the language attention network. Moreover, we conduct extensive experiments on three public datasets to validate the performance of the presented network.

Second, we present a Generative Adversarial Networks-based network to generate diverse and natural referring expressions. Referring expression generation mimics the role of a speaker to generate referring expressions for each detected region within images. For this task, we aim to improve the diversity and naturalness of expressions without sacrificing semantic validity. To this end, we propose a generator to generate expressions and exploit a discriminator to classify whether the generated descriptions are real or fake. We evaluate the performance of the proposed generation network via multiple evaluation metrics.

Third, inspired by the psychology term "affordance" and its applications in Human-Robot interaction, we draw support from object affordance to ground intention-related natural language queries. Formally, we first present an attention-based multi-visual features fusion network to recognize object affordances. The proposed network fuses deep visual features extracted from a pretrained CNN model with deep texture features encoded by a deep texture encoding network via an attention-based mechanism. We train and validate the performance of the object affordance detection network on a self-built dataset.

Moreover, we propose three natural language visual grounding architectures, which are based on referring expression comprehension, referring expression generation, and object affordance detection, respectively. We combine the referring expression comprehension and referring expression generation models with scene graph parsing to achieve complicated and unconstrained natural language queries grounding. Additionally, we integrate the object affordance detection network with an intention semantic extraction module and a target grounding module to ground intention-related natural language queries.

Finally, we implement extensive experiments to validate the effectiveness of the presented natural language visual grounding architectures. We also integrate with an online speech recognizer to complete target object grounding and manipulation experiments on a PR2 robot given spoken natural language commands.

Zusammenfassung

Natürliche Sprache bietet eine intuitive und effektive Interaktionsschnittstelle zwischen Mensch und Roboter. Eines der Kernprobleme dabei ist das Symbol Grounding in visueller Wahrnehmung, also die Zuordnung und Lokalisierung von Objekten in Bildern. Zwar gibt es bereits mehrere Ansätze, die das Symbol Grounding mit natürlicher Sprache für die Mensch-Roboter-Interaktion behandeln, aber die meisten dieser Arbeiten verwenden Dialogsysteme, um die Mehrdeutigkeit natürlicher Sprache zu verringern und die Zuordnung der Zielobjekte zu erreichen, was die Interaktionen umständlich und zeitaufwändig macht.

Das Ziel dieser Dissertation ist es, multimodale Lernverfahren für das visuelle Symbol Grounding natürlicher Sprache zu studieren und zu nutzen. Ausgehend von der Art und Weise, wie Menschen Objekte durch die Anfragen anderer verstehen und lokalisieren, werden in dieser Arbeit zunächst drei verschiedene Architekturen entwickelt und analysiert.

Erstens führen wir ein neuartiges "Semantic-aware deep neural network" für das Verständnis von Referenzausdrücken ein. Die Aufgabe ist, anhand eines Ausdrucks in natürlicher Sprache das jeweils relevanteste Objekt in einem Bild zu lokalisieren. Die vorgeschlagene Architektur legt den visuellen Bildinhalt über ein visuell-semantisches tiefes Netzwerk frei und verarbeitet den reichen sprachlichen Kontext der Referenzausdrücke mit einem Sprach-Aufmerksamkeitsnetzwerk. Wir führen Experimente an drei bekannten öffentlichen Datensätzen durch, um das vorgeschlagene Netzwerk zu validieren.

Zweitens stellen wir ein Generative Adversarial Networks (GANs) für die Erzeugung von Referenzausdrücken vor. Das Netzwerk imitiert die Rolle eines Sprechers, um Referenzausdrücke für erkannte Regionen in einem Bild zu generieren. Das vorgeschlagene System hat dabei zum Ziel, die Vielfalt und Natürlichkeit der erzeugten Referenzausdrücke gegenüber bekannten Methoden zu verbessern. Wir bewerten die Leistung des eingeführten Netzwerks mit mehreren Auswertungsmetriken.

Drittens führen wir ein "Multi-visual feature fusion network" für die Erkennung von Objekt-Anwendungscharakteren (Affordances) ein. Das vorgeschlagene Netzwerk kombiniert mehrere tiefe neuronale Netzwerke, um die Affordances von Objekten in RGB-Bildern zu erlernen. Die Architektur nutzt ein Aufmerksamkeitsnetzwerk, um visuelle Merkmale, die von einem vor-trainierten CNN-Modell extrahiert wurden, mit Textur-Merkmalen zu verschmelzen, die durch ein separates tiefes Netzwerk kodiert wurden. Wir testen die Leistung des Netzwerks mit einem selbst erstellten Datensatz.

Darüber hinaus schlagen wir Architekturen für interaktive Verarbeitung natürlicher Sprache vor, die jeweils auf dem Verständnis der Referenzausdrücke, der Erzeugung dieser Ausdrücke, und der Erkennung von Objekt-Affordances basieren. Dazu kombinieren wir diese Verfahren mit Szenengraph-Parsing, um ein ausgereiftes und uneingeschränktes interaktives visuelles Symbol Grounding natürlicher Sprache zu erreichen. Zusätzlich integrieren wir das Framework zur Affordance-Erkennung mit einem semantischen Extraktionsmodul, um absichtsbezogene Abfragen in natürlicher Sprache zu verarbeiten.

Schließlich führen wir umfangreiche Experimente durch, um die Wirksamkeit der vorgestellten visuellen Symbol Grounding Architekturen für natürliche Sprache zu validieren. Außerdem präsentieren wir Manipulationsexperimente mit Befehlen in natürlicher Sprache für einen PR2-Roboter und verschiedenen Zielobjekten.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Human beings live in a multimodal environment where natural language and vision are the dominant ways for communication and perception in our daily life. Humans often use natural language to indicate a specific person or object. For instance, "the man next to the car", or "the remote controller on the table". The listener can identify referred targets according to given natural language queries. Naturally, we would like to develop intelligent agents with the ability to communicate and perceive their working scenarios as humans do, and locate referred target objects within working scenarios. Natural language processing, computer vision, and the interplay between them are involved in the task to ground natural language queries in visual scenarios.

Natural Language and vision are two dominant channels to represent and exchange information in our daily life. In recent years, how to bridge the two domains has been attracting considerable research attention in the area of computer vision [67, 79, 147, 16], natural language processing [65, 112], and multimedia [19, 47].

One motivation of this thesis is that, grounding natural language in visual scenes provides a natural communication channel between humans, physical environments, and intelligent systems. We often refer to objects in the environment when we have a pragmatic interaction with others, and we have the ability to

comprehend and ground natural language queries in a wide range of practical applications. We also would like to endow intelligent agents the ability to comprehend and ground natural language instructions. For instance, in natural language-based human-robot interaction (HRI), robotic systems need to understand natural language instructions to locate the referred objects in their working scenarios. The ability to understand natural language commands prompts the robotic platforms to conduct natural language commands such as "pick up the red cup on the table", "pass me the remote controller near the TV", etc.

Another crucial motivation is the potential applications of natural language visual grounding. Natural language visual grounding can be widely used but not limited in human-computer interaction, robotics, and visual chatbot. One of the most representative example is the applications in robotics. Robots becoming omnipresent in varied human environments, such as factories, hospitals, and homes, the demand for natural and effective HRI has become urgent. Taking advantage of recent advances in machine perception, natural language visual grounding, and object manipulations, robots can draw support from these prerequisites to play more critical roles in human environments and perform diverse and dynamic tasks.

Natural language visual grounding requires a comprehensive understanding of natural language queries and visual scenarios, and the pivotal issue is to locate the referred objects in working scenarios according to the given queries. In order to ground target objects in different scenarios, intelligent agents have to deal with multiple challenging tasks, such as object detection, natural language understanding, and multimodal data fusion. However, nearly all of the proposed approaches do not consider the inherent ambiguity of natural language, or alleviate the ambiguity via dialogue systems. While the dialogue systems usually make the interaction cumbersome and time-consuming.

Natural language is the most straightforward and spontaneous medium in our daily communications with each other. As a special case of natural language queries, referring expressions depict objects within an image or a living environment from multiple perspectives, such as color, size, location, and the spatial re-

lations between their neighbor objects. Moreover, referring expressions are sufficiently easy for humans to locate the target objects during communication with others.

Within the realm of referring expressions, there are two related tasks, i.e., referring expression comprehension and referring expression generation. Referring expression comprehension plays the role of a listener to locate target objects within images given referring expressions, while referring expression generation mimics the role of a speaker to generate referring expressions for each detected object within images. Motivated by the role of referring expressions, in this thesis, we propose two architectures, which are based on referring expression comprehension and referring expression generation, to ground natural language queries. Moreover, we integrate scene graph parsing with referring expression compression and referring expression generation to ground complicated natural language queries.

Referring expression-based approaches can ground explicit natural language queries, such as "the left red apple on the table". While the target objects embedded in intention-related natural language commands cannot be located via referring expression-based frameworks, e.g. "I am thirsty, I want to drink some water." In order to ground the intention-related natural language queries, we draw support from a psychological term "affordance" that represents the association between the properties of an object and the capabilities of the object could possibly be used [94]. Inspired by the role of "affordance" and its applications in HRI, we introduce an object affordance detection-based framework to ground intention-related natural language queries.

In this thesis, we aim to achieve natural language grounding in a manner which is akin to end-to-end pattern and does not draw support from auxiliary information from human users. To this end, we propose three different architectures that are based on referring expression comprehension, referring expression generation, and object affordance detection, respectively.

1.2 Research Questions

Natural language visual grounding aims to understand the natural language queries and locate target object in images. Natural language grounding is a fundamental building block for many high-level tasks, such as image retrieval [15], video question answering [40], and natural language-based HRI [115], [44].

In real applications, natural language queries could be very complex and ambiguous, and working scenarios are complicated scenes and even challenging to analyze. In order to alleviate the ambiguity of natural language, some work employs dialogue systems [115, 44, 1] to locate target objects in their working scenarios, while the dialogue systems entail time cost and cumbersome interaction.

In this thesis, we exploit approaches to achieve natural language visual grounding without auxiliary information, such as dialogues between human users and intelligent agents, gestures, etc. Therefore, three critical issues are:

- how to disambiguate the natural language and exploit the rich linguistic context of natural language queries,
- how to excavate semantics embedded in visual images,
- how to build the mapping between natural language queries and visual regions to locate target objects given natural language queries,
- how to achieve natural language visual grounding without auxiliary information
- how to ground complicated and intention-related natural language queries

These questions will be addressed in this thesis one by one with the objective to achieve natural language grounding via joint learning visual features and language representations.

1.3 Novelty and Contribution

This study proposes approaches, experimental setups, and results for natural language grounding. The major contributions to the natural language grounding can be summarized in the following:

- **Referring expression comprehension via semantic-aware network**. Nearly all of the existing approaches only perform fine-grid spatial attention on extracted visual features to identify the most relevant objects, or resort to holistic associations between the referring expressions and the visual features. In this work, we propose a semantic-aware network for referring expression comprehension. The proposed semantic-aware network is composed of a visual semantic-aware network, a language attention network and a target localization module. The visual semantic-aware network excavates the visual semantics of extracted deep feature by fully utilizing the characteristics of the deep features, and the language attention network exploits the rich linguistic context of referring expressions and learns to assign different weights for each word in expressions. Moreover, the proposed referring expression comprehension network acquires competitive results on three public datasets in referring expressions.

- **Referring expression generation via adversarial training**. The existing methods employ Encoder-and-Decoder paradigms to generate expressions for image regions. These models process visual features by Convolutional Neural Networks (CNNs) and generate sequence words via Long Short-Term Memory (LSTM), and training with the objective to maximize the conditional likelihood of the training samples via Maximum Likelihood Estimation (MLE). However, the expressions generated by the CNN-LSTM paradigms are easy for humans to distinguish from natural descriptions because of their diversity and naturalness. In contrast, we adopt Generative Adversarial Networks (GANs) to generate more diverse and natural referring expressions. The generated expressions better imitate the way humans depict image re-

gions without sacrificing the semantic validity.

- **Object affordance recognition via attention-based multi-visual features fusion.** We propose an attention-based multi-visual features fusion architecture to learn object affordances from RGB images. The presented architecture employs an attention network to fuse deep visual features extracted from a pretrained CNN model with deep texture features encoded by a deep texture encoding network. The attention network learns attention weights automatically through sparse representations of the multi-visual features. Moreover, the attention-based fusion network takes into account the interaction of the multi-visual features and preserves the complementary nature of the different features. Furthermore, we introduce a dataset to train and validate the proposed object affordance detection network. Experimental results show that the attention-based multi-visual features fusion network outperforms other fusion scheme and affordance detection networks.

- **Interactive natural language visual grounding.** In order to achieve natural language visual grounding without auxiliary information, we propose three natural language grounding architectures that are based on referring expression comprehension, referring expression generation, and object affordance detection, respectively. We combine the trained referring expression comprehension and referring expression generation models with scene graph parsing to ground complicated and unrestricted interactive natural language queries, and we also integrate the object affordance detection network with an intention semantic extraction module to ground intention-related natural language instructions.

1.4 Thesis Structure

This thesis is organized into five main sections, they are described as follows:

1. Introduction. This section describes the motivations of this study, the re-

search questions and the research methodologies. Moreover, it also introduces the novelties of this work.

2. Referring expression comprehension via semantic-aware network. This chapter proposes a semantic-aware network for referring expression comprehension and elaborates on the details of the presented network. Moreover, this section lists the extensive experiments implemented to validate the performance of the introduced referring expression comprehension network and the acquired results.

3. Referring expression generation via adversarial training. This section presents an adversarial training-based network to generate diverse and natural referring expressions. The presented approach aims to generate natural referring expressions which are adequately easy for humans to locate the referred objects within images without sacrificing the semantic validity of expressions.

4. Object affordance recognition via attention-based multi-visual features fusion. This chapter introduces an object affordance detection network via attention-based multi-visual features fusion, and proposes a self-built dataset to learn human-centered object affordances.

5. Interactive natural language visual grounding. This section presents the details of three different natural language visual grounding architectures based on the three above introduced models, i.e., referring expression comprehension, referring expression generation, and object affordance detection. This chapter also introduces spoken instruction visual grounding by integrating with an online speech recognizer, and robotic applications conducted on a PR2 platform via the spoken instruction grounding framework.

6. Conclusion. This chapter summarizes the key ideas, insights, and approaches described throughout the thesis. After analyzing the acquired results, this section also describes the limitations of the presented architectures and provides future research directions.

Chapter 2

Referring Expression Comprehension via Semantic-Aware Network

2.1 Introduction

Referring expressions describe objects from diverse aspects, such as color, size, location, and spatial relations between their neighboring objects. Within the realm of referring expressions, there are two related tasks, i.e., referring expression comprehension and referring expression generation. The referring expression comprehension imitates the role of a listener to locate target objects within images given referring expressions. The inverse task is the referring expression generation which mimics the role of a speaker to generate discriminative referring expressions for objects or regions within images.

Referring expression comprehension aims to locate the most relevant objects or regions within images according to given referring expressions, and it requires a comprehensive understanding of natural referring expressions and images to locate target objects. Compared to image captioning and visual question answering, referring expression comprehension is widely used in image retrieval [15], video

question answering [40], and natural language-based HRI [115], [44].

Existing work adopts multiple approaches to tackle with referring expression comprehension. Baseline work [148] directly compares the visual and location difference to locate the most relevant object, [84] and [51] regard target object grounding as image retrieve where selects the most relevant region according to the ranking scores of generated expressions for detected region proposals. [107] completes grounding by reconstructing referring expressions based on local deep features, [149] proposes a Speaker-Listener model to generate and comprehend expressions, and takes advantage of RL (Reinforcement Learning) to discriminate the expressions, [16] leverages on external knowledge acquired by a fixed category detectors to assess language consistency and visual consistency, [28] introduces an accumulated attention network which accumulates the attention in image, object and referring expression to realize visual grounding.

In terms of representations of image regions and natural language referring expressions, existing approaches for referring expression comprehension can be generalized into two categories: 1) representations un-enriched models, which directly extract deep features from a pretrained CNN to be the visual representations for detected image regions [148, 84, 51, 149, 50, 28, 156, 159]. 2) representations enriched models, which enhance the visual representations by adding external visual information for regions. For instance, [78] leverages external knowledge acquired by an attributes learning model to enrich the information of regions. [151] trains on the Visual Genome dataset [64] to generate diversified and discriminative proposals. [147] extracts deep features from two different convolutional layers to predict region attribute cues and the predicted attributes are adopted to be auxiliary information for the extracted region deep features.

Although the existing models achieve promising results, they neglect two critical issues: 1) the essence of deep features extracted from a pretrained CNN, i.e., the features are spatial, channel-wise, and multi-layer [152], [18]. The existing methods focus on the spatial characteristics and perform fine-grid spatial attention to locate the most relevant object, while the importance of channel-wise traits is over-

looked. For example, in the process of predicting objects, the channel-wise features are generated by the convolutional filters relevant to represent visual semantics of objects. Therefore, the inherent semantics information of channel-wise features can be adopted to enhance the visual cues of regions. 2) the different contributions of each word in expressions to identify the target object. Nearly all existing approaches resort to a holistic association between the sentence and region feature. For instance, in the expression "a blue bus between two other buses", the word "blue" should be the one with the highest weight to locate the target "bus".

To address the issues as mentioned above, we fully utilize spatial and channel-wise characteristics of region deep features, and take into account the textual semantics of referring expressions for referring expression comprehension. Specifically, we propose a semantics-aware network that is composed of a visual semantic-aware network and a language attention network as illustrated in Figure 2.1. The crucial components of the visual semantic-aware network are the channel-wise attention and the region-based spatial attention. According to the characteristic of channel-wise features, the channel-wise attention can serve as a semantic attribute detector and is employed as an enrichment of visual representation for regions. For example, to predict a *horse*, the channel-wise attention pays more attention to channel-wise feature maps generated by the convolutional filters corresponding to represent visual semantics, such as furry texture and horse-like shape. While the region-based spatial attention mechanism attempts to focus on textual-semantic-related regions.

Moreover, the representation for words should be context-dependent, and the words in each expression contribute differently to locate the target object. To this end, we first extract word embeddings from a contextualized model, i.e., BERT[30], and then feed the extracted embeddings into a language attention network to acquire the different weights of each word in expressions. Additionally, the language attention network learns to parse expressions into three phrases that represent the target candidate, spatial location and relation between target and neighboring objects, respectively.

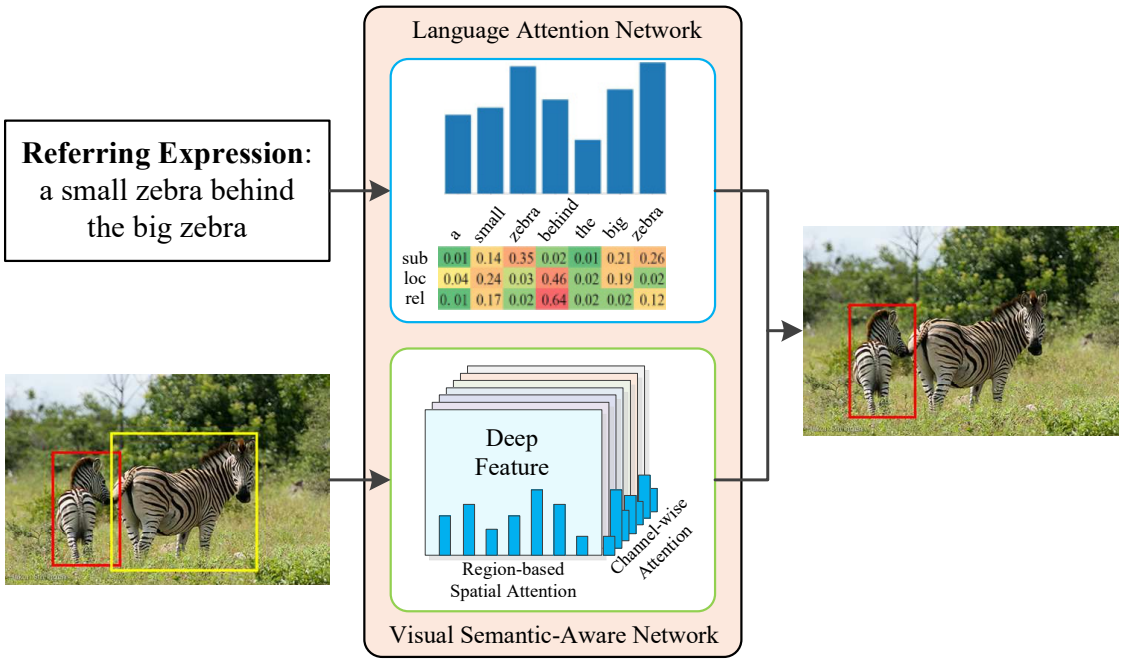

Figure 2.1: The illustration of the proposed semantic-aware network for referring expression comprehension. Given a referring expression, we first employ a language attention network to acquire different weights of each word in the expression and learn to parse the expression into three linguistic components. We perform channel-wise attention and region-based spatial attention, which are major constituents of the visual semantic-aware network, to generate keyword guided semantic-aware visual representation. We further combine the outputs of the two networks to locate target object.

In this work, we reformulate the proposed network for referring expression comprehension into three sub-modules: 1) a language attention network calculates different weights for each word in referring expressions and learns to parse expressions into three phrases; 2) a visual semantic-aware network incorporates the channel-wise attention and the region-based spatial attention to generate semantic-aware visual representation for regions under the guidance of attended words; 3) a target localization module coalesces the language attention network and the visual semantic-aware network to locate target objects.

We evaluate the proposed network on three popularized datasets: RefCOCO [148], RefCOCO+ [148], and RefCOCOg [84]. The introduced network acquires competitive results compared with the current state-of-the-art approaches. In summary, we propose a semantic-aware network, in which we exploit the rich linguistic context by a language attention network, and we excavate the inherent visual semantics in deep features via a visual semantic-aware network. We also conduct extensive experiments on the three public datasets to validate the performance of the introduced network.

2.2 Related Work

2.2.1 Referring Expression Comprehension

Different from visual relation detection [150], [63] and phrase grounding [17], [16], the pivotal point of referring expression comprehension is to locate the target objects according to given referring expressions, and usually several objects with the same category exist within images. Multiple methods have been proposed to tackle with referring expression comprehension. Baseline work [148] encodes the visual difference between objects of the same category within images, and through the comparison of visual difference to locate the target object. [84] combines CNNs with recurrent neural networks (RNNs) for joint understanding referring expressions. Through integrating spatial configurations and global scene-level contextual information into the network, [51] regards referring expression generation as object retrieval from the candidate objects. [149] proposes a Speaker-Listener-Reinforcer model to comprehend and generate referring expressions, and the reward-based reinforcer is used to guide the sampling of more discriminative expressions and further improve the grounding accuracy. [78] explores the role of visual attributes by incorporating them into referring expression comprehension. [69] adopts a visual context LSTM module and a sentence LSTM module to model bundled object context for referring expression. [156] presents a variational Bayesian framework

for referring expression comprehension, and the proposed model exploits the reciprocal relation between the referent object and context to reduce the context search space.

Attention mechanisms are first integrated with deep learning-based architectures in neural machine translation [6] and [139], and become an indispensable component in deep models to acquire superior results [3], [40]. Due to the excellent performance of attention mechanisms, it has also been utilized in referring expression comprehension [107, 50, 28, 159, 133]. [107] employs attention mechanism in referring expression comprehension by mapping phrase to image region and then through reconstructing a given phrase to realize referred object grounding. [50] parses the referring expressions into a triplet (subject, relationship, object) by an external language parser, and compute the weight of each part of parsed expressions with the soft attention mechanism. [28] introduces an accumulated attention network that accumulates the attention information in image, objects and referring expression to realize visual grounding. [159] argues that the image representation should be region-wise, and adopts a parallel attention network to ground target objects in variable length natural language descriptions, from short phrases query to long multi-round dialogs. [133] presents a graph attention that explicitly represents inter-object relations, and properties with flexibility and power impossible with competing approaches.

Weakly-supervised or unsupervised methods are also introduced for referring expression comprehension. [16] leverages to prompt weakly supervised visual grounding through drawing support from external knowledge which is acquired by parsing the referring expressions via a natural language processing (NLP) parser and retrieving the noun words, and then selecting the most probable class for each proposal. [143] develops a completely unsupervised framework for visual grounding by using hypothesis testing as a mechanism to link words to detected image concepts. [55] uses concept learning as a proxy task to obtain self-supervision, and the proxy task is utilized to decode the common concept present within each concept batch. [37] decomposes the referring expressions by an NLP parser and performs

compositional grounding progressively.

2.2.2 Visual and Textual Representations Augmented Models

The aforementioned models directly utilize the deep features extracted from a pretrained CNN as the visual representations for detected regions. Several studies draw support from external knowledge or different deep features to enrich the visual representations. [78] first learns attributes from objects and their paired descriptions, and then embeds the learned attributes with visual features into a common space. The target object is located via its Euclidean distance to the queried referring expression. Similar to [3], [151] adopts Faster R-CNN [105] to acquire diversified and discriminative proposals by training the detector on the Visual Genome dataset [64], so that the detected regions are augmented with external attributes. [147] proposes three modular attention networks to address language, subject, and relationship, respectively. The novelty of work [147] is that it employs deep features extracted from two convolutional layers to predict the attributes of regions, and the learned attributes are used for additional information to enhance the regions visual representations.

The authors of work [156] adopt GloVe [98] to represent words, and employs the hidden state of a two-layer BiLSTM (Bidirectional Long Short-Term Memory) [48] to calculate the referent-cue weights. GloVe is a context-free word representation model and generates word vectors in a vocabulary. Context-free word representation overlooks that the same word in different contexts expresses the different semantics. [30] proposes a contextual word representation model, BERT, which takes into account both left and right contexts to generate word representations.

Unlike the above mentioned approaches, we address the visual semantics of regions by taking advantage of the inherent semantic attributes of deep features, i.e., channel-wise and spatial characteristics of extracted deep features. Additionally, we explore the textual semantics by adopting BERT to generate word representa-

tions and employ a language attention network to learn to decompose expressions into multiple phrases to ground target objects.

2.3 Proposed Method

Given a referring expression r with M words $r = \{w_i\}_{i=1}^{M}$ and an image I with N regions of interest (RoIs) $I = \{o_j\}_{j=1}^{N}$, we model the relation between w_i and o_j to locate the target object. In this thesis, we decompose the proposed network for referring expression comprehension into three sub-modules: 1) a language attention network learns to assign different weights to each word in referring expressions, and learns to parse expressions into phrases that denote target candidate, relation between objects, and spatial location information; 2) a visual semantic-aware network generates semantic-aware visual representation, which is acquired by conducting the channel-wise attention and region-based spatial attention; 3) a target localization module achieves targets grounding by combining the outputs of the language attention network and the visual semantic-aware network with the relation and location representations. Figure 2.2 shows the details of the introduced semantics-aware network.

2.3.1 Language Attention Network

We propose a language attention network to compute the different weights of each word in referring expressions and learn to parse the expressions into phrases that embed target candidate r_{tar}, relation r_{rel}, and location r_{loc}, respectively.

For an expression r, we employ BERT [30] to tokenize and encode r into contextualized word embeddings $E_r = [e_1, e_2, ... , e_M]$, where $e_i \in \mathbb{R}^{1\times1024}$. We then feed E_r into an one-layer BiLSTM:

$$L_{out} = \text{BiLSTM}(E_r) \tag{2.1}$$

where L_{out} represents the final hidden representation for each word in referring expressions.

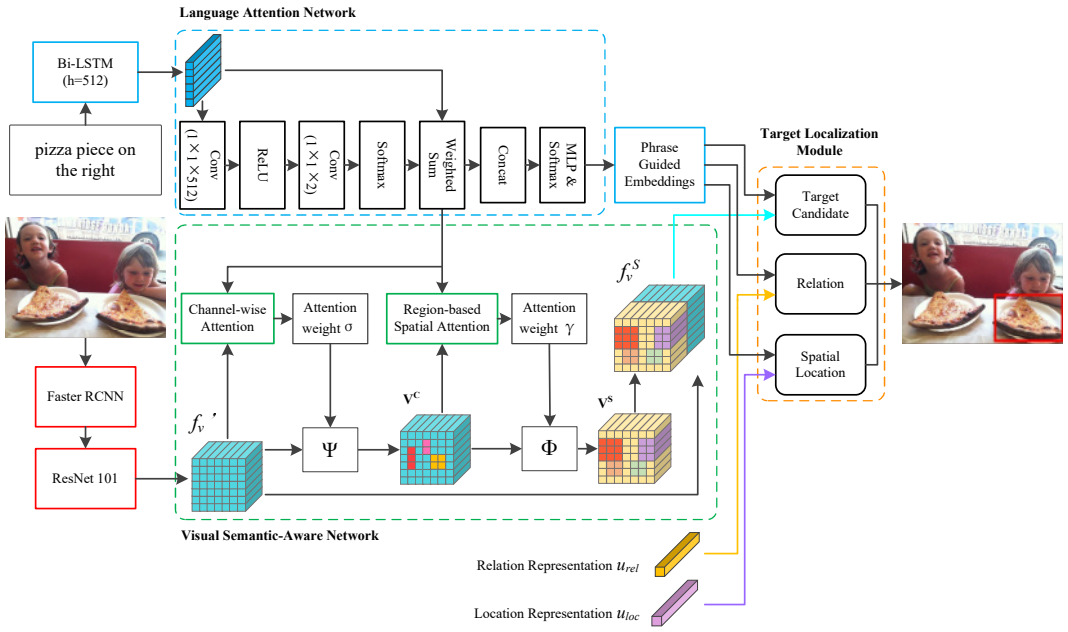

Figure 2.2: Architectural diagram of the proposed semantic-aware network for referring expression comprehension. We present a language attention network to compute the different weights of each word in expressions, and learn to parse the expressions into three phrases that embed the information of target candidate, relation, and spatial location, respectively. We conduct both channel-wise attention and region-based spatial attention to generate semantic-aware region visual representation. We further combine the outputs of the language attention network with the semantic-aware region visual representation, relation representation, and location representation to locate target objects. In the figure, $f_v^{'}$ denotes the projected deep features, $\mathbf{V}^C$ represents the channel-wise weighted deep feature, $\mathbf{V}^S$ is the spatial weighted feature, f_v^S is the generated semantic-aware visual representation by concatenating $f_v^{'}$ and $\mathbf{V}^S$, the details are described in section 2.3.2. The relation representation u_{rel}, the location representation u_{loc}, and the details of the target candidate module, the relation module, and the spatial location module are introduced in section 2.3.3. Ψ denotes a channel-wise multiplication for $f_v^{'}$ and the generated channel-wise attention weight σ, Φ represents element-wise multiplication for $\mathbf{V}^C$ and the acquired spatial attention weight γ (Best viewed in color).

In order to acquire the different weight of each word, we compute attention distribution over the expressions by:

$$\alpha_l = \text{softmax}(\mathcal{F}(L_{out})) \tag{2.2}$$

$$L = \sum_{i}^{g} \alpha_{l,i} L_{out,i} \tag{2.3}$$

where α_l denotes calculated attention weights, $\sum_{m=1}^{M} \alpha_l = 1$. In the implementation, $\mathcal{F}$ is modeled by two convolution layers, and the second convolution layer shares the parameters of the first layer. The glimpse number is set to $g = 2$, therefore, the generated expression representation $L \in \mathbb{R}^{\text{d}\times 2048}$, d is length of expressions in different dataset.

Hu *et al.* [50] decompose the referring expression into (subject, relationship, object) triplets, but not all expressions are well-posed like this construction. For example, an expression like "a bird with a red neck" reveals the target "bird" with specific attribute "red", while "the cow directly to the right of the largest cow" designates the spatial relation between target "cow" and object "largest cow". Expressions like the two exemplars, some words should be parsed to phrase to represent specific information, e.g., "with a red neck", "the right of", and "the largest cow", etc. To this end, we employ a single perceptron and a softmax layer to learn to parse the expressions into three module weights:

$$\overline{L} = \varphi(W_t L + b_t) \tag{2.4}$$

$$[w_{tar}, w_{rel}, w_{loc}] = \text{softmax}(\overline{L}) \tag{2.5}$$

where φ is a non-linear activation function, in the implementation, we adopt the hyperbolic tangent. W_t is a weight matrix and b_t represents a bias vector learned during training. w_{tar}, w_{rel}, w_{loc} represent weights guided by the target candidate phrase, relation phrase and spatial location phrase, respectively.

2.3.2 Visual Semantic-Aware Network

We take full advantage of the characteristics of deep features extracted from a pretrained CNN model, and we conduct channel-wise and region-based spatial attention to generate semantic-aware features for each detected region. This process can be deemed as visual representation enrichment for the detected regions.

RoI Features

Given an image, we adopt Faster R-CNN [105] to generate RoIs, and we extract deep feature $f_v \in \mathbb{R}^{7\times7\times2048}$ for each o_j from the last convolutional layer of the 4th-stage of ResNet101 [46], where 7×7 denotes the size of the extracted deep feature, 2048 is the output dimension of the convolutional layer, i.e. the number of channels. We then project the deep feature f_v into a 512-dimension subspace by a convolution operator with 1×1 kernel, i.e., the projected deep feature $f_v^{'} \in \mathbb{R}^{7\times7\times512}$.

Channel-wise Attention

Essentially, deep features extracted from pretrained CNN models are channel-wise, spatial, and multi-layer. Each channel of a deep feature correlates with a convolutional filter which performs as a pattern detector [18]. For example, the filters in lower layers detect visual clues such as color and edge, while the filters in higher layers capture abstract contents such as object component or semantic attributes. Accordingly, performing channel-wise attention on higher-layer features can be deemed as a process of semantic attributes selection.

We first reshape the projected RoI deep feature $f_v^{'}$ to $\mathbf{V}=[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{d_v}]$, where $\mathbf{v}_i \in \mathbb{R}^{7\times7}$ is the i-th channel of the deep feature $f_v^{'}$, d_v=512. We then perform average pooling on each channel to generate the channel-wise vector $V = [v_1, v_2, ..., v_{d_v}]$, where v_i represents the i-th channel feature.

After the feature pooling, we first utilize L2-normalization to process channel-wise vector V and expression representation r to generate more robust representa-

tions, we then perform channel-wise attention by a channel-wise attention network which is composed of an MLP (multi-layer perceptron) and a softmax layer. For the detected image region, the inputs of the channel-wise attention include the average-pooled feature V and the weighted expression representation L. The channel-wise attention weight is acquired by:

$$\begin{aligned} A_c &= \varphi((W_{v,c}V + b_{v,c}) \otimes (W_{t,c}L + b_{t,c})) \\ \sigma &= \text{softmax}(A_c) \end{aligned} \tag{2.6}$$

where $W_{v,c}$ and $W_{t,c}$ are learnable weight matrices, $b_{v,c}$ and $b_{t,c}$ are bias vectors, $W_{v,c}$ and $b_{v,c}$ are the parameters of the MLP for visual representation, while $W_{t,c}$ and $b_{t,c}$ for textual representation. $\otimes$ denotes outer product, $\sigma \in \mathbb{R}^{1\times 512}$ is the learned channel-wise attention weight which encodes the semantic attributes of regions. In the following, $W_{v,.}$ and $b_{v,.}$ represent the weight matrix and bias vector for visual, $W_{t,.}$ and $b_{t,.}$ are for textual.

Region-based Spatial Attention

The channel-wise attention attempts to address the semantic attributes of regions, while the region-based spatial attention is employed to attach more importance to the referring expression related regions. To acquire region-based spatial attention weights, we first combine the learned channel-wise attention weight σ with the projected deep feature $f_v^{'}$ to generate channel-wise weighted deep feature $\mathbf{V}^C$.

$$\mathbf{V}^C = \Psi(f_v^{'}, \sigma) \tag{2.7}$$

where Ψ is a channel-wise multiplication for deep feature channel and corresponding channel weights, $\mathbf{V}^C \in \mathbb{R}^{49\times 512}$.

We put the weighted channel-wise deep feature $\mathbf{V}^C$ and the weighted expressions into an attention network similar to the channel-wise attention to calculate the spatial attention γ:

$$\begin{aligned} A_s &= \varphi((W_{v,s}\mathbf{V}^C + b_{v,s}) \otimes (W_{t,s}L + b_{t,s})) \\ \gamma &= \text{softmax}(A_s) \end{aligned} \tag{2.8}$$

The acquired $\gamma \in \mathbb{R}^{49\times1}$ denotes the weights of each region related to the expressions, we further fuse the γ with channel-wise weighted feature $\mathbf{V}^C$ to obtain spatial weighted deep feature $\mathbf{V}^S$:

$$\mathbf{V}^S = \Phi(\mathbf{V}^C, \gamma) \tag{2.9}$$

where Φ denotes element-wise multiplication for regions of each deep feature channel and the corresponding region attention weights. $\mathbf{V}^S \in \mathbb{R}^{7\times7\times512}$ comprises the semantics guided by the channel-wise attention as well as the weight of each region. Therefore, we define $\mathbf{V}^S$ as semantic-aware deep feature. Finally, we concatenate $\mathbf{V}^S$ with projected feature $f_v^{'}$ to obtain semantic-aware visual representation for each region, i.e., $f_v^S = [f_v^{'}\ ;\ \mathbf{V}^S]$, $f_v^S \in \mathbb{R}^{7\times7\times1024}$, $[\cdot\ ;\ \cdot]$ denotes the concatenate operation.

2.3.3 Target Localization Module

In order to locate target objects for given expressions, we need to sort out the relevant candidates, the spatial location, and the appearance difference between the candidate and other objects. For instance, to ground the expression "the cow directly to the right of the largest cow", we need to understand the spatial location "the right of", and the appearance difference "largest" between the cows to identify the target "cow". To this end, we calculate the matching score of the target candidates, the relation, and the spatial location via a target candidate module, a relation module, and a spatial location module, respectively.

Target Candidate Module

We compute the target candidate phrase matching score by the target candidate module. Given a region semantic-aware representation f_v^S and a target candidate phrase guided expression embedding r_{tar}, we process them by L2-normalization and linear transform to compute the attention weights on each region:

$$\begin{aligned} t &= \varphi((W_v f_v^S + b_v) \otimes (W_t r_{tar} + b_t)) \\ \beta &= \text{softmax}(t) \end{aligned} \tag{2.10}$$

where β denotes the learned region-based attention weight.

We fuse β and f_v^S to obtain the target candidate phrase attended region visual representation u_{tar}, and we further compute the target candidate matching score s_{tar} by:

$$\begin{aligned} u_{tar} &= \beta \otimes f_v^S \\ \overline{u}_{tar} &= W_{v,tar} u_{tar} + b_{v,tar} \\ \overline{r}_{tar} &= W_{t,tar} r_{tar} + b_{t,tar} \\ s_{tar} &= \mathcal{D}(\overline{u}_{tar}, \overline{r}_{tar}) \end{aligned} \tag{2.11}$$

where $\mathcal{D}(\cdot\ ,\ \cdot)$ represents the consine distance measurement.

Relation Module

We adopt a relation module to obtain the matching score of a pair of candidates and relation guided phrase embedding r_{rel}. We use the average-pooled channel vector V as the appearance representation for each candidate. To tackle with the appearance difference between candidates, e.g., "the largest cow", we calculate the visual appearance difference representation $\delta v_i = \frac{1}{n} \sum_{j \neq i} \frac{v_i - v_j}{||v_i - v_j||}$ as [148], where n is the number of candidate chosen for comparison (in our implementation n = 5). We concatenate V and δv_i as the candidates visual relation representation u_{rel}, i.e., $u_{rel} = [V\,; \delta v_i]$. We calculate the relation matching score by:

$$\begin{aligned} \overline{u}_{rel} &= W_{v,rel} u_{rel} + b_{v,rel} \\ \overline{r}_{rel} &= W_{t,rel} r_{rel} + b_{t,rel} \\ s_{rel} &= \mathcal{D}(\overline{u}_{rel}, \overline{r}_{rel}) \end{aligned} \tag{2.12}$$

Spatial Location Module

We calculate the location matching score through the location module. To deal with the spatial relation of candidates in images, following [148], we adopt a 5-dimensional spatial vector $u_l = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}]$ to encode the top left position, bottom right position, and the relative size of the candidates in images. In order

to address the relative position expression like "the right of", "in the middle", we adopt the relative location vector $\Delta u_{ij} = [\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w_j \cdot h_j}{w_i \cdot h_i}]$ which is obtained by comparing with five surrounding objects and concatenate with u_l to generate candidate location representation $u_{loc} = [u_l \ ; \Delta u_{ij}]$.

Similar to the target candidate module, we process u_{loc} and location phrase r_{loc}, and then combine the transformed u_{loc} and r_{loc} to generate the location matching score s_{loc}:

$$\begin{aligned} \overline{u}_{loc} &= W_{v,loc} u_{loc} + b_{v,loc} \\ \overline{r}_{loc} &= W_{t,loc} r_{loc} + b_{t,loc} \\ s_{loc} &= \mathcal{D}(\overline{u}_{loc}, \overline{r}_{loc}) \end{aligned} \tag{2.13}$$

2.3.4 Learning Objective

Given an image I and expression r pair, we calculate the target candidate score, relation score and spatial location score, through the three above mentioned modules. We locate the target object by the final grounding score:

$$G(o_i|r) = w_{tar} s_{tar} + w_{rel} s_{rel} + w_{loc} s_{loc} \tag{2.14}$$

In the implementation, we adopt a combined max-margin loss as the objective function:

$$\mathcal{L}_\theta = \sum_i [max(0, \xi - G(o_i|r_i) + G(o_i|r_j)) + max(0, \xi - G(o_i|r_i) + G(o_k|r_i))] \tag{2.15}$$

where θ denotes the parameters of the proposed model to be optimized, ξ is the margin between positive and negative samples. During training, we set $\xi = 0.1$. For each positive target and expression pair (o_i, r_i), we randomly select negative pairs (o_i, r_l) and (o_t, r_i), where r_l is the expression for other objects, o_t is the other object in the same image.

2.4 Experiments

2.4.1 Datasets

The introduced network is trained and validated on three popular referring expression datasets: RefCOCO [148], RefCOCO+ [148], and RefCOCOg [84]. The images of the three datasets were collected from MSCOCO dataset [75]. The referring expressions in RefCOCO and RefCOCO+ were collected in an interactive manner [60], while RefCOCOg expressions were collected in a non-interactive way.

RefCOCO comprises 142,210 expressions for 50,000 referents in 19,994 images. The dataset is divided into training, validation, testA and testB which contains 120,624, 10,834, 5,657 and 5,095 referent-expression pairs, respectively.

RefCOCO+ has 141,564 expressions for 49,856 referents in 19,992 images. The split is same as RefCOCO and each subset contains 120,191, 10,758, 5,726, and 4,889 referent-expression pairs, respectively. Comparison with RefCOCO, RefCOCO+ discards absolute location words and attaches more importance to appearance differentiators.

RefCOCOg contains 95,010 expressions for 49,822 refs in 25,799 images. RefCOCOg was collected in a non-interactive pattern, therefore the referring expressions in RefCOCOg are longer than RefCOCO and RefCOCO+. RefCOCOg has two types of data splitting, [84] splits the dataset into train and validation sets, and no test set is published. Therefore, most existing work evaluates their performance on the validation set. We denote this data split as RefCOCOg "val*". Another data partition [89] splits the dataset as training, validation and test sets. We run experiments on this split and we denote as RefCOCOg "val" and "test".

The referring expressions in RefCOCO and RefCOCO+ were collected in an interactive manner [60], the average length of expressions in RefCOCO is 3.61, and the average number of words in RefCOCO+ expressions is 3.53. While RefCOCOg expressions were collected in a non-interactive way, therefore produces longer expressions than the RefCOCO and RefCOCO+, and the average length of RefCOCOg expressions is 8.43. From the perspective of expression length distri-

bution, 97.16% expressions in RefCOCO contain less than 9 words, the proportion in RefCOCO+ is 97.06%, while 56.0% expressions in RefCOCOg are less than 9 words.

2.4.2 Experimental Setup

In practice, the length of the sentences is set to 10 for the expressions in RefCOCO and RefCOCO+, and pad with "pad" symbol to the expressions whose length is smaller than 10. The length of the sentences is set to 20, and the same manner is adopted to process the expressions in RefCOCOg.

The "bert-large-uncased" model[1] is employed to generate contextualized word embedding E_r. According to [30], the word embedding from the sum of the last four layers acquire better results than the embedding extracted from the last layer. We select the embedding of the sum of the last four layers of BERT as E_r. Therefore, the obtained expression representation $q \in \mathbb{R}^{10\times1024}$ for RefCOCO and RefCOCO+, and $q \in \mathbb{R}^{20\times1024}$ for RefCOCOg.

For a given image and referring expression pair, the final ground score defined in Equation 2.14 is utilized to compute the matching score for each object in the image, and pick the one with the highest matching score as the correct one. IoU (Intersection over Unit) between the predicted region and the ground truth bounding box is computed, and the value larger than 0.5 is selected as the correct visual grounding.

The model is trained with Adam optimizer with coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 0.0004 and decay every 5,000 iterations with weight decay 0.0001, and the total number of iterations is up to 30,000.

2.4.3 Ablation Analysis

In Table 2.1, the different modules of the proposed network are combined to validate their performance and effectiveness. According to [151] and [147], the models

[1] `https://github.com/huggingface/pytorch-pretrained-BERT`

		RefCOCO			RefCOCO+			RefCOCOg	
		val(%)	testA(%)	testB(%)	val(%)	testA(%)	testB(%)	val(%)	test(%)
1	sub(ProjFeat)+loc	79.28	79.57	80.37	64.77	65.29	62.41	69.63	69.28
2	sub(ProjFeat)+loc+rel	79.99	80.24	80.82	64.89	66.00	63.57	70.14	69.96
3	sub(SemanAware)+loc	80.59	80.61	81.73	64.20	65.89	63.47	72.94	72.72
4	sub(SemanAware)+loc+rel	81.24	81.42	82.20	65.11	66.03	63.76	72.98	72.76
5	sub(ProjFeat)+loc+rel+LangAtten	81.83	82.10	82.20	66.42	67.46	63.84	73.33	72.81
6	sub(SemanAware)+loc+rel+LangAtten	**83.51**	**83.74**	**83.18**	**68.16**	69.66	**64.66**	**76.00**	**74.81**
7	sub(SemanAware)+loc+rel+LangAtten(I)	83.25	82.55	82.55	67.77	**69.70**	64.00	74.53	73.61

Table 2.1: Ablation studies of the proposed network using different module combinations.

trained by the deep features extracted from VGG16 [119] generates lower accuracy than the features from ResNet101, so the model is trained using the ResNet101 deep features rather than the VGG features.

First, the performance of the proposed model is validated from the visual perspective. The projected feature $f_v^{'}$ and location representation u_{loc} are concatenated as the visual representation for each region, and the output of the BiLSTM is used to the representation for expressions. This combination is deemed as the baseline, and the results are listed in Line 1. And then the relation representation is added to evaluate the benefits of the relation module, and the results are listed in Line 2.

Second, the effect of the visual semantic-aware module (section 2.3.2) is tested by selecting the semantic-aware visual representation f_v^S as the region visual representation. The f_v^S is combined with the spatial location and relation representation, respectively. Compared with Line 1 and Line 2, the results in Line 3 and Line 4 demonstrate the performance of the visual semantic-aware network. The results acquired by employing the semantic-aware visual representation f_v^S are improved by nearly 2% than the projected deep feature $f_v^{'}$.

Third, two manners are taken advantage to evaluate the performance of the language attention network. We first combine $f_v^{'}$ with the language attention, it is clear that the results outperform the results listed in Line 2. An interesting finding is that the results listed in Line 4 are close to Line 5, it also demonstrates

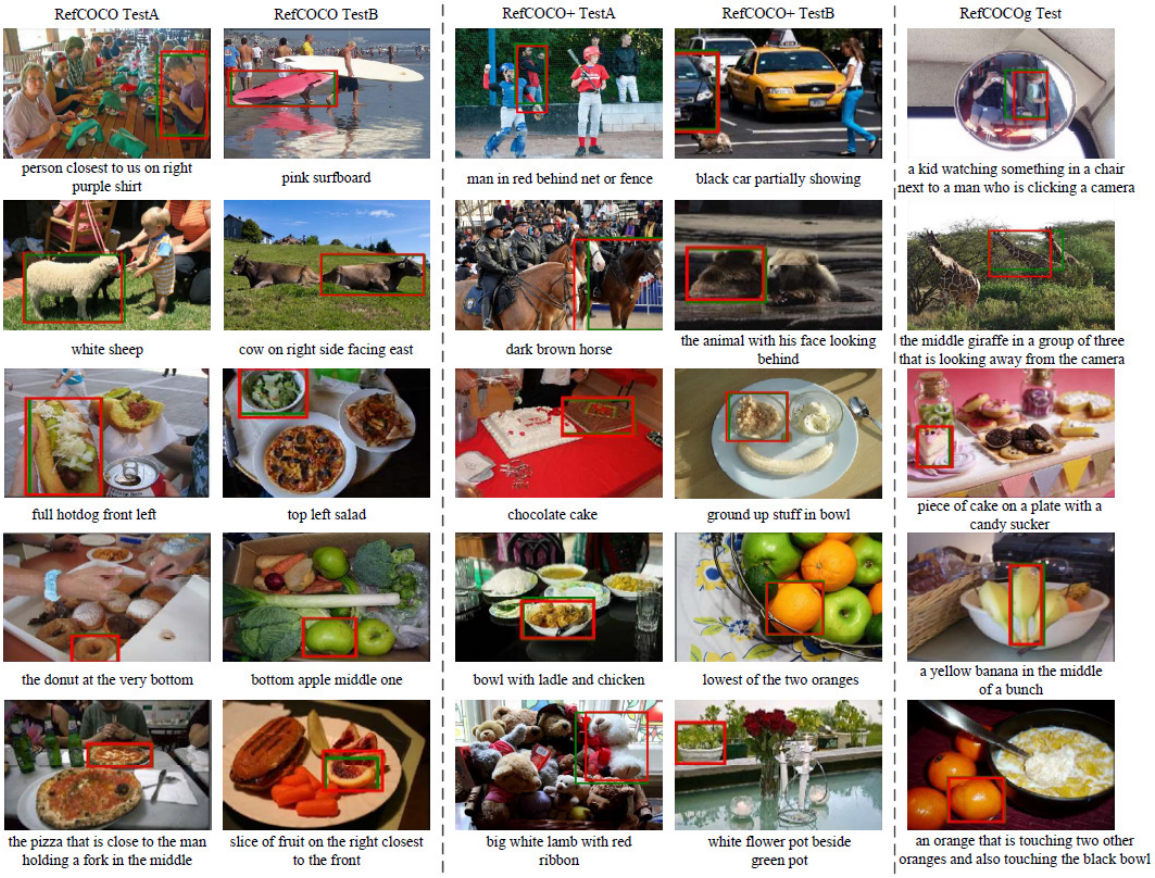

Figure 2.3: Example results acquired by the proposed semantic-aware network on RefCOCO, RefCOCO+, and RefCOCOg. Referring expressions are underlay the images. The red boxes show the correct groundings and the green bounding boxes denote the ground truth bounding boxes.

the benefits of the visual semantic-aware module. We then coalesce the language attention with f_v^S and the other three representations, this combination acquires the best accuracies on the three test datasets.

Fourth, the influence of the different word embeddings is compared by employing the word feature extracted from the different layers of BERT. The embeddings extracted from the last layer of BERT as the contextual representation and feed into the language attention, this word embedding is denoted as LangAtten(I). Line 7 lists the obtained results. Compared with the results in Line 6, it is demonstrated the accuracy benefits from the advantage of the embeddings from the sum of the

		RefCOCO			RefCOCO+			RefCOCOg		
		val(%)	testA(%)	testB(%)	val(%)	testA(%)	testB(%)	val*(%)	val(%)	test(%)
1	visdif[148]	-	67.57	71.19	-	52.44	47.51	59.25	-	-
2	MMI[84]	-	63.15	64.21	-	48.73	42.13	55.16	-	-
3	attr+MMI+visdif[78]	-	78.85	78.07	-	61.47	57.22	69.83	-	-
4	**Speaker**+Listener+Reinforcer[149]	79.56	78.95	80.22	62.26	64.60	59.62	72.63	71.65	71.92
5	Speaker+**Listener**+Reinforcer[149]	78.36	77.97	79.86	61.33	63.10	58.19	72.02	71.32	71.72
6	VC[156]	-	78.98	82.36	-	62.56	62.90	73.98	-	-
7	DDPN+VGG16[151]	76.9	67.5	73.4	67.0	50.2	60.1	-	-	-
8	DDPN+ResNet101[151]	80.1	72.4	76.8	70.5	54.1	64.8	-	-	-
9	CMN[50]	-	-	-	-	-	-	69.30	-	-
10	AccuAtten[28]	81.27	81.17	80.01	65.56	68.76	60.63	73.18	-	-
11	PLAN[159]	81.67	80.81	81.32	64.18	66.31	61.46	69.47	-	-
12	MAttNet+VGG16[147]	80.94	79.99	82.30	63.07	65.04	61.77	73.08	73.04	72.7
13	LGRANs [133]	82.0	81.2	**84.0**	66.6	67.6	**65.5**	-	75.4	74.7
14	VisSemanAware+LanAtten	**83.51**	**83.74**	83.18	**68.16**	**69.96**	64.66	-	**76.00**	**74.81**

Table 2.2: Comparison with the state-of-the-art approaches.

last four layers of BERT.

Finally, some example results of referring expression comprehension on the three datasets are shown in Figure 2.3. Incorporate the visual semantic-aware network with the language attention network, the introduced model is able to locate the target objects for complex referring expressions, as shown in the experimental results on RefCOCOg.

2.4.4 Comparison with State-of-the-art

Table 2.2 lists the results acquired by the proposed model and the state-of-the-art models. The table is split into two parts over the rows: the first part lists the approaches without introducing the attention mechanism. The second illustrates the results acquired by attention integrated models.

First, the proposed model outperforms the other approaches and acquire competitive results with the current state-of-the-art approaches. [147] extracts the features from the last convolutional outputs of the third stage and the fourth stage,

and utilizes the two different deep features to predict attributes of regions. The predicted attribute is concatenated with the features from the fourth stage as the visual representation for regions. The results of [147] benefit from the features extracted from two different layers. While in our implementation, we extract features from one layer and utilize the innate semantic attributes.

Second, through the experiments on the three datasets, the introduced model acquires better results on RefCOCO compared with the results on RefCOCO+ and RefCOCOg. The expressions in RefCOCO frequently utilize the location or other details to describe target objects, the expressions in RefCOCO+ abandon the location descriptions while adopts more appearance differences. While the expressions in RefCOCOg involves the surrounding objects of the targets and frequently use the relation of objects to depict the target objects.

Finally, some failure cases on the three datasets are shown in Figure 2.4. For complex expression, similar to "small table next to the chair", the proposed model generates closest weights for "table" and "chair". Moreover, to locate the object with vague visual features, such as the target for "black sleeves" in the first left image and "guy leg out" in the third image of the second row, our model frequently generates wrong predictions. For the long expression and image with the complex background, such as the two images in RefCOCOg, our model fails to generate correct predictions.

2.5 Discussion

In this section, we proposed a semantics-aware network for referring expression comprehension. Unlike the existing approaches, we excavated the visual semantic by taking full advantage of the characteristic of the extracted region deep features from a pretrained CNN model, and conducted channel-wise and region-based spatial attention to enrich region visual representation. Moreover, we exploited the rich linguistic structure of referring expression via contextualized word embeddings and a language attention network. Finally, we trained and validated the proposed

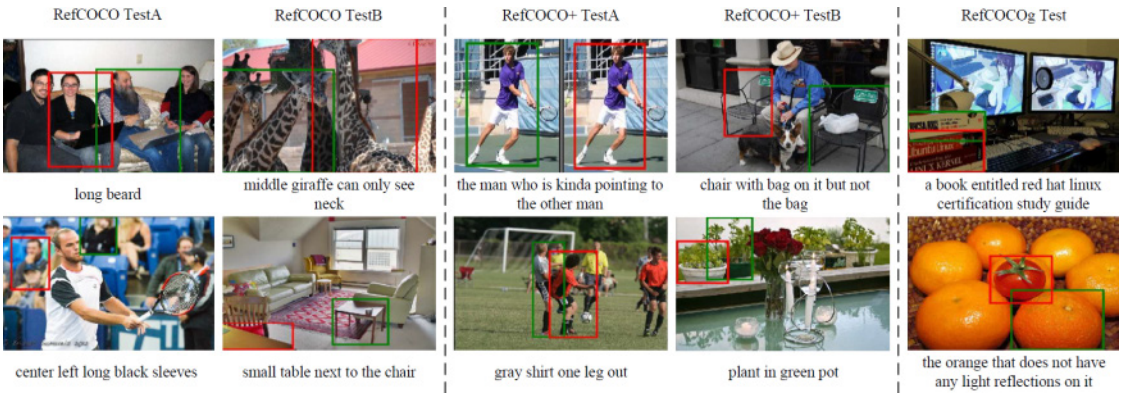

Figure 2.4: Examples of incorrect predictions. The red boxes show our wrong visual groundings, and the green boxes denote the ground truth bounding boxes.

network on three public datasets, RefCOCO, RefCOCO+, and RefCOCOg.

In the future, we will adopt different approaches to exploit the rich context of referring expressions, such as develop a method to parse the expressions in a more natural way. Additionally, we will address the interpretability and robustness of the presented model.

Chapter 3

Referring Expression Generation via Adversarial Training

3.1 Introduction

Referring expressions not only depict the attributes of objects, such as color, size, and location, but also describe spatial relationships between objects within images. Compared to generic image captions, referring expressions are context-aware and contain more accurate and rich descriptions for objects. Tasks that utilize textual descriptions or questions to help human beings understand or depict images and scenes are in agreement with the human desire to understand visual contents at a high semantic level. Examples of these tasks include dense captioning [56], visual question answering [5], referring expression comprehension [148], referring expression generation [148].

Referring expression comprehension imitates the role of a listener to ground target objects in given images, while referring expression generation mimics the role of a speaker to generate referring expressions for each detected region within images. Existing approaches mainly adopt Encoder-and-Decoder paradigms to generate expressions as provided by the ground-truth [148, 84, 149, 78]. The Encoder-and-Decoder models which adopt CNN to process visual features and employ LSTM to generate sequence words, and the training objective is to maximize the resem-

blance to the ground truth samples. Moreover, popular evaluation metrics, such as BLEU [95], ROUGE[73], and METEOR [29], mostly tend to match the n-grams with the ground-truth.

The existing approaches adopt generic generation paradigm and evaluation metrics to produce plausible expressions. However, the conventional methods bring about two crucial issues for referring expression generation. First, the generated expressions tend to be easy for humans to locate target objects in given images without taking into account the semantic validity of expressions. Moreover, the expressions generated by the generic approaches are easy for humans to discriminate from natural expressions. Second, the diversity of expressions is overlooked. Different people would probably depict image regions using different wording pattern, rather than follow the mode of training samples. While the diversity in expressions is an indispensable attribute of human language.

In this thesis, we aim to improve the diversity and naturalness of generated referring expressions, i.e., generating expressions that are adequately easy for humans to ground target objects within images without sacrificing the semantic validity. Inspired by the successful applications of Generative Adversarial Networks (GANs) in generating diverse and real-valued data, we introduce a GAN-based architecture to generate diverse and natural referring expressions.

GANs were originally introduced in [42] and have been widely used in image synthesis [104, 154, 52, 134, 11]. The crucial constituents of GANs are a generator and a discriminator, where the generator tries to generate realistic samples to coax the discriminator, while the discriminator tries to discriminate real samples from generated ones. In image synthesis, GANs learn a loss to classify if the generated images are real or fake, and simultaneously training a generative model to minimize the loss. Moreover, according to the results reported in [134, 11], the synthesized images are high-resolution and nearly indistinguishable from real photos without any hand-crafted losses or pretrained networks.

GANs are also adopted to generate captions [24, 113]. Compared to image synthesis, in which the transformation from the input vector to the synthesized

image is a continuous mapping process. In contrast, the process of generating referring expressions is a sequential sampling procedure, which poses a challenge when trying to update gradients during the generator training. [54, 82] propose Gumbel sampler to overcome the inability to apply re-parameterization trick to discrete data generation and allow for end-to-end training the generator. Motivated by these work, we adopt Gumbel-softmax for a GAN-based architecture to generate referring expressions for each detected image region.

In this thesis, the objective is to generate expressions in a pattern that better imitates the way humans depicting image regions while reserving the semantic validity of expressions. Thus, we formulate the expression generator as a generative adversarial network, and we propose a discriminator which encourages the generator to generate expressions to be diverse and natural. We train the generator with an adversarial loss with the discriminator. We train and evaluate the proposed network on RefCOCO, RefCOCO+, and RefCOCOg. The introduced generative network acquires competitive results compared with the current state-of-the-art approaches.

3.2 Related Work

3.2.1 Image Captioning

Different from referring expression generation, image captioning aims to generate natural language sentences to describe the general content of given entire images. Existing work adopts different approaches to address the multimodal task. These methods first detect object concepts by Conditional Random Field [25], or CNNs [36, 72], and then generate captions using sentence template [67], or retrieving sentences from existing data [36, 29].

Vinyals *et al.* [131] initially introduces an Encoder-and-Decoder model for caption generation, and the Encoder-and-Decoder-based approaches become popular [139, 145, 79, 142, 141]. These models first extract visual features through pre-

trained CNNs and generate captions sequentially via LSTM, and these models learn the parameters by maximizing the conditional log-likelihood of the training samples. A vital issue of the maximum likelihood principle-based models is that the generated captions often tend to replicate the generic sentence from a training set for given similar images [31].

Plenty of proposed approaches aim to generate captions with diversity and naturalness. [68] combines Maximum Mutual Information (MMI) with beam search to produce more diverse and interesting captions. [130] introduces diverse beam search which decodes diverse lists by decomposing the beam budget into groups and implementing diversity between groups of beams. [76] adopts a Natural Language Understanding component in training to optimize the specificity of the caption generation component, and employs multiple objective functions to generate diverse and meaningful captions. [153] argues that the existing models fail to capture visual contexts such as object relationships, and thereon introduces a context-aware visual policy network to generate context-aware descriptions for image regions.

Dai *et al.* [24] improves the naturalness and diversity of generated captions by employing conditional GAN to train the caption generator. This model jointly trains an evaluator in an adversarial way to discriminate irrelevant or artificial captions from natural ones. [113] formulates the caption generator as a generative adversarial network, and designs a discriminator to generate captions which are diverse and indistinguishable from human captions. While training an adversarial generator, caption generation is a sequential sampling process, and the operation is non-differential which poses a challenge to apply gradient back-propagation. [113] depends on the reinforcement rule to handle back-propagation, i.e., utilizes Monte Carlo rollouts [146] to compute the approximated future reward. While [113] employs Gumbel Sampler [54, 82] to achieve end-to-end training.

Although the approaches mentioned above acquire promising results, they generate captions for entire images and the captions cannot be grounded on a set of image regions. [80] introduces a novel captioning pattern which first generates a word-level sentence template with slot locations, and then the slots are filled by

object detectors with visual concepts detected in images. In this way, this model generates grounded image captioning. [23] depicts the same image by selecting concentrated regions in a different order and focuses on generating diverse captions via the control signal given as a sequence or as a set of image regions.

Another similar task is dense captioning, which aims to describe salient image regions within images in natural language. [56] initially introduces dense captioning, and this work proposes a Fully Convolutional Localization Network to locate objects and adopts a Recurrent Neural Network-based language model to generate label sequences. [140] exploits a dense captioning model in a methodical manner in which joint inference locates each visual concept accurately, and context fusion combines pooled features from image regions to produce better region descriptions. [70] learns a complementary object context for each caption region and transfers knowledge from objects to caption regions. In this way, this model generates context-aware descriptions for each image region. [144] investigates a context and attribute grounded dense captioning model that produces captions with context information, which includes the local, neighboring, and global data.

3.2.2 Referring Expression Generation

Compared to the captions generated by generic captioning models or dense captioning models, referring expressions not only depict image regions using properties such as color, size, and location, but also involve the interaction information between their neighboring objects. The goal of referring expression generation is to generate unambiguous natural language descriptions for detected objects or regions within given images. Thus, referring expression generation is more easily evaluated and can be used in interactive scenarios.

Referring expression generation has been studied for several years [62, 87]. [148] takes advantage of visual differences of objects and employs a CNN-LSTM paradigm to produce expressions. Similar to [148], [84] adopts the same pattern to generate expressions and uses beam search to approximately find the most probable descriptions. This work adds MMI to encourage the generator to produce better

descriptions for the target object than the other objects within the image. [149] utilizes the appearance similarity, size and location similarity to represent image regions and also produces expressions via a CNN-LSTM model. [78] explores the role of object attributes in expression generation and extends the generic CNN-LSTM model to hearten the generation bears more accurate attributes correlated with the input attributes. [81] first trains a comprehension module on human-generated expressions, and the trained model is selected to be a critic for referring expression generator. This work also follows the CNN-LSTM paradigm to generate expressions. [124] improves the method introduced by [149] and presents a new referring expression generation dataset. However, this model also follows the CNN-LSTM paradigm to generate referring expressions.

Unlike the existing approaches, we aim to generate diverse and natural referring expressions that are sufficiently easy for humans to locate the target objects and without sacrificing the semantic validity of generated expressions.

3.3 Evaluation Metrics

Accompanied by the development of the captioning generation approaches, multiple evaluation metrics have been introduced to evaluate the quality of generated natural language sentences. Classical metrics such as BLEU [95] focuses on precision and ROUGE [73] emphases on the recall of n-grams. These metrics show weak associations with human judgment [34, 67]. In order to measure the overall quality of generated descriptions, METEOR [29] evaluates both the precision and recall of n-grams, and [128] proposes CIDEr which computes the similarity of generated sentences against a set of ground truth written by humans. CIDEr shows high consistent with consensus as evaluated by humans.

In order to evaluate referring expressions generated by the proposed network, we adopt four types of evaluation metrics: BLEU@N [95], METEOR [29], ROUGE-L [73], and CIDEr [128].

BLEU (Bilingual Evaluation Understudy) [95], which is a score to compare the

N-gram overlapping of a candidate translation of the text to one or more reference translations. BLEU is widely used to evaluate text generation for a suite of natural language processing tasks. BLEU tries to compute the match average of variable length phrases between candidate translations and reference translations, and the acquired match averages are applied to assess the translation score. The BLEU metric requires to compute the brevity penalty BP by:

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ exp(1 - r/c), & \text{otherwise} \end{cases}$$

where c is the length of candidate translation, r represents the effective reference corpus length.

The basic BLEU is calculated as follows:

$$\text{BLEU} = \text{BP} \cdot exp(\sum_{n=1}^{N} w_n log p_n) \tag{3.1}$$

where w_n is positive weights summing to one, and p_n denotes the N-gram precision computed using N-grams with a maximum length of N. In our experiments, we use N = 1, 2, 3, 4 and the associated metrics are denoted as BLEU@1, BLEU@2, BLEU@3, and BLEU@4.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [29], which is utilized to calculate the harmonic mean of uni-gram matches' precision and recall between generated sentences and ground truths. METEOR calculates higher order N-grams, considers word-to-word matching, and applies arithmetic averaging for a final score. The METEOR metric is acquired by:

$$\begin{aligned} F_{mean} &= (\frac{PR}{\alpha P + (1-\alpha)R}) \\ Pen &= \gamma(\frac{ch}{m})^{\beta} \\ score &= (1 - Pen) \cdot F_{mean} \end{aligned} \tag{3.2}$$

where P and R denote the precision and recall of uni-gram matches, ch denotes the number of chunks and m is the number of matches. α controls the relative weight of precision and recall, γ determines the maximum penalty, β determines the

functional relation between the fragmentation ch/m and the penalty. In practice, the three parameters are set to $\alpha = 0.9$, $\beta = 3.0$, $\gamma = 0.5$ for maximizing correlation with human judgments.

ROUGE_L (Recall-Oriented Understudy for Gisting Evaluation) [73], which measures the longest common subsequences between a pair of sentences. ROUGE_L calculates the ratio between the size of two summaries' longest common subsequences and the size of the reference summary. ROUGE_L is computed by:

$$\begin{aligned} R_{lcs} &= \frac{LCS(X,Y)}{m} \\ P_{lcs} &= \frac{LCS(X,Y)}{n} \\ F_{lcs} &= \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \end{aligned} \tag{3.3}$$

where $LCS(X,Y)$ denotes the length of a longest common sequence of two given summaries X and Y, m and n represent the length of X and Y respectively. $\beta = P_{lcs}/R_{lcs}$, F_{lcs} is the calculated ROUGE_L score.

CIDEr (Consensus-based Image Description Evaluation) [128], which is proposed to evaluate the quality of image descriptions. CIDEr measures the consensus between candidate image captions and the reference sentences. CIDEr extends existing metrics with $tf - idf$ weighting over N-grams. The Term Frequency Inverse Document Frequency (TF-IDF) weighting $g_k(c_i)$ is calculated by:

$$\begin{aligned} TF(k) &= \frac{h_k(c_i)}{\sum h_l(c_i)} \\ IDF(k) &= log(\frac{N}{\sum_1^N min(1, \sum_1^M h_k(c_i))}) \\ g_k(c_i) &= TF(k) * IDF(k) \end{aligned} \tag{3.4}$$

where $h_k(c_i)$ represents n-gram occurs in candidate sentence c_i.

The CIDEr score CIDEr_n for n-grams of length n is calculated using the average cosine similarity between the candidate sentence c_i and the reference sentences $S_i = s_{i1,\ldots,s_{im}}$ by:

$$\text{CIDEr}_n = \frac{1}{M}\sum_{j=1}^{M}\frac{g^n(c_i)\cdot g^n(s_{ij})}{\|g^n(c_i)\|\,\|g^n(s_{ij}\|} \tag{3.5}$$

where $g^n(c_i)$ is a vector formed by $g_k(c_i)$, $\|g^n(c_i)\|$ denotes the magnitude of the vector $g^n(c_i)$.

3.4 Proposed Method

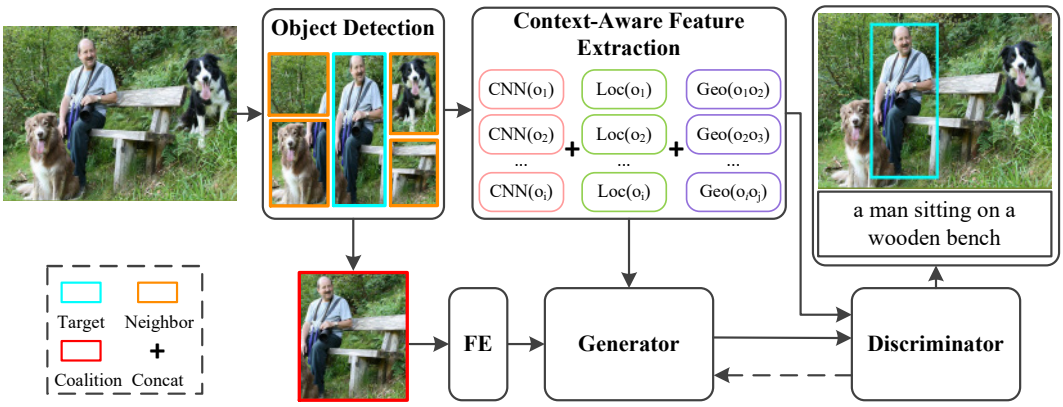

Figure 3.1: Diagram of the adversarial training-based network for referring expression generation. We generate context-aware visual representation for detected regions within an image. We propose a generator to generate expressions and a discriminator to classify whether the generated expressions are real sentences from the corpus or are generated by the generator. FE represents feature extraction.

Given an image I with N region of interests (RoIs) $I = \{o_i\}_{i=1}^{N}$, we generate referring expressions for each region o_i. In this work, we reformulate referring expression generation as a generative network which is composed of an expression generator to generate expressions and a discriminator to minimize the objective loss. Figure 3.1 illustrates the details of the generative architecture for referring expression generation.

3.4.1 Context-Aware RoI Representation

Similar to the proposed semantic-aware network for referring expression comprehension, we adopt Faster R-CNN [105] to detect RoIs, and extract deep features from the last convolutional layer of the 4th-stage of ResNet101 [46] , i.e. region

deep feature $f_v \in \mathbb{R}^{7\times7\times2048}$. We also perform average pooling on each channel to generate vector $V \in \mathbb{R}^{1\times2048}$ as the visual representation for each detected region.

In order to address the properties of regions, such as size and location, we adopt 5-dimensional spatial vector $u_l = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}]$ to encode the location and size of the region, where x and y are the top left and bottom right values, w and h represent the width and height of the region, and W and H are the width and height of the image.

For the interaction information between target and their neighbor objects, we employ a geometric feature for the subject and object pair. Given bounding boxes for subject $b_s = [x_s, y_s, w_s, w_s]$ and object $b_o = [x_o, y_o, w_o, w_o]$, where (x, y) are the center values of the box, and (w, h) are the width and height of the box, we adopt the geometric feature defined in [100] as:

$$u_g = [\frac{x_o - x_s}{\sqrt{w_s h_s}}, \frac{y_o - y_s}{\sqrt{w_s h_s}}, \sqrt{\frac{w_o h_o}{w_s h_s}}, \frac{w_s}{h_s}, \frac{w_o}{h_o}, \frac{b_s \cap b_o}{b_s \cup b_o}] \tag{3.6}$$

where $u_g \in \mathbb{R}^{1\times6}$. We process the spatial vector u_l and the geometric feature u_g by two fully connected layers to generate length uniformed representations, i.e. generated $u_l^{'} \in \mathbb{R}^{1\times64}$ and $u_g^{'} \in \mathbb{R}^{1\times64}$. We then concatenate the V, $u_l^{'}$ and $u_g^{'}$ as the context-aware representation for each detected object in an image, i.e. $V^{'} = [V; u_l^{'}; u_g^{'}]$, $[\cdot\ ;\ \cdot]$ represents the concatenate operation.

3.4.2 Gumbel-Softmax for Discreteness Problem

The authors of work [54, 82] introduce the Gumbel-Softmax trick that combines a continuous relaxation of the one-hot encoded vector for the discrete samples with the re-parameterization of the sampling process to achieve back-propagation. The Gumbel-Softmax trick attempts to tackle with the inability to apply the re-parameterization trick to generated discrete samples by GANs.

Given a random variable g, if g = -log(-log(u)) with $u \sim$ Uniform[0, 1], g obeys standard Gumbel distribution. The importance of the Gumbel distribution is that any discrete distribution can be parameterized in terms of Gumbel random variables. Lext X be a discrete random variable with distribution $P(X = k) \propto$

α_k (α_k is a random variable), and g_k is a sequence obeying standard Gumbel distribution. We can calculate X by:

$$X = \arg\max_k(log(\alpha_k) + g_k) \tag{3.7}$$

Although the argmax operation relates the Gumbel samples, the α_k and the realization of the discrete distribution are not continuous. As suggested in [54] and [82], one way of circumventing this is to relax the discrete set by considering random variables taking values in a large set. Note that any discrete random variable can always be expressed as a one-hot vector (i.e., a vector filled zeros except for an index where the coordinate is one). By mapping the realization of the variable to the index of the non-zero entry of the vector, we can compute the probability simplex by the convex hull of the set of the one-hot vector as follows:

$$\Delta^{K-1} = \left\{ x \in \mathbb{R}_+^K, \sum_{k=1}^{K} x_k = 1 \right\} \tag{3.8}$$

where Δ^{K-1} denotes (K-1)-dimensional simplex.

Through computing the probability simplex, we can construct the relaxation of discrete samples. Thus, a natural way to relax a discrete random variable is to take values in the probability simplex. Both [54] and [82] present to consider the softmax map indexed by a temperature parameter by:

$$f_\tau(x)_k = \frac{exp(x_k/\tau)}{\sum_{k=1}^{K} exp(x_k/\tau)} \tag{3.9}$$

where τ represents the softmax temperature, and $\tau \in (0, \infty)$.

Instead of the discrete valued random variable X, we can calculate the sequence of simplex-valued random variable X^τ by:

$$X^\tau = f_\tau(log(\alpha) + g) = \left(\frac{exp((log(\alpha_k) + g_k)/\tau)}{\sum_{k=1}^{K} exp((log(\alpha_i) + g_i)/\tau)} \right) \tag{3.10}$$

The generated X^τ obeys the concrete distribution, denote as $X^\tau \sim$ Concrete(α, τ). The density of the Gumbel-Softmax distribution is given by:

$$p_{\alpha,\tau}(x) = (n-1)!\tau^{n-1} \prod_{k=1}^{K} \left(\frac{\alpha_k x_k^{-\tau-1}}{\sum_{k=1}^{K} \alpha_i x_i^{-\tau}} \right), x \in \Delta^{K-1} \tag{3.11}$$

With the softmax temperature τ approaching to 0, samples from the Gumbel-Softmax distribution become one-hot and the Gumbel-Softmax distribution becomes identical to the categorical distribution. Therefore, by replacing categorical samples with Gumbel-Softmax samples, we can use back-propagation to calculate the gradients.

3.4.3 Expression Generator

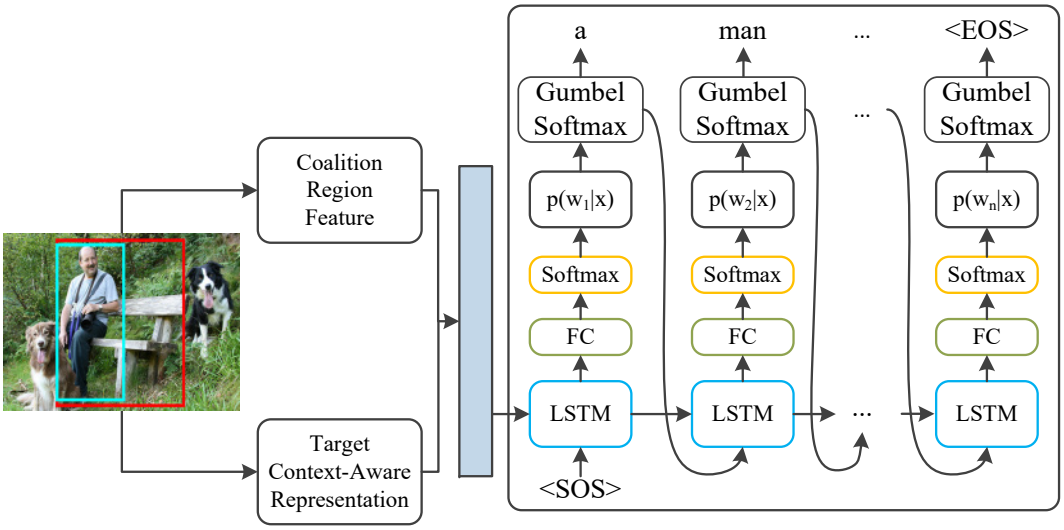

Figure 3.2: Architectural diagram of the generator. The generator employs LSTM-based language model to produce referring expressions, and the LSTM cells take the region context-aware representation and the coalition region feature as input to predict next word. The Gumbel Softmax updates gradients during adversarial training.

According to [140], the global context of an image improves the quality of generated captions. Motivated by this, we adopt the coalition region $b_c = (b_o \cup b_s)$ to be a supplementary region to leverage the global text. Figure 3.2 shows details of the expression generator G. We employ a LSTM [48] based language model to produce sequential words. The generator takes $V^{'}$ and b_c as inputs, where b_c is input to the LSTM at only the zeroth time step, and $V^{'}$ is input to the LSTM at

all time-steps. On top of the generator, we utilize softmax layer to compute the probability distribution over the vocabulary $p(w_t|w_{t-1}, x)$ at each step as follows:

$$\begin{aligned} h_t^G &= \text{LSTM}^G(w_{t-1}, V^{'}, h_{t-1}^G, c_{t-1}) \\ h_t &= \text{FC}(h_t^G) \\ p(w_t|w_{t-1}, x) &= \text{softmax}(h_t) \end{aligned} \tag{3.12}$$

where LSTM^G is the LSTM for generator, h_t^G is the hidden state of the generator LSTM, c_t represents the cell state at time t, w_t denotes the generated word at time step t, $t \in 1, ..., n$, n is the length of the generated sentence.

Unlike image synthesis, expression generation is a discrete procedure, i.e., it is a non-differentiable operation. In order to tackle with this issue, available solutions include: (1) recursively feed back the previously sampled word until to generate the end-of-sentence (EOS) token, and select the sentence with the highest probability as in work [33]; (2) use greedy search method like beam search. However, taking these discrete samples as input to the discriminator does not allow for back-propagation and achieve end-to-end training.

The Gumbel-Softmax approximation does not require auxiliary steps to approximate the gradients and can be deemed as a plug into the models as a differential node. In this thesis, we adopt the straight-through variation of the Gumbel-Softmax approximation [54] to be the output of the generator for sampling word sequence during the adversarial training.

3.4.4 Discriminator

Taking the region context-aware representation $V^{'}$ and a set of generated expressions $E_p = e_1, ..., e_p$ by the generator as input, the discriminator aims at classifying whether the E_p is a real sentence from the corpus or is generated by the generator. Besides, the discriminator should encourage the generator to generate more diverse expressions during adversarial training. To this end, we design the discriminator making decisions on the criteria in two perspectives. First, the discriminator should

guarantee the generated expressions E_p describe the image regions correctly. Second, the discriminator should endeavor to generate expressions with high diversity.

In order to generate more diverse referring expressions, we employ two different critics in the discriminator, i.e. validity critic and diversity critic. The validity critic calculates the distance between the object region and generated expression e_i, $i \in (1, \mathrm{p})$. We encode the generated expressions E_p into length-uniformed vector $v_e \in \mathbb{R}^M$ by an one layer LSTM, where M denotes the word number in each expression. We also process the object visual representation to a vector $v_o \in \mathbb{R}^M$ by a fully connected layer. The distance between v_e and v_o is formulated as follows:

$$\begin{aligned} v_e &= \mathrm{LSTM}(e_i) \\ v_o &= \mathrm{FC}(V^{'}) \\ d_{oe} &= \|v_o - v_e\|_2 \end{aligned} \tag{3.13}$$

where $\|\cdot\|_2$ denotes the Euclidean distance calculator.

The diversity critic computes the difference between the expressions. We use the same approach to process the e_i and e_j, $i, j \in (1, \mathrm{p})$. We calculate the cosine distance between the expressions vectors by:

$$\begin{aligned} v_{e,i} &= \mathrm{LSTM}(e_i) \\ v_{e,j} &= \mathrm{LSTM}(e_j) \\ d_{ee} &= \mathcal{D}(v_{e,i}, v_{e,j}) \end{aligned} \tag{3.14}$$

where $\mathcal{D}$ represent the cosine distance calculator.

The distance vectors d_{oe} and d_{ee} capture the correctness of the expressions for each object and the diversity of expressions, respectively. We concatenate the two vectors and feed into a softmax layer to generate the output probability.

3.4.5 Adversarial Training

We train the generator G and discriminator D alternatively. The discriminator D aims at classifying $E_p^r \in \mathbb{R}(x)$ as real and $E_p^g \in \mathbb{R}(x)$ as fake. Additionally, we also add some random variables E_p^f to the training samples for augmenting the diversity of generated expressions. We define the loss function of D as follows:

$$L(D) = -log(D(E_p^r, x)) - log(1 - D(E_p^g, x) - log(1 - D(E_p^f, x))) \tag{3.15}$$

The objective of the generator G is to coax the discriminator classifying E_p^g as real. We adopt an l_2 loss to match the expected value of distance vectors d_{ee} and d_{eo} between the real samples and the generated data. We difine the loss function of generator by:

$$L(G) = -log(D(E_p^g, x)) + \|\mathbb{E}[d_{ee}] - \mathbb{E}[d_{eo}]\|_2^2 + \|\mathbb{E}[d_{ee}] - \mathbb{E}[d_{eo}]\|_2^2 \tag{3.16}$$

where $\mathbb{E}$ denotes the expectation.

3.5 Experiments

We implement experiments on RefCOCO, RefCOCO+, and RefCOCOg to evaluate the introduced referring expression generation network, and we select seven evaluation metrics to assess the generated expressions.

3.5.1 Datasets

We train and validate the proposed generation network on RefCOCO, RefCOCO+, and RefCOCOg. For fair comparison, we use the same train/validation/test splits as [148] and [89].

3.5.2 Experimental Setup

We pretrain the generator via standard maximum likelihood training to generate consistent sentences. We also pretrain the discriminator to classify correct object-expression pairs. We found that, through the pretrain of the generator and the discriminator, the generative network can generate more coherent and human-friendly expressions.

	RefCOCO		RefCOCO+		RefCOCOg
	testA	testB	testA	testB	val
BLEU@1	73.67	75.58	63.48	49.83	42.75
BLEU@2	57.54	57.78	46.34	30.97	26.71
BLEU@3	41.32	41.36	30.33	18.24	16.50
BLEU@4	23.35	26.27	18.27	9.70	10.36
METEOR	30.67	33.82	23.62	21.93	16.48
ROUGE_L	65.30	69.57	56.42	50.20	39.01
CIDEr	84.16	133.2	66.82	79.71	77.02

Table 3.1: Performance of the proposed network on the three datasets under different evaluation metrics. All values are reported as percentage (%).

According to [82], the Gumbel temperature τ should be in the range of (0.1, 0.8). If τ is beyond the range, the training process is unstable. In practice, we set $\tau = 0.5$.

We train the generator using the RMSProp optimizer, and we set the initial learning rate to 1e-6, the decay rate to 0.999 and the smooth eps to 1e-8. We train the discriminator for 10 iterations for each generator update.

3.5.3 Results on the Three Datasets

We evaluate the generative network using multiple evaluation metrics. In image captioning, BLEU@N, METEOR, ROUGE, and CIDEr are standard metrics and have been widely used to assess the generated captions. We also adopt these metrics to evaluate our model. Table 3.1 lists the acquired results by the proposed network under different evaluation metrics. Because the length of the generated expressions, the introduced generative network obtains lower performance under the BLEU@N metrics.

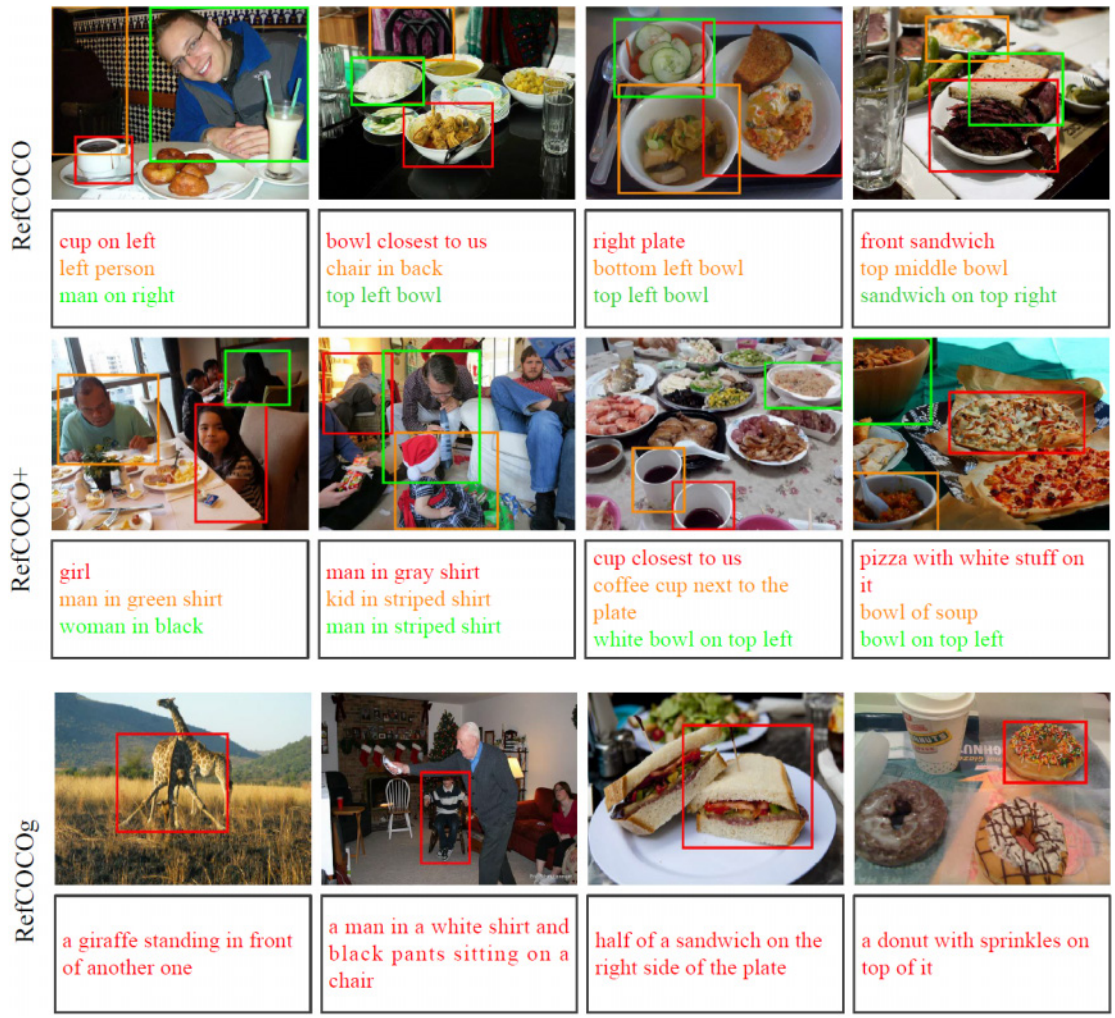

Figure 3.3: Generated expression examples on the test sets of RefCOCO, RefCOCO+, and RefCOCOg. The generated expressions are listed in rectangles. Each sentence shows the generated expression for each detected object within an image. Same color between an expression and bounding box of an object indicates correspondence.

Figure 3.3 shows some obtained example results of the referring expressions generation network on the test splits of the three datasets. As shown in the results, the generated expression of each target object is associated with the target and has a weak connection with the other objects within the image.

		RefCOCO				RefCOCO+				RefCOCOg	
		testA		testB		testA		testB		val	
		METEOR	CIDEr	METEOR	CIDEr	METEOR	CIDEr	METEOR	CIDEr	METEOR	CIDEr
1	visdif+tie[148]	18.9	-	24.9	-	15.0	-	14.3	-	15.1	-
2	attr+visdif[78]	22.2	-	25.8	-	15.5	-	15.5	-	16.0	-
3	SLR+MMI+rerank[149]	29.6	77.5	34.0	132.0	21.3	52.0	21.5	73.5	15.9	66.2
4	SLR+rerank[124]	**31.3**	83.7	**34.1**	132.9	**24.2**	66.4	**22.8**	78.7	**17.0**	**77.7**
5	our	30.67	**84.16**	33.82	**133.2**	23.62	**66.82**	21.93	**79.71**	16.48	77.02

Table 3.2: Comparison with the state-of-the-art approaches. All values are listed as percentage (%).

3.5.4 Comparison with State-of-the-art

Table 3.2 presents the performance of different models on the three referring expression datasets. The exiting methods do not employ BLEU@N and ROUGE_L to evaluate their performance, so we just list the METEOR and CIDEr results in the Table 3.2.

Additionally, it is worth noting that the existing work generates referring expressions by adopting the popular CNN-LSTM paradigm and utilizes different region visual representation. Specifically, [148] and [149] extract visual features from VGG-fc7 [119] and append global context representation, location feature, location difference representation, and visual appearance difference representation as the object visual representation. [124] uses the feature from the last convolutional layer of the fourth stage of the ResNet 152 [46], and affixes the global context substitution representation and the other four representations same as [148] to represent the detected object. [78] utilizes the concatenation of the region feature extracted from VGG-fc7 and the location representation, and distill pretrained attribute to generate referring expressions.

In our experiments, we extract the deep feature from the last convolutional layer of the fourth stage of ResNet 101 for the detected object and append the coalition region representation. Overall, the results across seven evaluation metrics indicate that the proposed generation network acquires better performance than the existing methods.

3.6 Discussion

In this chapter, we proposed a generative network for referring expression generation. Unlike the existing methods which adopt the popular CNN-LSTM paradigm to produce referring expressions, we aimed to generate diverse and natural expressions via adversarial training. We introduced a discriminator with two different critics to classify whether the generated expressions are real or fake, this approach also prompted the generator to generate expressions with more diversity and naturalness. Moreover, we conducted experiments to evaluate the performance of our referring expression generation network.

In this thesis, we attempted to generate diverse and natural referring expressions via adversarial training. Although the introduced network obtains promising results on three datasets, we will improve the network to generate more human-friendly referring expressions. Subject to the scale of the referring expressions datasets, we will exploit a generative approach for dense captioning which aims at generating descriptions for each detected visual elements within images, and the published dataset is sufficient to train a model to deal with the diverse and complicated human environments.

Chapter 4

Object Affordance Recognition via Attention-based Multi-Visual Features Fusion

4.1 Introduction

When new objects come into our sight, we can deduce their function according to multiple visual properties, such as shape, size, color, texture, and material. The capacity to infer functional aspects of objects or object affordance is crucial for us to describe and categorize objects more easily. Affordance is widely used in multiple tasks, [125] fuses visual features and affordance to improve robustness for sensorimotor object recognition, [12] demonstrates affordance could improve the quality of natural HRI, [86] utilizes affordance to prompt a robot to understand human spoken instructions.

Psychologist James J.Gibson initially introduced affordance in 1976 [41]. He suggested that affordance encodes the "action possibilities" in the environment for a given agent. While cognitive psychologist Don Norman discussed affordance from the design perspective [94] as:

"*The term affordance refers to the relationship between a physical object and a*

person (or for that matter, any interacting agent, whether animal or human, or even machines and robots). An affordance is a relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used."

Don Norman argued that affordance refers to the fundamental properties of an object which determines how the object could possibly be used. According to Norman's view, drinks afford *drinking*, eating utensils afford *eating*, and readings such as text documents are for *reading*.

Following Norman's standpoint, in this thesis, we generalize ten affordances (*calling*, *drinking(I)*, *drinking(II)*, *eating(I)*, *eating(II)*, *playing*, *reading*, *writing*, *cleaning*, and *cooking*) for objects that are commonly used in indoor environments. Although drinkware and drinks can be used for drinking, the drinkware affords different function with drinks, i.e., the affordance of drinkware is different from drinks. The same situation also exists between foods and eating utensils. Therefore, different labels are utilized to discriminate the different affordance between drinkware and drinks, i.e., *drinking(I)* denotes the affordance of drinkware, *drinking(II)* is for drinks, *eating(I)* is for eating utensils, and *eating(II)* is for foods, respectively.

Most of the existing work adopts mono features, such as geometric features [61], visual attributes [158] or deep features extracted from a pretrained CNN [93] to recognize the object affordances. Even though these presented frameworks achieved substantial results for recognizing object affordances, the mono feature is not sufficient to recognize the affordance in some situations, for example, the features from a partially occluded object may downsize the recognition accuracy. It is clear that different features may indicate different task-relevant information and can be complementary to complete a given task. Moreover, the existing approaches do not pay attention to the multi-visual features that can improve affordances recognition.

Inspired by the complementary nature of the multiple features, we adopt multi-visual features, deep visual features extracted from a pretrained CNN and deep texture features encoded by a deep texture encoding network, to learn the human-

centered object affordances. The primary issue of fusing multi-visual features is that the fusion scheme should reserve the complementary nature of the features. Fusing different features through naive concatenation may fail to learn the relevance of multiple features, bring about redundancies and may lead to overfitting during the training period. Consequently, to preserve the complementary nature of the multi-visual features in the process of affordance learning, the interaction information between the multi-visual features is employed and an attention-based architecture is proposed to fuse the multi-visual features. Additionally, no dataset is published for recognizing human-centered object affordances. Therefore, a dataset in which a large portion of images originate from MSCOCO and ImageNet is collected to learn the aforementioned human-centered affordances.

To summarize, we propose an attention-based multi-visual features fusion architecture to fuse the deep visual features and deep texture features for learning object affordances. To the best of our knowledge, this work is the first attempt to combine the multi-visual features to recognize human-centered affordance. Accordingly, the contributions of this work involve: (1) a first attempt to fuse the deep visual features and deep texture features for learning human-centered object affordances; (2) an attention-based multi-visual features fusion architecture; (3) a dataset collected for learning object affordances. We conduct extensive experiments on the self-built dataset to train and validate the introduce object affordance recognition network, and the experimental results show that the proposed attention-based multi-visual fusion network outperforms features naive concatenation, feature extracted from VGG, RetinaNet, and YOLO V3.

4.2 Related Work

4.2.1 Object Affordance

Existing work utilizes multiple approaches to infer object affordances. [123] predicts object affordances through human demonstration, [61] deduces affordance

through extracted geometric features from point cloud segments, [158] reasons affordance through querying visual attributes, physical attributes, and categorical characteristics of objects in a pre-built knowledge base. [88] perceives affordance from local shape and geometry primitives of objects. These methods adopted visual characteristics or geometric features to infer object affordances, so the scalability and flexibility of these approaches are limited.

Several recently published methods adopt deep learning-based approaches to detect object affordance. [27] proposes a denoising auto-encoder to actively learn the affordances of objects and tools through observing the consequences of actions performed on objects and tools. [108] uses multi-scale CNN to extract mid-level visual features and combines them to segment affordances from RGB images. Unlike [108], [110] regards affordance perception as semantic image segmentation and adopted a deep CNN based architecture to segment affordances from weakly labeled images. [92] extracts deep features from a CNN and adopts an encoder-decoder architecture to detect affordances for object parts. [86] utilizes the deep features extracted from different convolutional layers to recognize object affordances. [93] combines an object detector, a CNN with dense conditional random fields to detect object affordance from RGB images. Similar to [93], [32] also utilizes deep features and employs a universal object detection framework and deconvolution to recognize object affordance.

The aforementioned work utilized the geometric features or the deep features extracted from a pretrained CNN to infer object affordance, and did not take into consideration that the features from another source can be applied to improve affordance recognition accuracy. Unlike the work mentioned above, we utilize multi-visual features to learn the object affordances.

4.2.2 Multiple Features Fusion

The fundamental purpose of multi-visual feature fusion is to enhance model performance by exploiting the complementary information of different features. According to fusion approaches, the fusion frameworks are divided into three types,

i.e., early fusion, late fusion, and intermediate fusion [129]. In early fusion, the original features are obtained from different sensors and integrated into a single representation vector, so the early fusion is often deemed to be data fusion or multisensor fusion. A crucial issue of early fusion is that the dimension of fused vectors may be huge and may contain redundancies [101]. To reduce the dimension and redundancies, [85] and [132] apply a hash-based framework to learn compact multimodal representations for the data from different modalities.

Late fusion, is also known as decision level fusion, refers to the integration of features extracted from different modalities by deep learning-based models. Moreover, according to [118] and [137], the late fusion acquires better results than the early fusion. The intermediate fusion constructs a shared representation layer to merge the learned features using deep neural networks (DNN). Unlike the early fusion and late fusion, the primary advantage of intermediate fusion is its flexibility and excellent performance due to the integration with DNN architectures, as demonstrated in [53] and [90].

Although the schemes mentioned above acquired promising results regarding various practical problems, the interaction information of the multiple features is ignored. [106] proposes Factorization Machines (FM) which can model interactions between different features via factorized parameters, and has the capability to assess the interactions from sparse data. Moreover, [6] initially introduces an attention mechanism to acquire different weights for different parts of input features, and can automatically search the most relevant parts to acquire better results from source features. Due to its performance, the attention mechanisms have been widely employed in multiple research fields, such as image captioning [139], visual question answering (VQA) [10], video description [49], etc.

Inspired by [106] and [49], we propose an attention-based architecture to fuse the deep visual features and deep texture features through a soft attention mechanism. The introduced fusion architecture takes the sparse representations of the multi-visual features as input and achieves attention-based dynamic fusion of the multi-visual features.

4.3 Proposed Method

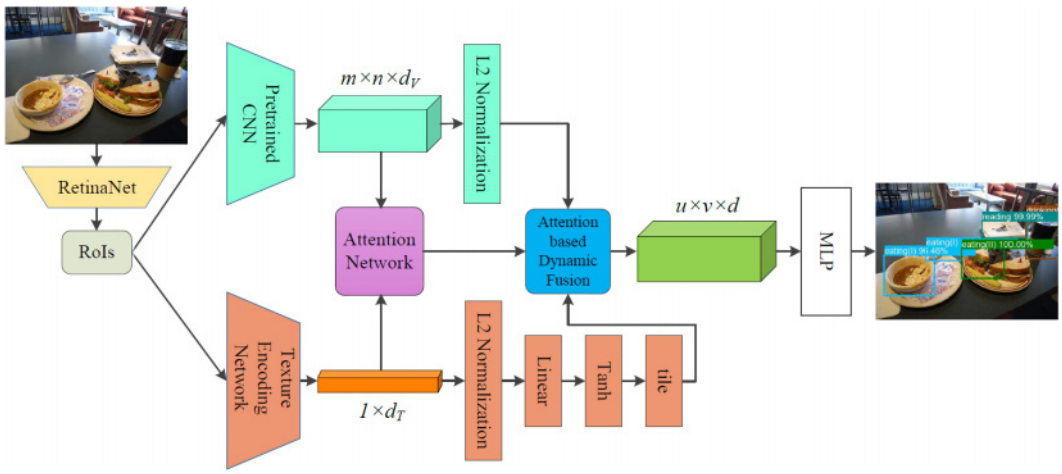

Figure 4.1: Architectural diagram of the object affordance detection via attention-based multi-visual features fusion. The RetinaNet is adopted to detect RoIs from raw images, and then for each detected RoI, the deep visual features and deep texture features are extracted by a pretrained CNN and a texture encoding network, respectively. In order to reserve the complementary nature of the different features and avoid causing redundancies during the multi-visual features fusion, an attention-based fusion mechanism is applied to fuse the multiple visual features. Through the attention-based fusion, the fused features are fed into an MLP to learn object affordances.

Following Norman's viewpoint, we generalize ten affordances for ordinary household objects, and we propose an attention-based multi-visual features fusion architecture, which can be trained end-to-end, to learn the human-centered affordances. Figure 4.1 illustrates the details of the proposed multi-visual features fusion architecture. The presented architecture is composed of a RoIs detection network (RetinaNet), a deep features extraction module, an attention network, an attention-based dynamic fusion module, and an MLP. Two different deep networks are employed to extract the multi-visual features, the attention network learns dynamic attention weights through the sparse representations of the extracted multi-visual features, while the dynamic fusion module fuses the multi-visual features by

integrating them with the generated attention weights, and the MLP is applied to learn the object affordances. We introduce the details of each component of the proposed architecture in this section.

4.3.1 Deep Features Extraction

Deep Visual Feature Extraction

RetinaNet [74] acquires better detection accuracy on MSCOCO [75] than the state-of-the-art two-stage detectors. Considering the performance of RetinaNet, we adopt Retinanet to generate RoIs from raw images. And several pretrained CNN models, such as AlexNet [66], VGGNet [117] and ResNet [46], can be applied to extract deep visual features from RGB images. In this work, we adopt VGGNet to extract deep feature for detected RoIs. The deep visual feature f_v is extracted by a pretrained CNN for each RoI I_R:

$$f_v = CNN(I_R) \tag{4.1}$$

where $f_v \in \mathbb{R}^{m \times n \times d_v}$, $m \times n$ denotes the size of the extracted deep features, d_v is the output dimension of the CNN layer. In order to improve learning dynamics and reducing training time, L_2 normalization is utilized to process the extracted deep visual features.

Deep Texture Feature Extraction

Multiple presented texture recognition networks can be used to encode texture features, e.g., [20] generates texture features through Fisher Vector pooling of a pretrained CNN filter bank, [155] proposes a texture encoding network for material and texture recognition, the texture encoding network encodes the deep texture features through a texture encoding layer which is integrated on top of convolutional layers and is capable of transferring CNNs from object recognition to texture and material recognition. Furthermore, the texture encoding network achieves state-of-the-art on the material dataset MINC2500 [9]. Due to the good performance of the texture encoding network introduced in [155], we select it to

encode the texture feature for each detected RoI and convert the texture feature to vector $\mathbf{v}_t$:

$$\mathbf{v}_t = TexNet(I_R) \tag{4.2}$$

where $\mathbf{v}_t \in \mathbb{R}^{1\times d_t}$, d_t is the output size of the texture encoding network.

The extracted texture vector $\mathbf{v}_t$ is also processed by L_2 normalization. For modeling convenience, a single perceptron which is comprised of a linear layer and a tanh layer is employed to transform $\mathbf{v}_T$ into a new vector:

$$\hat{\mathbf{v}}_t = tanh(W\mathbf{v}_t + b) \tag{4.3}$$

where $\hat{\mathbf{v}}_t \in \mathbb{R}^{1\times d_l}$, W is a weight matrix, b denotes a bias vector for the linear layer, and d_l is the dimension of the linear layer. From [10] and the experimental results, hyperbolic tangent produces slightly better results.

For fusing convenience, we adopt tile operation to expand the texture vector $\hat{\mathbf{v}}_t$ to generate the deep texture representation f_t which has the same dimension with the deep visual feature f_v, i.e., the generated $f_t \in \mathbb{R}^{m\times n\times d_v}$.

4.3.2 Attention-based Multi-visual Features Dynamic Fusion

Factorization Machines were proposed for recommendation system [106], and aimed at solving the problem of feature interactions under large-scale sparse data. Given feature vector list, FM predicts the target through modeling all interactions between each pair of features:

$$\hat{y}(x) = w_0 + \sum_{i=1}^{t} w_i x_i + \sum_{i=1}^{t}\sum_{j=i+1}^{t} \hat{w}_{ij} x_i x_j \tag{4.4}$$

where $w_0 \in \mathbb{R}$ is the global bias, x_i and x_j denote the i-th and j-th feature in the given feature list, $w_i \in \mathbb{R}^t$ is the weight of x_i, $\hat{w}_{ij}$ models the interaction between x_i and x_j and is calculated by:

$$\hat{w}_{ij} = \mathbf{v}_i^T \mathbf{v}_j \tag{4.5}$$

where $\mathbf{v}_i$, $\mathbf{v}_j \in \mathbb{R}^s$ are the sparse representations, i.e., embedding vectors for the nonzero elements of x_i and x_j, s denotes the dimension of the embedding vectors.

Although the interaction information of the features can be modeled by $\hat{w}_{ij}$, the different weights of the multi-visual features in the process of features fusion cannot be achieved by FM. To this end, we draw support from soft attention mechanism to achieve weighted dynamic fusion.

In light of the FM, the $\hat{w}_{ij}$ comprises the interaction information of different features, and is represented by the sparse nonzero elements of the different features. Formally, we extract the nonzero element set from f_v and $\mathbf{v}_t$, and adopt an embedding layer to acquire the sparse representations e_v for f_v and e_t for $\mathbf{v}_t$, respectively. We calculate the interacting matrix k_{vt} which embeds the interaction information between f_v and $\mathbf{v}_t$ by:

$$k_{vt} = e_v^T e_t \tag{4.6}$$

where $k_{vt} \in \mathbb{R}^{p \times p}$, e_v and $e_t \in \mathbb{R}^{1 \times p}$, p denotes the output size of the embedding layer.

In order to avoid causing information redundancies during features fusion, we integrate an attention mechanism with k_{vt} to complete features fusion. By learning attention weights, the attention mechanism endows the model with the ability to emphasize the different weights of the multi-visual features during learning affordance. And the attention weights can be parametrized by an attention network which is composed of an MLP and a softmax layer. The input of the attention network is the interacting matrix k_{vt}, the generated weight encodes the interaction information between the different features. The attention weights τ_{att} can be acquired by:

$$\tau_{att} = \frac{exp(A_{vt})}{\sum exp(A_{vt})} \tag{4.7}$$

and

$$A_{vt} = \alpha^T tanh(W_{att} k_{vt} + b_{att}) \tag{4.8}$$

where $\tau_{att} \in \mathbb{R}^{1 \times p}$, W_{att} , b_{att}, and α are weight matrices, bias vector and model parameters for the attention network, respectively.

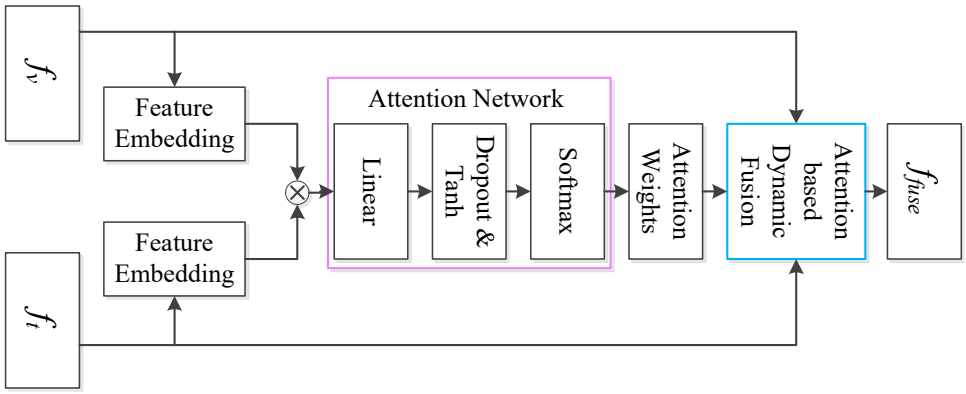

Figure 4.2: Attention-based multi-visual features fusion network. The feature embedding layers process the sparse representations of the deep visual features and deep texture features, and the outputs of feature embedding layers are applied to generate the interaction information of the multi-visual features. Subsequently, the interaction information is fed into the attention network to acquire the attention weights, which are adopted to complete attention-based dynamic fusion.

By means of the learned τ_{att}, f_v and f_t are fused to produce feature f_{fuse} that is used to learn object affordances. The fused feature f_{fuse} is generated by:

$$f_{fuse} = (1 - \tau_{att}) f_v \oplus (\tau_{att}) f_t \tag{4.9}$$

where $f_{fuse} \in \mathbb{R}^{m \times n \times d}$, $\oplus$ denotes concatenation. Figure 4.2 shows the details of the attention-based multi-visual features fusion.

4.4 Experiments

4.4.1 Dataset

In MSCOCO and ImageNet [109], there are only a few indoor scenes and few objects associated with the introduced ten affordances. Therefore, we create a dataset to train and evaluate the proposed object affordance recognition architecture. The proposed dataset[1] is composed of images collected by a Kinect V2 sensor, indoor scenes collected from MSCOCO and ImageNet datasets.

[1] `https://tams.informatik.uni-hamburg.de/research/datasets/index.php`

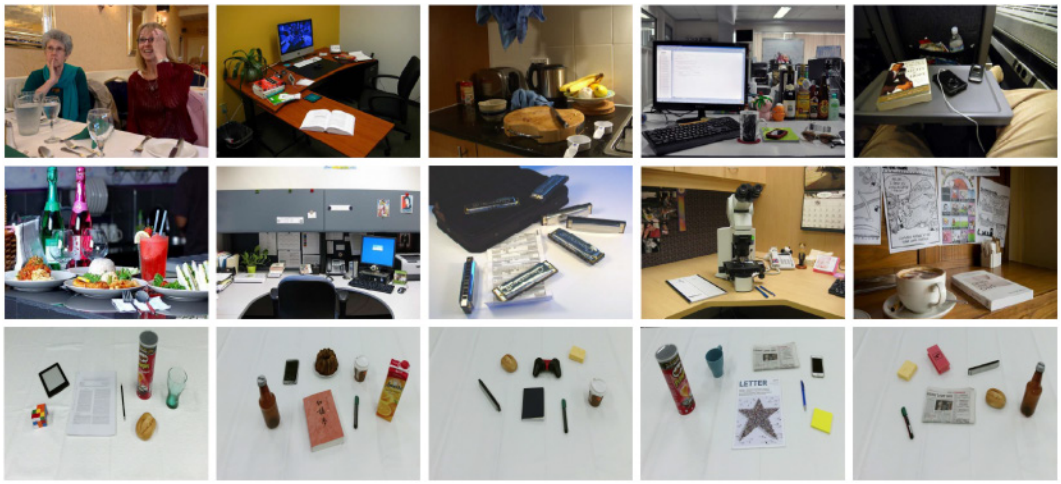

Figure 4.3: Example images of the proposed dataset.**Top row:** images from MSCOCO. **Middle row:** images from ImageNet. **Bottom row:** images taken by Kinect V2.

The dataset contains in total of 12349 RGB images and 14695 bounding box annotations for object affordance detection (in which 3378 annotations are from MSCOCO and ImageNet). 56.1% regions (8250) from the dataset are randomly selected for training, 22.1% regions (3253) are for validation and the remaining 21.8% regions (3192) are for testing. Figure 4.3 shows some example images of the proposed dataset.

As mentioned above, ten affordances are labeled which are related to ordinary household objects. Figure 4.4 illustrates the affordance distribution in the presented dataset. There are few *writing* and *cleaning* objects included in the images in the MSCOCO and ImageNet datasets, so a large portion of these two categories images are taken by a Kinect V2 sensor.

4.4.2 Experimental Setup

We utilize the available source[2] which is an implementation of RetinaNet [74], and employ ResNet 50 to be the backbone to detect RoIs from RGB images. Then, we extract the deep visual features from the last pooling layer of VGG19 [117]

[2]`https://github.com/fizyr/keras-retinanet`

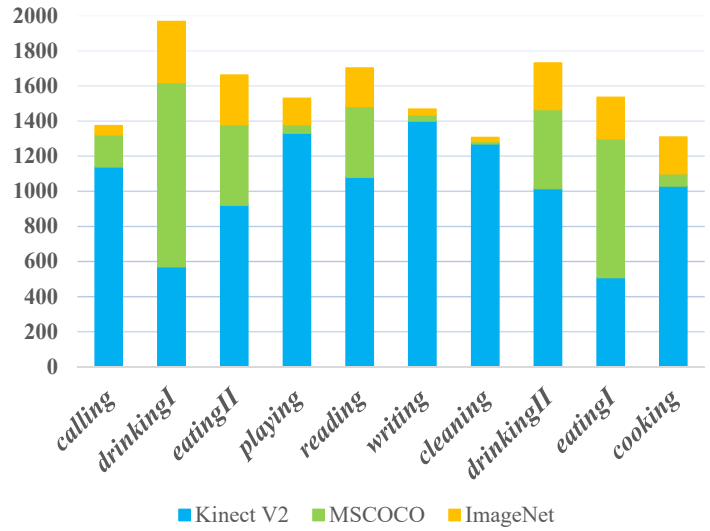

Figure 4.4: The affordance distribution in the presented dataset. Y-axis denotes the region number of each affordance.

trained on Imagenet [109]. To produce a length-uniformed feature map for RoIs with different size, we rescale the detected RoIs to 224×224 pixels. Accordingly, the dimension of the extracted deep visual feature for each RoI is 7×7×512, i.e., $f_v \in \mathbb{R}^{7\times7\times512}$.

The deep texture encoding network [155] trained on the material database MINC2500 is adopted to generate deep texture representations. The texture features are extracted from the texture encoding layer for RoIs. The output size of the texture encoding layer is 32×128, so the dimension of $\mathbf{v}_t$ is 1×4096. The output size of the single perceptron is set to d_l=512, therefore, the dimension of the transformed texture vector $\hat{\mathbf{v}}_t$ is 1×512. Through the tile operation, the dimension of the generated deep texture representation $f_t \in \mathbb{R}^{7\times7\times512}$.

For modeling convenience, the size of the embedding layer is set to $p = 512$, the generated sparse representation for the deep visual feature and the deep texture feature, e_v and e_t, are vectors with the dimension of 1×512, and the dimension of produced interacted matrix $k_{vt} \in \mathbb{R}^{512\times512}$. The produced k_{vt} is fed into the attention network, so the size of the generated attention weights $\tau_{att} \in \mathbb{R}^{1\times512}$. Through the attention weights based dynamic fusion, the dimension of fused feature f_{fuse} is

7×7×1024, i.e., $f_{fuse} \in \mathbb{R}^{7\times 7\times 1024}$.

The fused features are fed into the MLP to learn affordances. The parameters of the MLP include: Cross Entropy loss function, Rectified Linear Unit (ReLU) activation function, and Adam optimizer. The structure of the MLP is 50176-4096-1024-10. In practice, the standard error back-propagation algorithm is adopted to train the model. The learning rate is set to 0.0001 and batch size to 32. in order to prevent overfitting, dropout is employed to randomly drop 50% neurons during training.

4.4.3 Results

The architecture is trained in PyTorch. After 100 epochs training, the proposed network acquires 61.38% average accuracy on the test set. Fig.4.5 shows the confusion matrix of the acquired results by the proposed network, and Figure 4.6 shows some acquired example results of object affordance detection on the test set.

From Figure 4.5, the affordances *writing*, *cleaning*, and *cooking* have relative low accuracies compared to the other affordances. The shapes and textures of the selected objects in the three categories are significantly different from each other. Therefore, we deduce the primary cause that leads to the low accuracy of the three affordances is the great shape and texture differences, so that the similarities between the deep features in one category are difficult to generalize and learn.

4.4.4 Ablation Study and Comparison Experiments

Except validating the attention-based multi-visual features fusion network on the presented dataset, we also compare the results acquired by different deep features, different features fusion approach, and different networks to demonstrate the performance of the proposed affordance detection network.

VGG19 Deep Features: In order to verify the effectiveness of the multi-visual features fusion for object affordances learning, the results generated by the attention-based fusion network and a model trained by the deep visual features

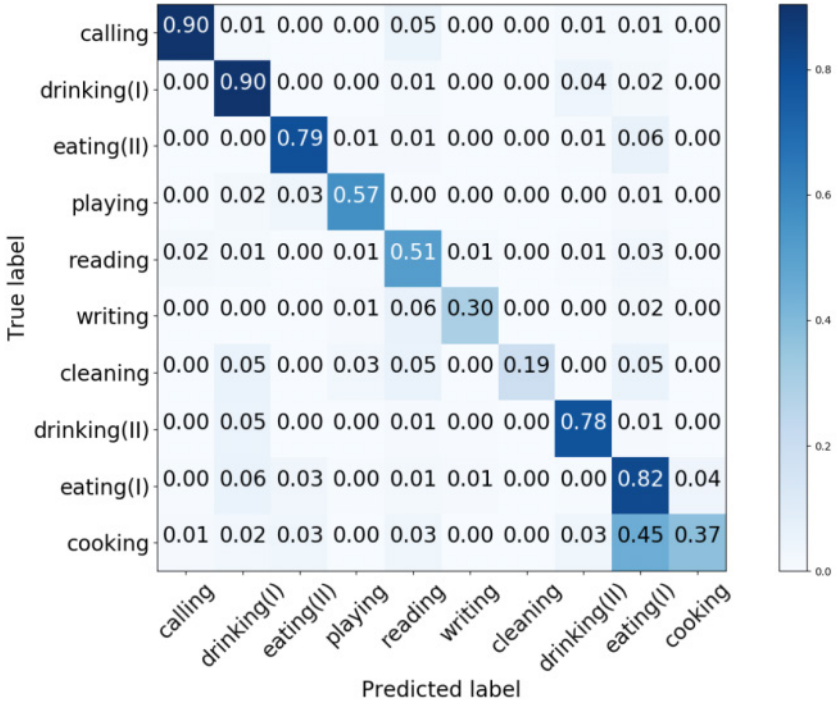

Figure 4.5: Generated confusion matrix of object affordance detection on the test set.

extracted from VGG 19 is compared. In this case, the deep features with the shape of 7×7×512 are fed into an MLP with the structure of 25088-4096-1024-10 to learn the affordances. After 100 epochs training, the model acquires 55.54% on the test set.

Naive Concatenation: For validating the performance of attention-based fusion scheme, the deep visual features and the deep texture features are naive concatenated to generate the fused representations of the multi-visual features. The concatenated features are with the shape of 7×7×1024, which are fed into the MLP with the same structure in the multi-visual fusion architecture to recognize affordances. After 100 epochs, the generated model acquires 58.21% on the test set.

RetinaNet: We also directly train the RetinaNet [74] (available source[2]) on the proposed dataset. For a fair comparison, the backbone also utilizes ResNet 50. After 100 epochs training, the RetinaNet obtains 58.92% average accuracy on the

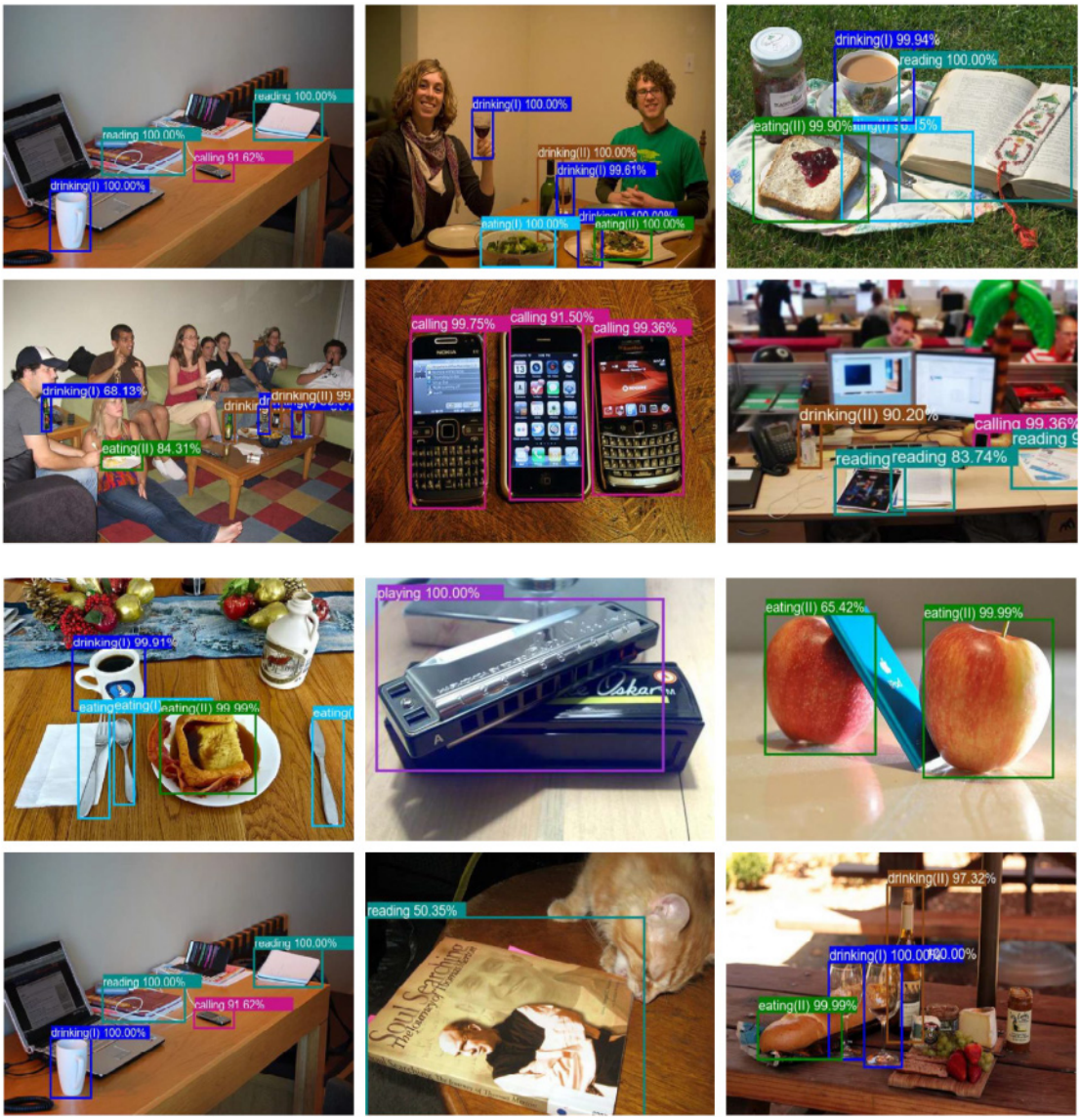

Figure 4.6: Example results of object affordance detection on the test set.

test set.

YOLO V3: We also adopt the original pretrained weights to train YOLO V3 [103] (available code[3]) on the dataset. After 100 epochs training, the YOLO V3 model obtain 49.63% average accuracy on the test set. Table 4.1 lists the results acquired by the different deep features, different feature fusion mechanism, and the different networks.

[3] `https://github.com/qqwweee/keras-yolo3`

	Attention Multi-Visual Features Fusion	VGG Deep Features	Naive Concatenation	RetinaNet	YOLO V3
calling	0.9036	**0.9096**	0.8723	0.7747	0.5783
drinkingI	**0.8991**	0.7785	0.8195	0.7806	0.4771
eatingII	**0.7943**	0.7658	0.7569	0.6829	0.5696
playing	0.5676	0.4791	0.5305	**0.8305**	0.7871
reading	0.5148	0.4938	0.5297	**0.6424**	0.652
writing	**0.2995**	0.2028	0.286	0.2628	0.2028
cleaning	0.1875	0.1625	0.175	**0.375**	0.3327
drinkingII	**0.7838**	0.7627	0.7248	0.6128	0.5824
eatingI	**0.8162**	0.7103	0.7049	0.6738	0.4837
cooking	0.3719	0.2893	**0.4214**	0.2562	0.2968
Average	**0.6138**	0.5554	0.5821	0.5892	0.4963

Table 4.1: Object affordance detection results acquired by the proposed network, VGG deep features, multiple feature fusion via naive concatenation, RetineNet, and YOLO V3.

From the experimental results, the accuracies of affordance classes *writing*, *cleaning* and *cooking* acquired by the four different approaches are relatively lower than the other affordances. Nonetheless, our architecture acquires the best recognition accuracies on five affordance categories and the best average accuracy on the test set. The results demonstrate the multi-visual features and attention-based fusion improve the model performance for learning object affordances.

4.5 Discussion

We presented an attention-based multi-visual features fusion architecture to learn human-centered object affordances. Different from the existing affordance detection frameworks, our fusion architecture fused deep visual features and deep texture features to recognize object affordances from RGB images. The attention-based fusion architecture, which took into account the interaction of the multi-visual

features, preserved the complementary nature of the multi-visual features extracted from different networks and avoided producing information redundancies during features fusion. We trained and validated the proposed attention-based multi-visual features fusion network on our self-built dataset, and the experimental results demonstrated the effectiveness of multi-visual features and attention-based fusion for learning affordances.

Currently, the introduced architecture learns ten affordances through fusing the deep visual features and the deep texture features. In the future, we will employ meta-learning to learn more affordances from a smaller amount of annotated images, and employ a network-based framework to learn the different contributions of the different features for object affordances learning.

Chapter 5

Interactive Natural Language Visual Grounding

5.1 Introduction

Human beings often refer to objects in the physical world when they have interactions with others, and they can interpret the other's motivation even though many details are omitted in utterances. Naturally, we anticipate intelligent agents have the ability to interact with human users using the most intuitive and effective pattern, understand natural language instructions, and carry out assigned tasks.

Natural language and vision are the two crucial ways to exchange information in our daily life, and are also the essential channels to achieve communication between humans and intelligent agents. Bridging the two domains has been attracting substantial research attention [67, 112, 47, 79, 147]. Natural language visual grounding is a challenging task which aims at locating target objects within visual images or scenarios according to given natural language queries. Moreover, natural language visual grounding can establish a natural communication channel to facilitate the interaction between humans, physical environments, and intelligent agents.

A representative application of natural language visual grounding is natural language-based HRI. Natural language-based HRI has been attracting considerable research attention, and a number of approaches have been proposed [97, 59, 86, 115,

44, 1, 96]. Since the properties of natural language visual grounding, in addition to the applications in robotics, it has been widely used in VQA [71, 157], visual image search [43], visual chatbot [26].

Natural language visual grounding requires a comprehensive understanding of natural language queries and visual scenarios, and the pivotal issue is to locate the referred objects in working scenarios according to given instructions. In real applications, natural language queries are complicated and ambiguous, and visual scenarios are also sophisticated. Although the existing models achieve promising results, some of them either do not take into consideration the inherent ambiguity of natural language [97, 59, 86, 96], or alleviate the ambiguity via dialogues between human users and robots [115, 44, 1]. However, dialogue systems entail time cost and cumbersome interactions.

A crucial aim of this thesis is to achieve natural language visual grounding without auxiliary information, such as dialogues, gestures. Motivated by the roles of referring expression comprehension and referring expression generation, we propose two architectures in which draw support from referring expression comprehension and referring expression generation to ground natural language queries.

Considering the richness and diversity of natural language and the relatively simple expressions in the three referring expression datasets (RefCOCO, RefCOCO+, and RefCOCOg), we integrate the trained referring expression comprehension and referring expression generation models with scene graph parsing to ground complicated natural language queries. Formally, we first employ scene graph parsing to parse the sophisticated natural language instructions into scene graph legends, and combine the parsed scene graph legends with the referring expression comprehension and referring expression generation models to achieve unconstrained and sophisticated natural language commands grounding.

Referring expression-based approaches can ground explicit natural language queries, but referred objects embedded in intention-related natural language instructions can not be located via the referring expression-based approaches. Inspired by the affordance and its application in HRI [116], [86], we introduce an

intention-related natural language queries grounding architecture based on object affordance detection.

In order to ground the intention-related natural language queries, we introduce an intention semantic extraction module and integrate it with the object affordance detection to achieve intention-related natural language queries grounding. Specifically, we first extract verb words which embed intention semantics from intention-related natural language queries, and then by calculating the semantic relatedness between the extracted semantic verbs and detected object affordances to locate the referred target objects within working scenarios.

5.2 Related Work

5.2.1 Natural Language Understanding for HRI

With applications of robots becoming omnipresent in varied human environments, such as factories, hospitals, and homes, natural language understanding for HRI attracts great research interest. Researchers have proposed different approaches and representation formalisms to parse and understand deep semantics of natural language. [111] presents a flexible system for robust natural language interpretation to facilitate natural HRI in a domestic service robotics domain. The introduced system first syntactically pre-processes the given utterance into an internal representation, and then adopt decision-theoretic planning to acquire the most likely interpretation of the utterance. [8] introduces a discriminative approach, which integrates a standard linguistic pipeline with discriminative learning and distributional semantics, to understand the spoken natural language. Moreover, this work combines grounded information with a learning algorithm to improve the performances of natural language understanding.

The authors of work [7] introduce two kinds of existing natural language understanding approaches for HRI, i.e., grammar-based approaches and data-driven approaches. [120] introduces Combinatory Categorial Grammar (CCG), and de-

scribes the polynomial-time parsing algorithm. CCG has well-defined connections between syntax and semantics, and the λ-term is adopted to represent the semantics. [122] and [121] present Fluid Construction Grammar (FCG), in which the key component is an FCG-interpreter. The FCG-interpreter carries out basic operations, such as syntactic parse, production, linguistic aid research, and so on. [38] introduces Embodied Construction Grammar (ECG), which utilizes a precise formalism and technical notation to present the grammar and meaning of natural language queries. [35] employs ECG to parse the deep semantics in natural language and integrates with Robot Operating System (ROS) to prompt multiple robots to understand natural language. However, these systems focus on a constrained domain and are difficult to broaden and extend [135], and the process to build a grammar-based system is a tough business.

Data driven-based methods mainly employ Statistical Learning (SL) to address natural language understanding. [39] proposes a deep question answering (DeepQA) architecture and an artificial intelligence system Watson, which can answer questions in natural spoken language. The performance of Watson on the Jeopardy quiz show indicates its promising accuracy, high question processing speed, and strong confidence. [14] exploits a general SL-based framework to interpret navigation instructions given only sample observations of humans following such instructions.

The authors of work [136] adopt the Dempster-Shafer (DS) theory for inferring the intentions from human utterances in a specific context and generating utterances from intentions in contexts. In the presented system, the semantic interpretation is passed to a new component for a pragmatic inference that uses contextual and general knowledge to unearth the intention underlying the literal semantics. [77] develops an object functional role perspective method to enable the robot has the ability to understand the comprehensive behavior of human beings. The role-based method is adopted to model the human user's cognitive process during task performing by analyzing object selection. This work enables a robot to know how and why the human is doing, rather than only to help the robot rec-

ognize what the human is doing. Nonetheless, the data driven-based approaches have the data sparseness problem, and the significant barrier is the demand of a large amount of labeled training data. These attributes constrain the applications of the data driven-based approaches in real-world [135].

5.2.2 Natural Language Visual Grounding for HRI

Natural language provides an intuitive and natural interaction channel between human beings and robots. And multiple approaches are proposed to address natural language grounding for HRI. [97] proposes a probabilistic model named adaptive distributed correspondence graph to understand abstract spatial concepts, and introduces an approximate inference procedure to realize concrete constituents grounding. [96] utilizes a distributed correspondence graph to infer the environment representation in a task-specific approach. [59] introduces a statistical semantic mapping method that enables the robot to connect multiple words embedded in spoken utterance with a place in a semantic mapping processing. However, these existing methods do not take into consideration the inherent vagueness of natural language queries. [86] first presents an object affordances detection model, and then integrates the object affordances detection with a semantics extraction module for grounding intention-related spoken language instructions. This model subjects to limited classes of affordance, so it can not ground unconstrained natural language commands.

Shridhar *et al.* [115] adopt a pretrained captioning model DenseCap [56] to generate expressions for each detected region within uncluttered working scenarios, and by conducting K-means clustering to identify the relativeness of input instructions and the generated expressions. The expressions generated by [56] do not include the interaction information between objects, such as the spatial relationship with each other, so [115] employs gestures and dialogs with human users to handle ambiguity in spoken instructions. [44] draws support from a referring expression comprehension model [149] to identify target objects, and tackles with the ambiguity of spoken instructions via the referred object solely defined conversations

between human users and robots. [1] employs hourglass network [91] to generate position heatmaps for input images, and combines the generated heatmaps with a question generation module to find referred objects. [126] translates the spoken instructions into discrete robot actions, and through clarification dialog to improve objects grounding. Nevertheless, dialogue systems make the interaction between human users and robots cumbersome and time-consuming.

Thomason *et al.* [127] take into account visual, haptic, auditory, and proprioceptive data to predict target objects, and the natural language grounding supervised by an interactive game. However, this model needs to gather language labels for objects to learn lexical semantics. [83] presents a multimodal classifier generative adversarial network to identify target areas according to given linguistic commands, task context, and scene context.

Unlike the approaches mentioned above, we propose three architectures to address natural language visual grounding. Specifically, we integrate the referring expression comprehension and referring expression generation network with scene graph parsing to ground complicated natural language queries. We also combine the object affordance detection network with an intention semantic extraction module to ground intention-related natural language commands.

5.2.3 Natural Language Parsing

Extracting rich linguistic context from natural language sentences has a wide range of practical applications, including VQA [4], phrase grounding [138], and referring expression comprehension [21]. These methods employ dependency parsing to generate syntactic representations for questions, phrases, and referring expressions. Dependency parsing assigns a parent word to each word in a sentence, and each such connection is assigned with a label. In recent years, dependency parsing with neural networks acquires impressive performance [13, 114].

Scene graph was introduced in [57], in which the scene graph is used to describe the contents of a scene. A scene graph consists of nodes that represent an object with attributes and edges that expresses the connection and association be-

tween nodes. [112] adopts this paradigm and introduces the scene graph in natural language parsing. Compared with dependency parsing, scene graph parsing generates less linguistic compositions. Moreover, multiple vision tasks prove the value of scene graph generation, such as image retrieval [57], caption quality evaluation [2], etc. Inspired by the role of the scene graph, we adopt scene graph parsing to extract rich contexts from natural language queries to facilitate complicated and unconstrained natural language queries grounding.

5.3 Scene Graph Parsing

Given a natural language sentence, scene graph parsing aims to parse the natural language sentence into nodes and edges, where nodes comprise objects with their attributes, and edges express the relationships between objects. For instance, for the sentence "red apple next to the bottle", the output of scene graph parsing contains node ("red apple") and node ("bottle"), and edge ("next to").

Formally, a scene graph is defined as a tuple $\mathcal{G}(S) = (\mathcal{N}(S), \mathcal{E}(S))$, where $\mathcal{N}(S) = \{N_1(S), N_2(S), ..., N_n(S)\}$ is a set of nodes that encode objects with attributes, and $\mathcal{E}(S) = \{E_1(S), E_2(S), ..., E_m(S)\}$ is a set of edges that express the relationships between objects. Specifically, a node $N_i(S) \subseteq n_i \times \mathcal{A}_i$ represents attribute $\mathcal{A}_i$ of an object n_i (e.g., red apple). An edge $E_i(S) \subseteq (n_o \times R \times n_s)$ denotes the relationship R between a subject n_o and an object n_s (e.g., next to).

In general, a scene graph parser can be constructed on a corpus consisting of paired node-edge labels. However, no such dataset is released for natural language grounding. In order to ensure the precision of results acquired by the scene graph parser, we adopt a simple yet reliable rule, i.e., word-by-word match, to achieve scene graph alignment. Specifically, for a generated scene graph, we check the syntactic categories of each word in a node and an edge by part of speech. A correct parsed node should consist of a noun word or an adjective, and an edge contains adjective or adverb. In practice, we adopt the language scene graph [112] and the natural language toolkit [99] to complete scene graph generation and alignment.

5.4 Interactive Natural Language Grounding via Referring Expression Comprehension and Scene Graph Parsing

We present an natural language grounding architecture which combines the referring expression comprehension network with scene graph parsing to ground complicated natural language queries. Specifically, we parse the given queries into scene graph legends via a scene graph parser, and locate target objects within images by the trained referring expression comprehension network. Moreover, we validate the effectiveness of the introduced natural language grounding architecture on multiple household object scenarios with diverse natural language queries.

5.4.1 Architecture Overview

Natural language provides the most intuitive and natural interaction interface between humans and intelligent systems. For grounding unrestricted and complicated natural language queries in an end-to-end manner, we propose a novel architecture via referring expression comprehension and scene graph parsing as shown in Figure 5.1. We decompose the natural language grounding into two subtasks: 1) parse the natural instructions into scene graph legends by scene graph parsing. The scene graph legend is a data structure composed of nodes that denote objects with attributes, and edges that indicate the relationships between nodes; 2) ground the parsed natural language instructions by the referring expression comprehension network.

In this thesis, we aim to locate the target referents in working scenarios given natural language commands without auxiliary information. The inputs consist of a working scenario given as an RGB image and a natural instruction given as text, and the outputs are the bounding boxes of target objects. We generate scene graph legends for the input natural language instructions by scene graph parsing, and we ground the parsed scene graph legends via the trained referring expression

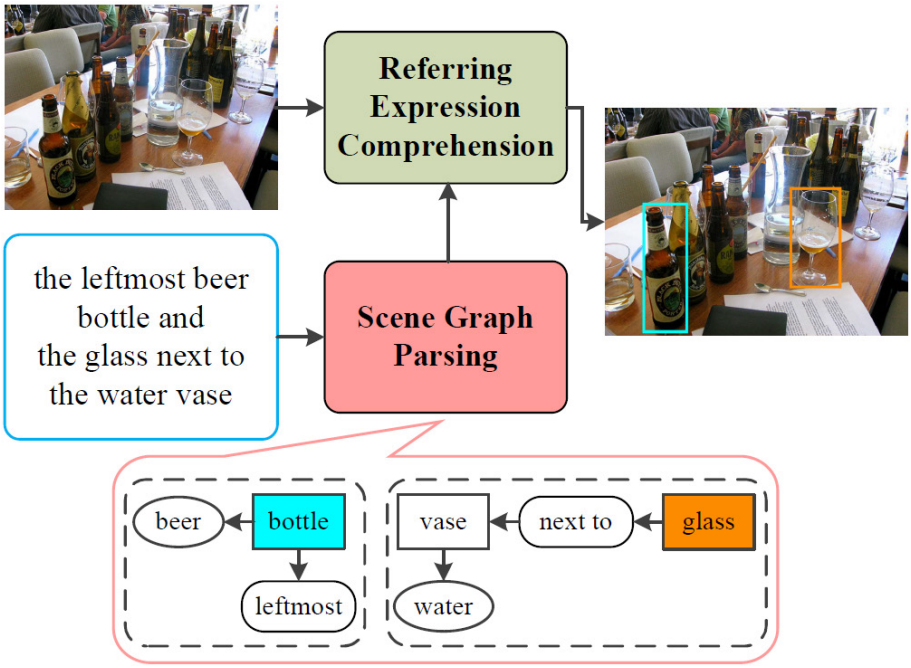

Figure 5.1: The architectural diagram of natural language grounding via referring expression comprehension and scene graph parsing. We first parse the natural language instructions into scene graph legends by the scene graph parsing. We then ground the generated scene graph legends via the referring expression comprehension network. The mark rectangle in bottom encompasses the scene graph parsing result for the input natural language query. The scene graph consists of: rounded rectangles with black dashed lines denote the parsed scene graph legends, color shaded rectangles represent referents, no color shaded rectangle is an object, ovals indicate objects attributes, rounded rectangles act for edges with relationships between other objects. The same color of the bounding boxes in the output image and the referents in parsed scene graph legends denote a grounding.

comprehension model.

5.4.2 Experiments

We validate the effectiveness of the presented natural language grounding architecture in two different manners. First, we select 133 indoor scenarios from the test

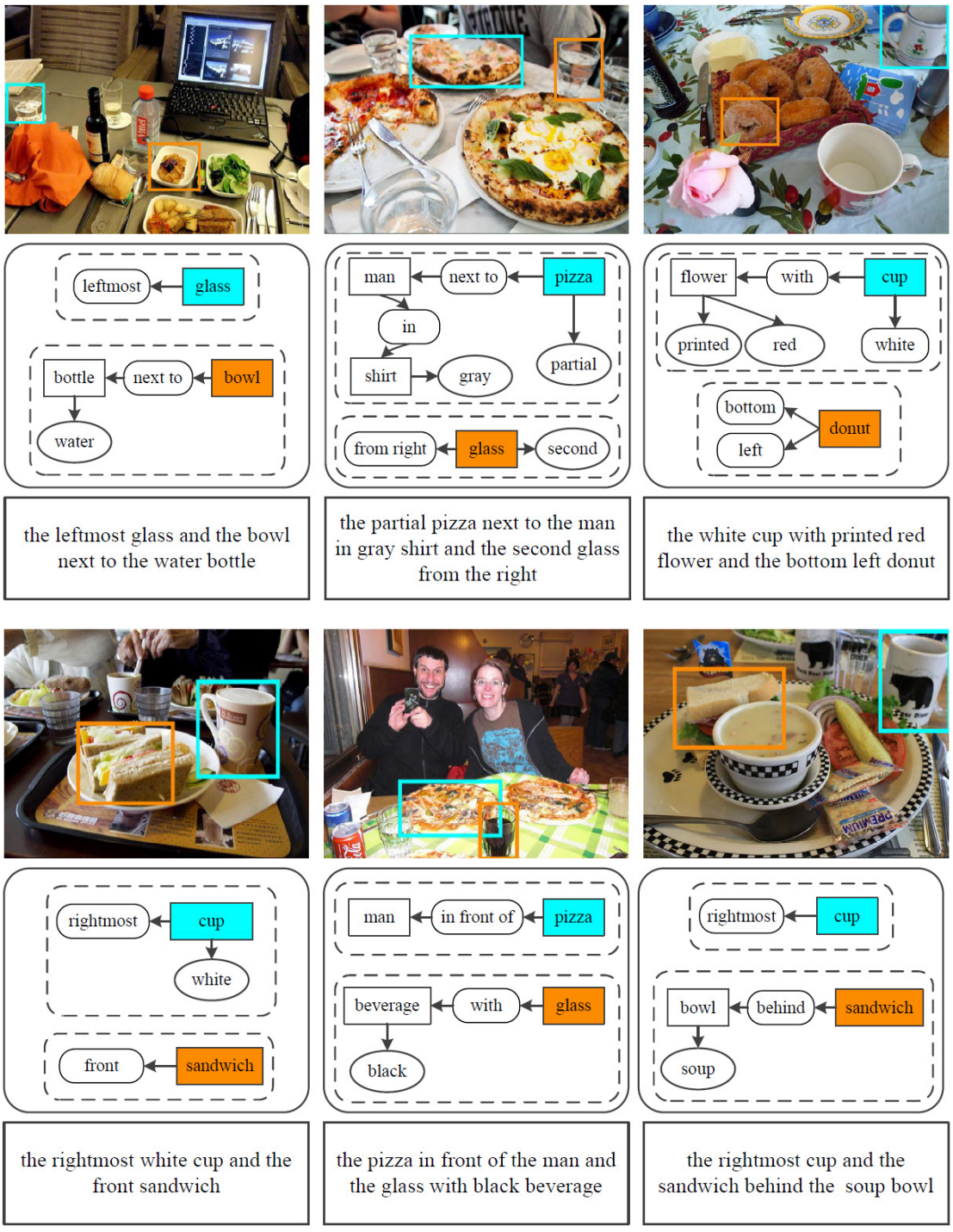

Figure 5.2: Example results of natural language grounding via referring expression comprehension and scene graph parsing on MSCOCO images. The input natural language commands are listed in the rectangles in the third row, the scene graph parsing results are shown in the rounded rectangles in the second row.

datasets of RefCOCO, RefCOCO+, and RefCOCOg, and collect 187 expressions that contain 2 referents for the selected images. The average length of the expressions for MSCOCO images is 10.75. Second, we collect 30 images via a Kinect V2 camera, and these images consist of the household objects that can be manipulated by robots. We collect 220 expressions, which contain 128 expressions with 2 referents and 92 expressions with 3 targets, for the self-collected images. The average number of words in these expressions is 14.31.

In order to collect diverse expressions for the collected images, we recruit 10 participants and show them different scenarios. For the MSCOCO images, we ask the participants to give expressions to depict two specific targets for each scenario, such as "the bottom row second donut from the left and the bottom rightmost mug". Figure 5.2 lists some grounding results of the MSCOCO images. We adopt the referring expression comprehension network trained on the three datasets to ground the expressions. The accuracies of the collected expressions grounding for MSCOCO images acquired by the three models are RefCOCO 86.63%, RefCOCO+ 79.41%, and RefCOCOg 80.48%.

For the self-collected scenarios, we ask the participants to give expressions with two or three referents for each image. For instance, "move the red apple outside the box into the box and take the second water bottle from the right". Figure 5.3 shows the example results for the self-collected scenarios. The grounding accuracies attained by the three models are RefCOCO 91.63%, RefCOCO+ 87.45%, and RefCOCOg 88.44%. From these experimental grounding results, it is clear that the trained referring expression comprehension model has superior robustness.

Through analyzing the failure grounded expressions, we found that the expressions with more "and" cannot be parsed correctly. For instance, the expression "take the apple between the bottle and the glass and the red cup" will be parsed into four nodes "apple", "bottle", "glass", and "red apple", while the relationship between "apple", "bottle", and "glass" is lost, which leads to the failure grounding.

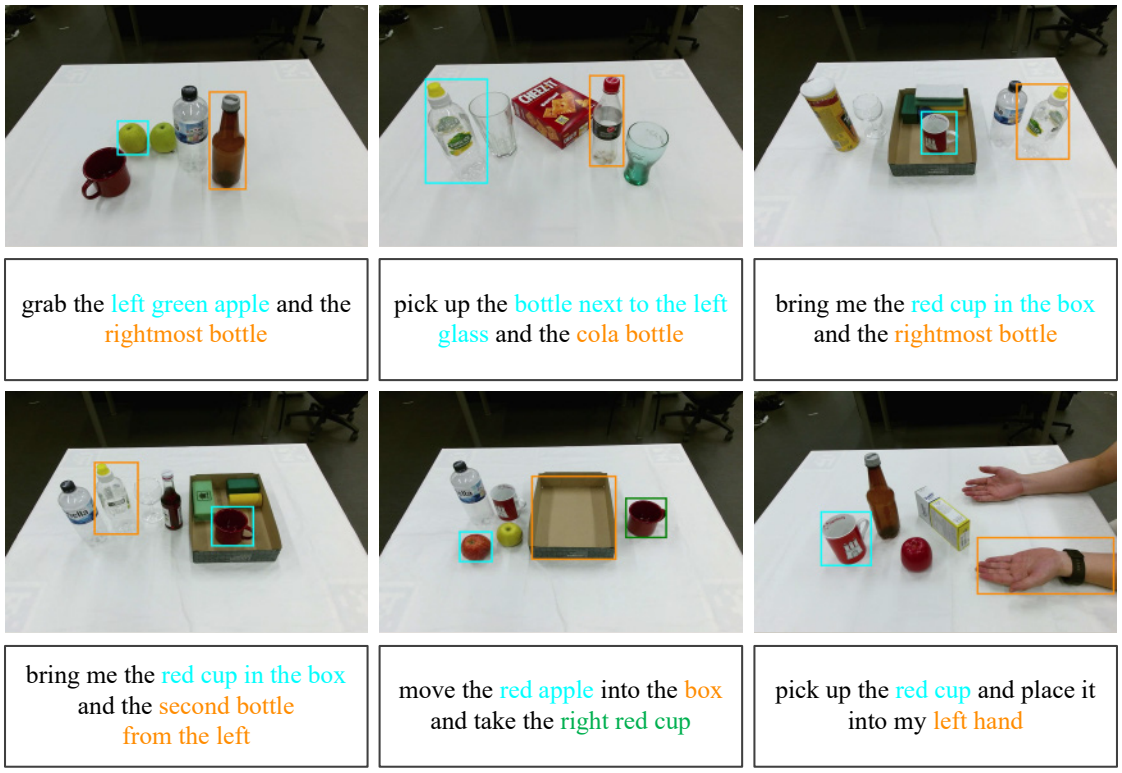

Figure 5.3: Example results of natural language grounding via referring expression comprehension and scene graph parsing on self-collected scenarios. The input natural language instructions are listed in the rectangles, and the parsed scene graph legends are covered with the corresponding color of target objects in the output images.

5.5 Interactive Natural Language Grounding via Referring Expression Generation and Scene Graph Parsing

We also integrate the referring expression generation model with scene graph parsing to ground complicated natural language instructions. Unlike the referring expression comprehension-based grounding architecture, the generation-based framework locates targets within images by identifying semantic relatedness between input queries and generated expressions of image regions. Thus, the generation-based

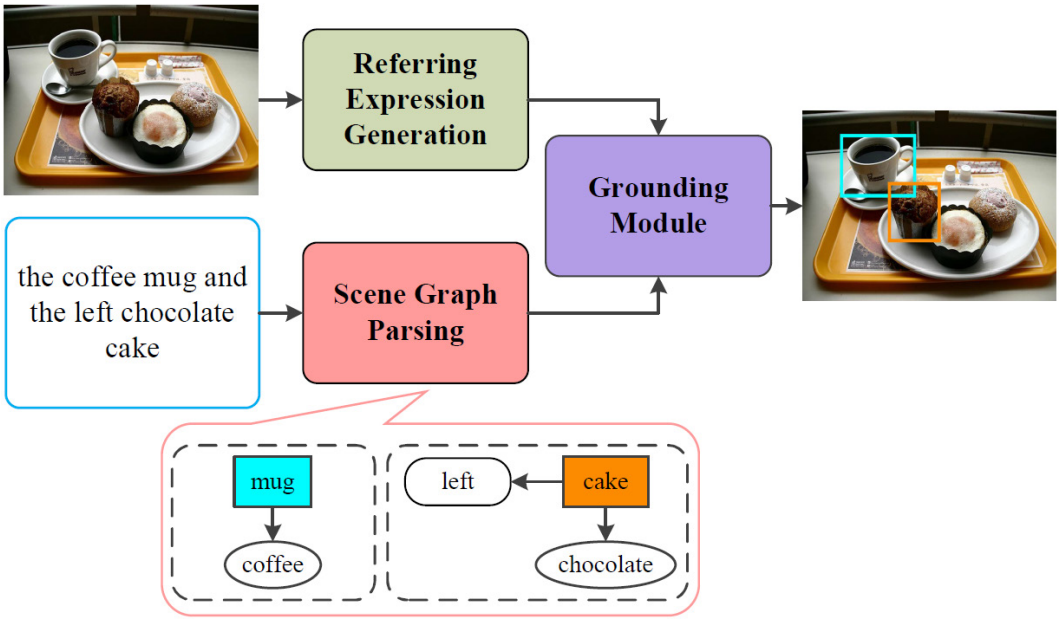

Figure 5.4: The architecture of natural language grounding via referring expression generation and scene graph parsing. The referring expression generation network generates referring expressions for image regions, as elaborated in chapter 3. The grounding module takes the generated expressions and the parsed scene graph legends as inputs to locate target objects.

framework needs a grounding module to acquire the semantic similarity between generated expressions and parsed natural language queries. We validate the effectiveness of the referring expression generation-based natural language grounding architecture on the indoor working scenarios and natural language queries used in the experiments for validating the referring expression comprehension-based grounding architecture.

5.5.1 Architecture Overview

The architectural diagram of natural language grounding via referring expression generation and scene graph parsing is shown in Figure 5.4. The referring expression generation network generates expressions for each detected region within images, the grounding module takes the generated expressions and the scene graph leg-

ends parsed from complicated natural language queries to achieve target objects grounding.

Under this framework, we reformulate the natural language grounding as three subtasks: (1) generate referring expression for each detected region within images, (2) parse the complex input queries into scene graph legends, (3) calculate the semantic similarity between the generated expressions and parsed scene graph legends to locate target objects.

5.5.2 Target Grounding

An essential step to realize natural language grounding via referring expression generation and scene graph parsing is to acquire the semantic relatedness of the generated expressions and the parsed natural language commands. Inspired by the Latent Semantic Analysis (LSA) which is used to measure the similarity of words and text documents meaning, we propose a sentence semantic metric measuring-based approach to build the mapping between the generated expressions and the natural language queries.

We first transform each word in a sentence into a 300-D vector by GloVe [98], and then adopt InferSent [22], which is a sentence embedding approach and provides semantic representation for sentences, to generate a representation for the entire sentence. We further calculate the semantic relatedness between the vectorized representations of generated expressions and input queries to select the most related targets.

Given an input queries and generated expressions for the regions within working scenario, x and y represent the representations generated by InferSent for the queries and expressions, i.e., $x \in \mathbb{R}^{1\times4096}$, $y \in \mathbb{R}^{1\times4096}$. Then, the semantic similarity of them is calculated by:

$$Sim(x, y) = \frac{x \cdot y}{\|x\|_2 \cdot \|y\|_2} \tag{5.1}$$

where $\|\cdot\|_2$ denotes L_2 normalization operation.

Through the semantic similarity calculation, the object with the maximum semantic similarity of input query and the generated referring expression is selected as the target.

5.5.3 Experiments

We evaluate the performance of the referring expression generation-based natural language grounding architecture on the self-collected working scenarios and expressions in referring expressions comprehension-based grounding experiments. The grounding accuracies acquired by the three models are RefCOCO 80.23%, RefCOCO+ 77.80%, and RefCOCOg 78.24%.

We also implement grounding experiments on self-collected working scenarios, Figure 5.5 shows some grounding example results. The grounding accuracies obtained by the models trained on the three datasets are RefCOCO 86.32%, RefCOCO+ 81.22%, and RefCOCOg 83.78%, respectively.

5.6 Intention-related Natural Language Grounding via Object Affordance Detection and Intention Semantic Extraction

In our daily communication, we use explicit queries to convey our objective, such as referring expressions. We also express our intention in relatively vague expressions, for example, "I want to drink some water". The explicit natural language queries can be grounded through the referring expression comprehension and referring expression generation models. Unlike the target-specified natural language commands grounding, no dataset is published for grounding intention-related natural language instructions. While grounding intention-related natural language queries is also a crucial component of natural language visual grounding. Inspired by the affordance and its applications in natural HRI, we introduce an object affordance detection-based architecture to address the intention-related natural language grounding.

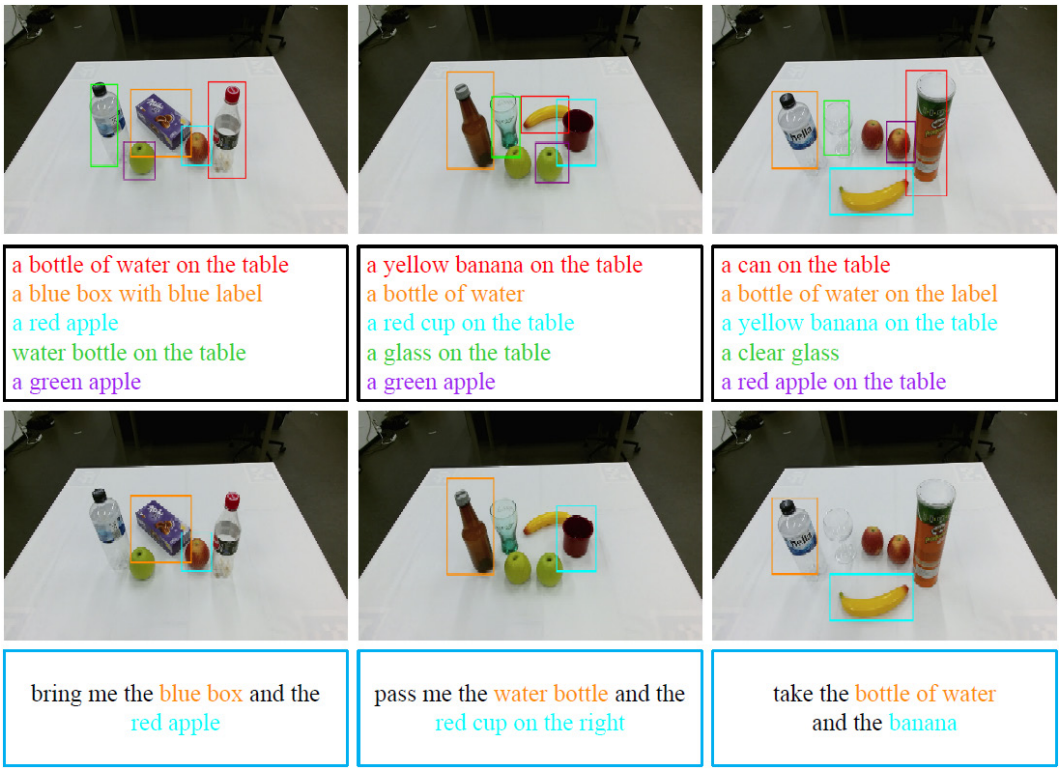

Figure 5.5: Example results of natural language grounding via referring expression generation and scene graph parsing. The rectangles with black outlines in the second row encompass the generated expressions for the detected objects in the self-collected working scenarios. The same color between the bounding boxes and the generated expressions represent correspondence. The input natural language instructions are listed in the rectangles with blue outline in the fourth row, the parsed scene graph legends are noted with the same color with the target objects as shown in the images in the third row. The same color of the bounding boxes in the images and the scene graph legends denotes a grounding.

5.6.1 Architecture Overview

Given an intention-related natural language command, such as "I am hungry, I want to eat something", and an image composed of multiple household objects, the objective of intention-related natural language grounding is to locate the most related object within the image. In this thesis, we decompose the intention-related

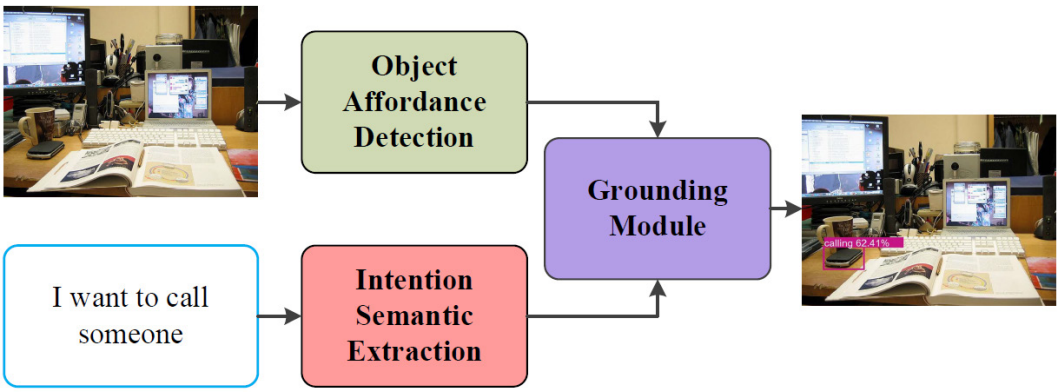

Figure 5.6: The architecture of intention-related natural language grounding via object affordance detection and intention semantic extraction. The introduced object affordance detection network detects objects affordances from visual images, as described in chapter 4. The intention semantic extraction module calculates the different weights of each word in given natural language queries, and extracts the intention semantic word. The grounding module locates the target object by combining the detected object affordances with the extracted intention semantic words.

natural language grounding into three subtasks: 1) an intention semantic extraction module extracts the intention semantic from the natural language instructions; 2) an object affordance detection network detects object affordances from RGB images; 3) a grounding module integrates outputs of the intention semantic extraction module and the object affordance detection network to locate the referred object. Figure 5.6 illustrates the details of the proposed intention-related natural language grounding architecture.

5.6.2 Intention Semantic Extraction

Each word plays a different role to represent the semantics of a natural language sentence, so we argue that each word should have different weights in a sentence for expressing the semantic. In order to acquire the different weights, we present a self-attentive network to calculate the weight of each word in natural language queries.

We acquire the weight in three steps. Given a natural language sentence S, we first tokenize S into words by NLTK [99] toolkit, i.e., $S = w_1, w_2, ..., w_n$, n denotes the word number of S. And the lexical category of each tokenized word w_i, $i \in (1, n)$ is generated by a POS-tagger (part of speech tagger) of NLTK. Second, we transfer w_i into a 300-D vector v_i by GloVe [98] as word representation, $v_i \in \mathbb{R}^{1\times 300}$. And then, these word representation vectors are concatenated as the representation of the sentence, i.e., $V = (v_1, v_2, ..., v_n)$, $V \in \mathbb{R}^{\mathrm{n}\times 300}$. Finally, the generated sentence representation V is fed into a self-attentive network to calculate the weight of each word. The self-attentive network adopts an attention mechanism over the hidden vector of a BiLSTM to generate a weight score α_i for w_i. The self-attentive network is defined as:

$$\begin{aligned} h_t &= \mathrm{BiLSTM}(V) \\ u_i &= tanh(W_w h_t + b_w) \\ \alpha_i &= \frac{exp(u_i)}{\sum_t exp(u_i)} \end{aligned} \tag{5.2}$$

where h_t represents the hidden vector of the BiLSTM, u_i is the transformation vector generated by an MLP with weight W_w and bias b_w.

In practice, we adopt the weight trained on the supervised data of the Stanford Natural Language Inference dataset [22] to be the initiate weight of the BiLSTM in the self-attentive network. For instance, the weight visualization of the sentence "I am thirsty, I want to drink some water" generated by the self-attentive network is shown in Figure 5.7.

The sentence is then re-ordered according to the acquired α_i, and the verb with the largest weight is selected to present the semantic of intention-related instructions, and the selected verb is feed into the grounding module to complete target object grounding. For instance, the word "drink" is selected as the representation of the intention-related spoken instruction "I am thirsty, I want to drink some water".

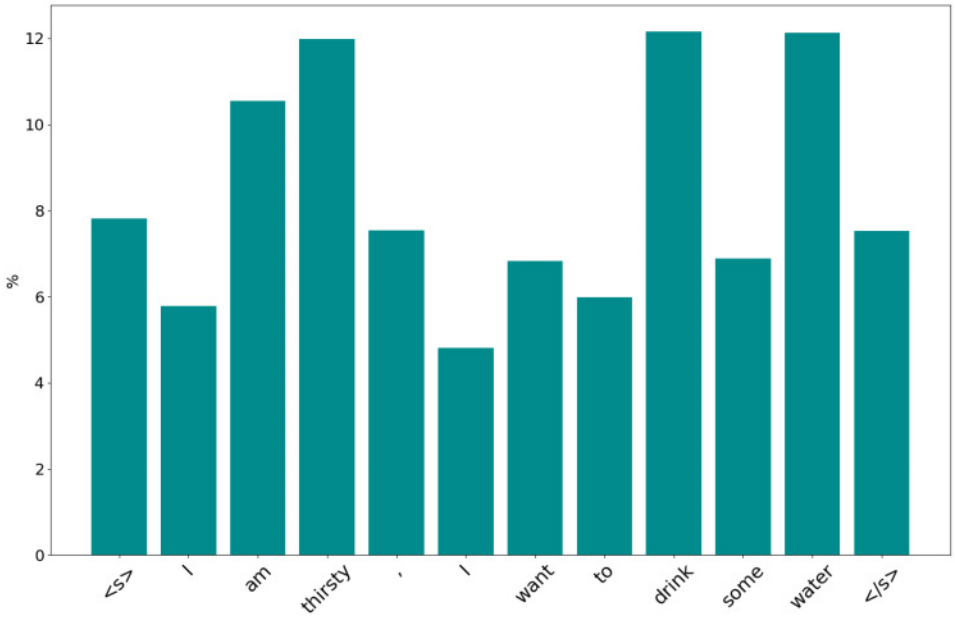

Figure 5.7: Visualisation of words weight of the sentence "I am thirsty, I want to drink some water". <s> and </s> represent the beginning of sentence token and the end of sentence token.

5.6.3 Target Grounding

In order to ground intention-related natural language instructions, we adopt the same mechanism which is introduced in the referring expression generation-based grounding architecture. Specifically, we first transfer the extracted semantic words and the detected affordances into 300-D vectors by GloVe, and then calculate the semantic similarity between the acquired intention semantic vectors and affordance vectors to complete targets grounding. Through the semantic similarity calculation, the extracted intention semantics are mapped into the corresponding human-centered object affordances.

5.6.4 Experiments

We select 100 images from the test set of the self-built dataset, and collect 150 natural language queries. Figure 5.8 lists some example results.

Figure 5.8: Example results of intention-related natural language query grounding via object affordance detection and intention semantic extraction. The first row lists example results of grounded target objects. The bar charts in the second row show the different weights of each word in given natural language instructions. The rectangles in the third row encompass the intention-related natural language queries, and the extracted intention semantics are covered with the related color with the detected affordances.

5.7 Spoken Instructions Visual Grounding and Robotic Applications

We train an online speech recognizer under Kaldi [102] with WSJ corpus, and integrate the trained recognizer with the three above introduced natural language visual grounding architectures to achieve spoken instruction grounding. We also conduct multiple spoken instruction grounding and target object manipulation experiments on a PR2 robot.

5.7.1 Online Speech Recognizer

In natural language-based HRI, the spoken utterance needs to be translated into text through speech recognition. In the open source toolkit Kaldi , the state-of-the-art techniques, such as Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transform (MLLT), Speaker Adaptive Training (SAT), Maximum Mutual Information (MMI), Minimum Phoneme Error (MPE), Deep Neural Networks (DNN), are applied in the AM training to acquire better recognition accuracy. Due to the properties of Kaldi, we train an online speech recognizer using the Kaldi to translate spoken language commands into text.

The WSJ corpus (LDC93S6B (WSJ0) and LDC94S13B (WSJ1)) provided by the Linguistic Data Consortium (LDC) is selected to be the corpus. The critical components of an online speech recognizer contain an acoustic model (AM), a language model (LM) and a lexicon. To acquire a better recognition accuracy, the Hidden Markov Model (HMM) + DNN AM and Tri-gram-based LM have been adopted in this thesis. The framework of the online speech recognizer is shown in Figure 5.9.

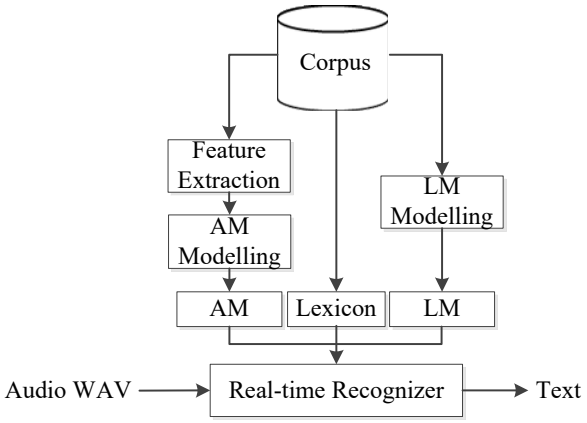

Figure 5.9: Framework of the online speech recognizer.

To evaluate the performance of the introduced online speech recognizer, we employ Word Error Rate (WER) to calculate the speech recognition accuracy. WER can be calculated as follows:

$$WER = \frac{D + I + S}{N} \tag{5.3}$$

where D denotes the number of deletions, I represents the number of insertions, S dedicates the number of substitutions, and N is the total number of words in the spoken sentences which have been translated into text.

In this thesis, we adopt the HMM + DNN based AM and Tri-gram based LM. According to [58], the WER of WSJ corpus is 6-7%. Because of the excellent WER of the adopted corpus, the trained online speech recognizer acquires a high recognition accuracy. With the high spoken command recognition accuracy, the performance of the intention semantics extraction module can be guaranteed.

5.7.2 Spoken Instruction Grounding and Target Object Segmentation

We also perform several target object groundings for spoken instructions by integrating the aforementioned architectures with the online speech recognizer. In the experiments, we use an ASUS wireless microphone to collect spoken instructions. These experiments aim to ground spoken instructions in working scenarios of robots and prompt natural HRI.

In order to achieve target object manipulations during robotic applications, we need to acquire the 3D localization of target objects. To this end, we draw support from instance segmentation. Formally, we first adopt Mask RCNN [45], which completes object detection and instance segmentation in one network, to segment the grounded target objects. We then integrate the segmented target objects with the depth data acquired from a Kinect V2 camera to realize 3D object localization. Figure 5.10 lists some grounding and segmentation results.

5.7.3 Robotic Applications

We conduct a number of target objects grasping experiments on a PR2 platform. The experimental setup for spoken instructions grounding and target objects manipulations is shown in Figure 5.11.

We complete objects manipulation planning in MoveIt and implement tar-

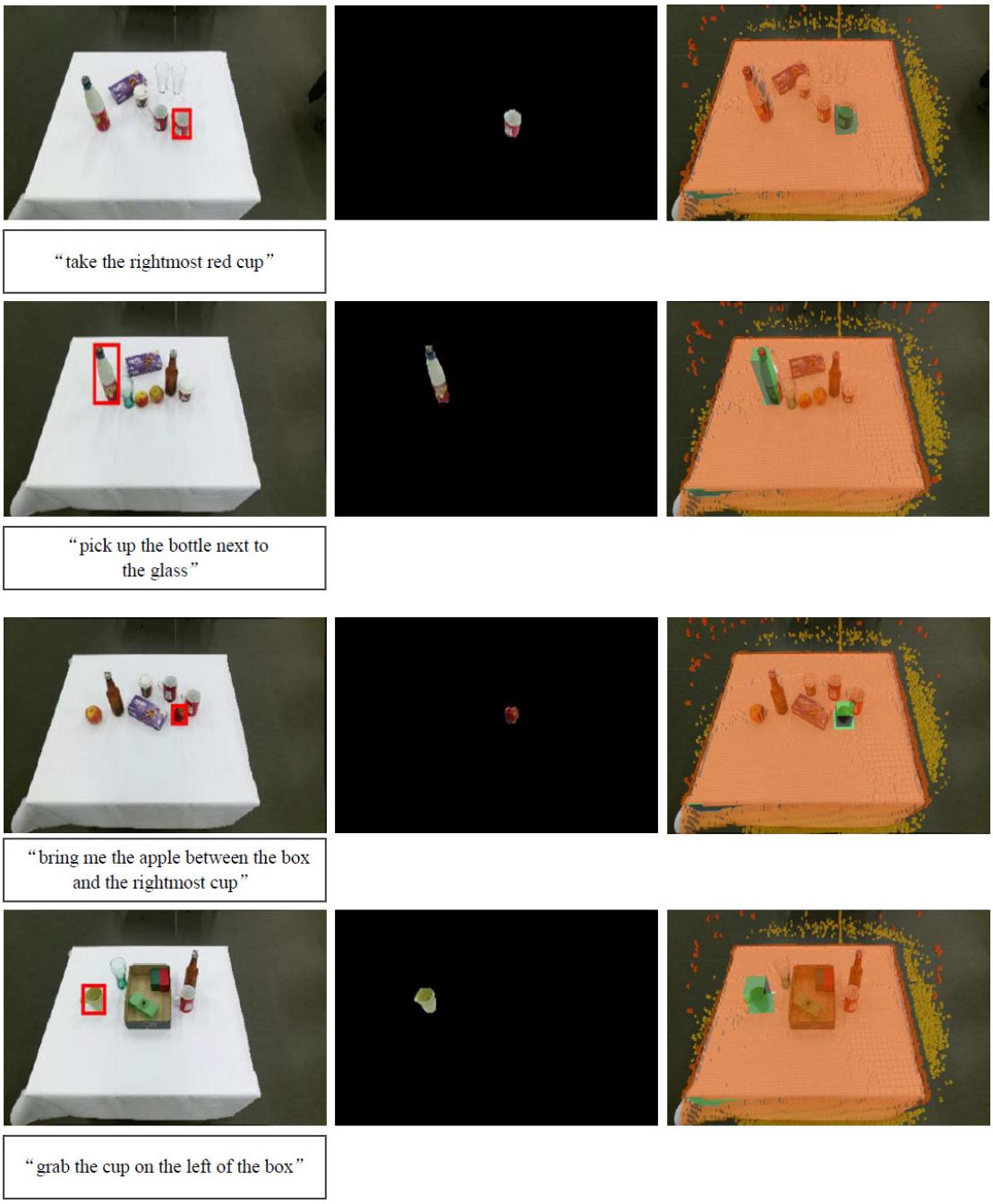

Figure 5.10: Example results of spoken instructions grounding and target object segmentation. The first image in each row shows the grounding result of a given spoken instruction, the second image is the segmentation acquired by Mask RCNN, and the third image is the segmentation in 3D (covered with green cubic) which is combined the point cloud data with the Mask RCNN segmentation. The spoken instructions are listed in the rectangles under the grounding results.

Figure 5.11: Experimental setup for spoken instructions grounding.

gets manipulation by a two-finger gripper of a PR2 platform. Figure 5.12 shows some target object manipulations by PR2, and the implementations mentioned above can be found on the following link:`https://www.youtube.com/watch?v=LbujSM6G5yY`.

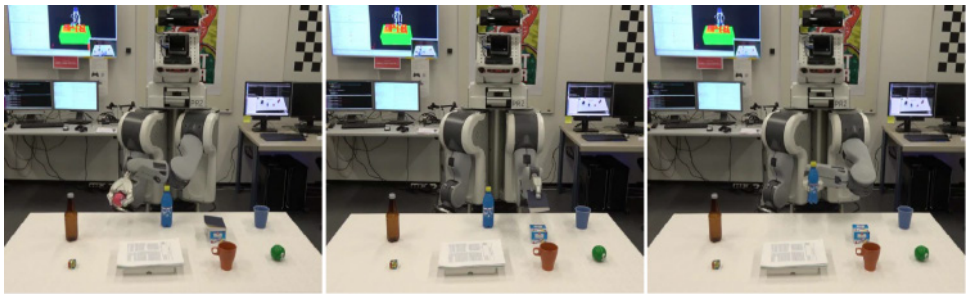

Figure 5.12: Target object grasping experiments conducted on a PR2 robot.

In this thesis, the pivotal point is to realize natural language visual grounding. Thus, in the grasping experiments, we attach more importance to locate the target objects rather than pay more attention to the grasping strategy or trajectory. A two-finger gripper implements grasping tasks. Because of the limitation of the gripper, the diversity of objects can be selected in the grasping experiments is heavily restricted.

5.8 Discussion

In this chapter, we proposed three natural language visual grounding architectures. In order to achieve unrestricted and compound natural language grounding, we integrated the trained referring expression comprehension and referring expression generation models with scene graph parsing. Compared with the existing methods for natural language visual grounding, the referring expression comprehension and referring expression generation-based approaches grounded and disambiguated the natural language instructions in a manner which is akin to an end-to-end pattern, and the introduced methods did not draw support from dialog systems and other auxiliary information.

Moreover, we presented a framework that combines the object affordance detection network with an intention-semantic extraction module and a grounding module to achieve intention-related natural language queries grounding. As far as we know, the introduced object affordance detection-based framework is the first attempt to ground intention-related natural language queries via human-centered affordances. We validated the three proposed natural language visual grounding architectures by implementing extensive experiments on household working scenarios with multiple natural language instructions.

Afterwards, we will improve the performance of the introduced architectures, such as exploring sophisticated approaches to address natural language grounding. Moreover, the introduced scene graph parsing module performs poorly when parsing some complex natural language queries, we will exploit a learning-based method to generate scene graphs. Additionally, we will exploit more effective methods to ground more complicated natural language queries and achieve more natural HRI.

Chapter 6

Conclusion

Natural language serves as the most straightforward medium in our daily communication, understanding natural language queries and grounding target objects are an essential skills for intelligent agents to communicate with humans. Natural language visual grounding can create a natural communication channel between humans, physical environments, and intelligent systems. Moreover, natural language visual grounding is widely used in multiple tasks. The primary objective of the thesis is to achieve natural language visual grounding without dialogue systems and other auxiliary information. As a foundation to address the research question, we focus on vision and natural language-based multimodal learning.

6.1 Thesis Summary

In order to ground natural language queries, we propose three different architectures that build upon three networks. First, we proposed a semantic-aware network for referring expression comprehension which imitates the role of a listener to ground the most related objects within images given referring expressions. On this basis, we combined the referring expression comprehension network with scene graph parsing to achieve sophisticated and unconstrained natural language queries grounding. We conducted experiments on three public referring expression datasets to evaluate the performance of the proposed referring expression comprehension

network, and we also implemented experiments on household working scenarios with diverse natural language queries to validate the effectiveness of the presented natural language grounding architecture.

Second, we presented an adversarial network for referring expression generation that mimics a speaker to generate referring expressions for each detected object within images. We validated the diversity and naturalness of expressions generated by the proposed referring expression generation network using multiple evaluation metrics. We also integrated the referring expression generation network with scene graph parsing and a grounding module to ground complex natural language queries. We evaluated the effectiveness of the referring expression generation-based natural language grounding framework using the working scenarios and the natural language queries collected for validating the referring expression comprehension-based grounding architecture.

Additionally, we introduced an object affordance detection network via attention-based multi-features fusion and a dataset for learning human-centered affordances. We combined the object affordance detection network with an intention semantic extraction module and a grounding module to achieve intention-related natural language instruction grounding. We validated the performance of the object affordance detection network on the self-built dataset, and we also evaluated the performance of the affordance detection-based intention-related natural language grounding framework on diverse working scenarios with multiple natural language queries.

6.2 Discussion

The research approaches introduced in this thesis relate to the interdisciplinary aspects of computer vision and natural language processing and aim to develop multimodal learning architectures for natural language visual grounding. In the following sections, we discuss the primary modeling aspects of the proposed architectures and the acquired results, and the limitations need to be addressed.

6.2.1 Referring Expression Comprehension

Referring expression comprehension requires a comprehensive understanding of natural referring expressions and visual images to locate referred objects in images. Unlike the other natural language and vision-based multimodal tasks, referring expression comprehension has been widely used in multiple applications.

Plenty of models have been proposed for referring expression comprehension [148, 84, 151, 16, 28, 147], etc. However, the existing approaches neglect two critical issues. First, the essence of deep features extracted from pretrained CNN models. According to [152], deep features are spatial, channel-wise, and multi-layer. The existing methods focus on the spatial characteristics and perform fine-grid spatial attention to access the most relevant object, while the importance of channel-wise characteristics is overlooked. For example, in the process of predicting objects, the channel-wise features are generated by the convolutional filters relevant to represent the visual semantics of objects. Therefore, the inherent semantic information of channel-wise features can be adopted to enhance the visual representations of detected regions. Second, the different contributions of each word to an expression. The existing approaches resort to holistic associations between the referring expressions and the region features, rather than take into account the different weights of words in expressions to locate target objects.

In contrast to the existing methods, we proposed a semantics-aware network for referring expression comprehension, where we excavated the visual semantic by taking full advantage of the characteristic of deep features and exploited the rich linguistic context of referring expressions. In chapter 2, we reformulated the referring expression comprehension into three sub-modules: 1) a language attention network learns to assign different weights for each word in expressions and learns to parse expressions into three phrases that embed the information of target candidate, relation and spatial location, respectively; 2) a visual semantics-aware network incorporates channel-wise and region-based spatial attention to generate semantic-aware visual representation for regions under the guidance of attended

words; 3) a target localization module coalesces the language attention network and the visual semantic-aware network to identify the target object.

Although the proposed network acquires relative lower accuracy than the state-of-the-art [147], which trained additional attributes for each detected regions and the attributes are employed as an enrichment for the region visual representation. While the proposed network utilized the inherent semantic of region deep features, and addressed the different contribution of each word in referring expressions to locate target objects.

6.2.2 Referring Expression Generation

Unlike the generic captions to describe given entire images, referring expressions depict each detected objects within images from diverse perspectives, such as color, size, location, and the interaction information between their neighbor objects. Thus, the referring expressions are closer to the pattern which humans take to describe objects.

The existing methods adopt generic CNN-LSTM paradigms to generate referring expressions [148, 149, 78]. Because of the generation paradigm and the evaluation metrics, the generated expressions are easy for humans to discriminate from human created ground truth. In addition, the existing approaches focus on locate the targets while sacrificing the semantic context. Inspired by these existing limitations, we introduce an adversarial training-based network to generate diverse referring expressions that are easy enough for humans to locate the target in images and reserve the semantic validity.

Motivated by the superior performance of GANs in image synthesis, and the synthesized images even cannot be distinguished by humans, we introduced GANs to generated referring expressions. In chapter 3, the primary goal is to improve the diversity and naturalness of generated referring expressions. We presented a generator to produce expressions and a discriminator to classify the generated expressions are real or fake. In order to generate expressions with more diversity and naturalness, we employed two critics in the discriminator to prompt the generator

producing diverse and natural expressions.

Unlike the Encoder-and-Decoder paradigm utilized in existing work, the introduced referring expression generation network produced more diversity and naturalness expressions. We also used multiple evaluation metrics to assess the generated referring expressions.

6.2.3 Object Affordance Detection

Affordance is a psychological term that refers to the fundamental properties of an object, while the properties determine how the object could possibly be used [94]. Moreover, affordance is widely used to achieve different objectives, such as improves the robustness of object recognition[125], augments the quality of HRI [12], and prompts the robot to understand natural language instructions [86]. Inspire by the affordance and its applications in HRI, we proposed an object affordance detection network and a dataset to learn household objects affordance.

Existing methods adopt different features to recognize affordances, such as geometric features, visual attributes, and deep feature extracted from pretrained CNN. Unlike the existing methods which take mono-feature as input to recognize object affordance, we took advantage of the multi-visual features, i.e., deep visual features extracted from a pretrained CNN model and deep texture features encoded by a deep texture encoding network, to recognize human-centered object affordances. Moreover, we proposed an attention network to fuse the multi-visual features to avoid bringing redundancy while preserving the complementary nature of the multiple features.

Besides, we proposed a dataset to train and validate the presented object affordance recognition network. We collected indoor scenarios from the MSCOCO and ImageNet datasets, and we also collected some household working scenarios via a Kinect V2 camera.

6.2.4 Interactive Natural Language Visual Grounding

Natural language plays a predominant role and provides the most straightforward medium in our daily communication. With the application of robots in human environments becoming omnipresent, the demand for natural language based HRI turns into urgent. Motivated by the applications of natural language visual grounding, we proposed three architectures to achieve natural language grounding without auxiliary information.

First, we combined the referring expression comprehension network with scene graph parsing to ground sophisticated natural language instructions. The scene graph parsing aims to parse complicated natural language commands into scene graph legends, which are composed of object with attributes and relations between objects. By this combination, the referring expression comprehension-based architecture can ground unconstrained and sophisticated natural language instructions.

Second, we integrated the referring expression generation with scene graph parsing and a grounding module to achieve natural language visual grounding. The grounding module takes referring expressions generated by the introduced referring expression generation network and the parsed scene graph legends to ground the target objects.

Third, we utilized the object affordance detection-based architecture to ground intention-related natural language queries. Specifically, we combined the detected object affordances with the extracted intention semantic words to ground intention-related natural language queries, such as "I am thirsty, I want to drink some water".

Moreover, we collected multiple indoor working scenarios and diverse natural language queries to validate the performance of the proposed natural language visual grounding architectures. We performed extensive grounding experiments and the experimental results demonstrated the effectiveness of the introduced grounding architectures. Additionally, we conducted spoken language visual grounding and target objects manipulation experiments on a PR2 robot.

6.3 Conclusion

To sum up, this thesis contributes to the field of natural language visual grounding by exploiting multimodal learning approaches, and aims to locate the referred target objects in working scenarios given natural language queries without auxiliary information. To this end, we proposed three architectures for natural language visual grounding, and these architectures are based on referring expression comprehension, referring expression generation, and object affordance detection, respectively. We explored the different roles of the three crucial components to realize natural language visual grounding.

Reported experimental results show the performance of the introduced networks for referring expression comprehension and referring expression generation, and object affordance detection. In addition, we implemented multiple experiments to validate the effectiveness of the three natural language visual grounding architectures.

In conclusion, our experiments demonstrated that referring expression comprehension and referring expression generation, and object affordance detection-based grounding architectures can be employed to achieve natural language visual grounding without auxiliary information.

6.4 Future Work

Compared to the existing work for referring expression comprehension, the proposed network acquires relative lower accuracy than the state-of-the-art on the three datasets. In future work, we will exploit the rich linguistic components of referring expressions and integrate with visual representation to acquire better accuracy, as well as improve the interpretability of the network. For the referring expression comprehension-based natural language grounding architecture, we will focus to improve the performance of the scene graph parsing to generate more accurate results for sophisticated natural language instructions.

The proposed referring expression generation network is the first attempt to generate diverse and natural expressions via adversarial training, and acquires promising results on three datasets. Afterwards, we will exploit a generative approach for dense captioning which aims to generate descriptions for each detected region within images, and the published datasets for dense captioning are more plentiful than the referring expression datasets. Moreover, the dense captioning model trained on big datasets is sufficient for grounding natural language queries.

The introduced object affordance detection network can detect ten human-centered affordances through fusing the deep visual features and the deep texture features. In the future, we will employ meta-learning to learn more affordances from a smaller amount of annotated images. Furthermore, we will integrate the referring expression generation approach with object affordance detection to generate affordance-aware referring expressions. In this way, the specific and intention-related natural language queries can be grounded by one architecture.

Appendix A

List of Abbreviations

HRI Human-Robot Interaction

CNN Convolutional Neural Network

LSTM Long Short-Term Memory

MLE Maximum Likelihood Estimation

IoU Intersection over Unit

GANs Generative Adversarial Networks

RL Reinforcement Learning

RNNs Recurrent Neural Networks

BiLSTM Bidirectional Long Short-Term Memory

RoIs Regions of Interest

MLP Multi-Layer Perceptron

MMI Maximum Mutual Information

EOS End-of-Sentence Token

DNN Deep Neural Networks

FM Factorization Machines

VQA Visual Question Answering

ReLU Rectified Linear Unit

CCG Combinatory Categorial Grammar

FCG Fluid Construction Grammar

ECG Embodied Construction Grammar

ROS Robot Operating System

SL Statistical Learning

DeepQA Deep Question Answering

DS Dempster-Shafer theory

LSA Latent Semantic Analysis

LDA Linear Discriminant Analysis

MLLT Maximum Likelihood Linear Transform

SAT Speaker Adaptive Training

MPE Minimum Phoneme Error

LDC Linguistic Data Consortium

AM Acoustic Model

LM Language Model

HMM Hidden Markov Model

WER Word Error Rate

Appendix B

Collected Working Scenarios and Natural Language Queries

In order to validate the referring expression comprehension and referring expression generation-based natural language visual grounding architectures, we collect 133 indoor scenarios from the test datasets of RefCOCO, RefCOCO+, and RefCOCOg, and 187 expressions for the MSCOCO images, as shown in Fig B.1. We also collect 30 images which are composed of the commonly used household objects by a Kinect V2 camera, and 228 expressions for the 30 working scenarios, as shown in Fig B.2.

1. the coffee mug and the left food plate
2. the water glass and the top left food plate
3. the top food plate and the coffee mug

1. the container with fruits and the glass with juice
2. the container with foods and the mug next to the juice glass
3. the clear glass and the croissant

1. the rightmost white cup and the sandwich behind the soup bowl
2. the soup bowl and the top left mug
3. the soup bowl and the spoon

Figure B.1: Working scenarios seleceted from MSCOCO and collected natural language instructions for validating the referring expression comprehension and referring expression generation-based natural language grounding architectures.

1. pick up the green apple and the red apple
2. grab the water bottle and the cola bottle
3. take the red apple and the right bottle
4. bring me the water bottle and the green apple
5. pick up the left bottle and the right bottle
6. move the right bottle to left of the water bottle and pick up the red apple
7. pick the water bottle and place it to right of the cola bottle and take the green apple
8. move the green apple to right of the right bottle and take the water bottle
9. move the right apple to left of the water bottle and grab the green apple
10. pick up the left bottle and place it to right of the cola bottle and bring me the red apple

1. bring me the rightmost red cup and the leftmost red cup
2. pick up the right glass and the right apple
3. take the right glass and the left apple
4. grab the leftmost red cup and the left apple
5. take the rightmost cup and the right apple
6. move the right glass to left of the leftmost red cup and pick up the right apple
7. pick up the left apple and place it to right of the rightmost red cup and take the left red cup
8. move the right cup to left of the left red cup and then grab the right glass

1. bring me the red box and the right bottle
2. pick up the left bottle and the left glass
3. take the right glass and the right bottle
4. grab the water bottle and the right glass
5. pick up the left glass and the right bottle
6. move the right glass to left of the left bottle and pick up the right bottle
7. move the left glass to right of the right glass and grab the left bottle
8. pick up the red box and place it to right of the right box and take the right bottle
9. move the right bottle to left the left bottle and then grab the red box
10. pick up the left glass and place it to right of right glass and then take the left bottle

Figure B.2: Collected working scenarios via a Kinect V2 camera and natural language queries.

Appendix C

Publications Originating from this Thesis

C.1 Journal Articles

- Jinpeng Mi, Song Tang, Zhen Deng, Michael Goerner, Jianwei Zhang. Object Affordance based Multimodal Fusion for Natural Human-Robot Interaction. Cognitive Systems Research, 54:128–137, 2019.

- Jinpeng Mi, Jianwei Zhang. Interactive Natural Language Grounding via Referring Expression Comprehension and Scene Graph Parsing. Frontiers in Neurorobotics (submitted).

- Jinpeng Mi, Hongzhuo Liang, Nikolaos Katzakis, Changshui Zhang, Jianwei Zhang. Intention-Related Natural Language Grounding via Object Affordance Detection and Intention Semantic Extraction. Frontiers in Neurorobotics (submitted).

C.2 Conferences

- Jinpeng Mi, Yu Sun, Yu Wang, Haiyang Jin, Liang Li, Jianwei Zhang. Gesture Recognition based Teleoperation Framework of Robotic Fish. IEEE In-

ternational Conference on Robotics and Biomimetics (ROBIO), 2016, 137–142.

- Jinpeng Mi, Yannick Jonetzko, Fuchun Sun, Jianwei Zhang. Speech-based Object Grounding and Grasping for natural Human-Robot Interaction. International Conference on Cognitive Systems and Information Processing (ICCSIP), 2018.

- Zhen Deng, Jinpeng Mi, Zhixian Chen, Lasse Einig, Jianwei Zhang. Learning human compliant behavior from demonstration for force-based robot manipulation. IEEE International Conference on Robotics and Biomimetics (ROBIO), 2016, 319–324.

- Song Tang, Lijuan Chen, Jinpeng Mi, Mao Ye, Qingdu Li, Jianwei Zhang. Adaptive pedestrian detection by modulating features in dynamical environment. IEEE International Conference on Robotics and Biomimetics (ROBIO), 2017, 62-67.

- Song Tang, Yunfeng Ji, Jianzhi Lyu, Jipeng Mi, Qingdu Li, Jianwei Zhang. Visual Domain Adaptation Exploiting Confidence-Samples. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2019.

Appendix D

Acknowledgements

I would like to acknowledge to those who have supported me to complete my Ph.D. studies and researches at the Technical Aspects of Multimodal Systems research group, University of Hamburg in the last four years.

First of all, I would like to sincerely thank my supervisor Professor Jianwei Zhang, who has offered me the chance to chase doctor title in Germany and provided guidance, supports, and suggestions.

Second, I gratefully thank the DAAD German Academic Exchange Service for the Cognitive Assistive Systems project (Kz:A/13/94748) which offers the scholarship to me. With the financial aid, I had the chance to go to Germany, learn German in Berlin, and chase my doctorate in the University of Hamburg. I also would like to thank Professor Stefan Wermter, who is the coordinator of the DAAD project, has offered me the chance to research in Germany.

Third, I also would like to thank my current colleagues in TAMS who provide helpful assistance for me, Michael Goerner, Yannick Jonetzko, and Hongzhuo Liang help me to conduct object grasping experiments on PR2 robot and UR5 robotic platform, discuss the specific topics debug the codes with Song Tang and other colleagues, and Lu and Norman help to review and revise my papers. Studying for a doctorate abroad was a dream for me, I tried multiple times and I finally acquired the chance to chase my doctor tile in the University of Hamburg. I sincerely appreciate the support from my family members, the chance provided by

my supervisor, and the ardent help from my colleagues. A Chinese proverb said, "A man leaves his impression behind wherever he stays, just as a goose utters its cry where it flies". All the lovely colleagues and the efforts provided by them will be a good memory in my future life.

On a more personal perspective, I sincerely appreciate my loving grandparents, parents and other family members who always give me selfless love, support and care. I live in Germany for five years, I know the miss and solicitude fill all time in their life. Here, I just want to say, thank you my Yeye, Nainai, Baba, Mama and the other relatives, your selfless love is the originate of my impetus keep forwarding. I heartily wish all of the family members good health, happy life, prosperous family and all the best.

When I try to end the journey of my student life and start a new expedition of my life, multiple emotions welled up in my mind. One of my most favorite sentence from Su Shi's Song Poems said, "Turning my head, I see the deary beaten track. Let me go back! Impervious to wind, rain or shine, I will have my will". And I wish in my future life, "Someday, with my sail piercing the clouds, I will mount the wind, break the waves, and traverse the vast, rolling sea".

Bibliography

[1] Hyemin Ahn, Sungjoon Choi, Nuri Kim, Geonho Cha, and Songhwai Oh. Interactive text2pickup networks for natural language-based human–robot collaboration. *IEEE Robotics and Automation Letters*, 3(4):3308–3315, 2018.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[6] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on learning and Representation (ICLR)*, 2015.

[7] Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Roberto Basili, and Daniele Nardi. Effective and robust natural language understanding for human-robot interaction. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, pages 57–62, 2014.

[8] Emanuele Bastianelli, Danilo Croce, Andrea Vanzo, Roberto Basili, and Daniele Nardi. A discriminative approach to grounded spoken language understanding in interactive robotics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2747–2753, 2016.

[9] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3479–3487, 2015.

[10] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 3, 2017.

[11] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on learning and Representation (ICLR)*, 2015.

[12] Hande Celikkanat, Güner Orhan, and Sinan Kalkan. A probabilistic concept web on a humanoid robot. *IEEE Transactions on Autonomous Mental Development*, 7(2):92–106, 2015.

[13] Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

[14] David L Chen and Raymond J Mooney. Learning to interpret natural language navigation instructions from observations. In *Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 859–865, 2011.

[15] Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. Amc: Attention guided multi-modal correlation learning for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2644–2652, 2017.

[16] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4042–4050, 2018.

[17] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 824–832, 2017.

[18] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5659–5667, 2017.

[19] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886, 2015.

[20] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3828–3836, 2015.

[21] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Using syntax to ground referring expressions in natural images. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 6759–6764, 2018.

[22] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, 2017.

[23] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8307–8316, 2019.

[24] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979, 2017.

[25] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3076–3086, 2017.

[26] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2951–2960, 2017.

[27] Atabak Dehban, Lorenzo Jamone, Adam R Kampff, and José Santos-Victor. Denoising auto-encoders for learning of objects and tools affordances in continuous space. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4866–4871, 2016.

[28] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7746–7755, 2018.

[29] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, 2014.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1*, pages 4171–4186, 2019.

[31] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, 2015.

[32] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5. IEEE, 2018.

[33] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015.

[34] Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1292–1302, 2013.

[35] Manfred Eppe, Sean Trott, and Jerome Feldman. Exploiting deep semantics and compositionality of natural language for human-robot-interaction. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 731–738. IEEE, 2016.

[36] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1473–1482, 2015.

[37] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6378–6388, 2019.

[38] Jerome Feldman, Ellen Dodge, and John Bryant. Embodied construction grammar. In *The Oxford handbook of linguistic analysis*. 2009.

[39] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

[40] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5275, 2017.

[41] James J Gibson. *The theory of affordances*. Hilldale, USA, 1977.

[42] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[43] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision (ECCV)*, pages 241–257. Springer, 2016.

[44] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781, 2018.

[45] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[47] Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7):1363–1377, 2016.

[48] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[49] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4203–4212, 2017.

[50] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1115–1124, 2017.

[51] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4555–4564, 2016.

[52] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[53] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3125, 2016.

[54] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

[55] Syed Ashar Javed, Shreyas Saxena, and Vineet Gandhi. Learning unsupervised visual grounding through semantic self-supervision. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–5, 2018.

[56] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574, 2016.

[57] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015.

[58] Kaldi-ASR. Kaldi. *http://kaldi-asr.org/doc/examples.html*, 2011.

[59] Yuki Katsumata, Akira Taniguchi, Yoshinobu Hagiwara, and Tadahiro Taniguchi. Semantic mapping based on spatial concepts for grounding words related to places in daily environments. *Frontiers in Robotics and AI*, 6:31, 2019.

[60] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014.

[61] David Inkyu Kim and Gaurav S Sukhatme. Semantic labeling of 3d point clouds with object affordance for robot manipulation. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5578–5584, 2014.

[62] Emiel Krahmer and Kees Van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012.

[63] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6867–6876, 2018.

[64] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[65] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.

[66] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[67] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[68] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 110–119, 2016.

[69] Xiangyang Li and Shuqiang Jiang. Bundled object context for referring expressions. *IEEE Transactions on Multimedia*, 20(10):2749–2760, 2018.

[70] Xiangyang Li, Shuqiang Jiang, and Jungong Han. Learning object context for dense captioning. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[71] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6116–6124, 2018.

[72] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Pro-*

ceedings of the IEEE International Conference on Computer Vision (ICCV), pages 1261–1270, 2017.

[73] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81, 2004.

[74] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2980–2988, 2018.

[75] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[76] Annika Lindh, Robert J Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D Kelleher. Generating diverse and meaningful captions. In *International Conference on Artificial Neural Networks (ICANN)*, pages 176–187. Springer, 2018.

[77] Chaofeng Liu, Zachary G Neale, and Guozhong Cao. Understanding electrochemical potentials of cathode materials in rechargeable batteries. *Materials Today*, 19(2):109–123, 2016.

[78] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4856–4864, 2017.

[79] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 375–383, 2017.

[80] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7219–7228, 2018.

[81] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7102–7111, 2017.

[82] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*, 2017.

[83] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. A multimodal classifier generative adversarial network for carry and place tasks from ambiguous language instructions. *IEEE Robotics and Automation Letters*, 3(4):3113–3120, 2018.

[84] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 11–20, 2016.

[85] Jonathan Masci, Michael M Bronstein, Alexander M Bronstein, and Jürgen Schmidhuber. Multimodal similarity-preserving hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):824–830, 2014.

[86] Jinpeng Mi, Song Tang, Zhen Deng, Michael Goerner, and Jianwei Zhang. Object affordance based multimodal fusion for natural human-robot interaction. *Cognitive Systems Research*, 54:128–137, 2019.

[87] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, 2016.

[88] Austin Myers, Ching Lik Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381, 2015.

[89] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision (ECCV)*, pages 792–807, 2016.

[90] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

[91] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016.

[92] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Detecting object affordances with convolutional neural networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2765–2770, 2016.

[93] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915, 2017.

[94] Don Norman. *The design of everyday things*. New York, NY, USA: Basic Books, 1988.

[95] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting on Association for Computational Linguistics (ACL), pages 311–318, 2002.

[96] Siddharth Patki, Andrea F Daniele, Matthew R Walter, and Thomas M Howard. Inferring compact representations for efficient natural language understanding of robot instructions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6926–6933, 2019.

[97] Rohan Paul, Jacob Arkin, Derya Aksaray, Nicholas Roy, and Thomas M Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms. *The International Journal of Robotics Research*, 37(10):1269–1299, 2018.

[98] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[99] Jacob Perkins. *Python text processing with NLTK 2.0 cookbook*. Packt Publishing Ltd, 2010.

[100] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5179–5188, 2017.

[101] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2539–2544, 2015.

[102] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr

Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[103] Joseph Redmon and Ali Farhadi. Yolo v3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[104] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, pages 1060–1069, 2016.

[105] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

[106] Steffen Rendle. Factorization machines. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*, pages 995–1000, 2010.

[107] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 817–834, 2016.

[108] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *European Conference on Computer Vision (ECCV)*, pages 186–201. Springer, 2016.

[109] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[110] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2017.

[111] Stefan Schiffer, Niklas Hoppe, and Gerhard Lakemeyer. Natural language interpretation for an interactive service robot in domestic domains. In *International Conference on Agents and Artificial Intelligence*, pages 39–53. Springer, 2012.

[112] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[113] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4135–4144, 2017.

[114] Tianze Shi, Liang Huang, and Lillian Lee. Fast (er) exact decoding and global training for transition-based dependency parsing via a minimal feature set. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12–23, 2017.

[115] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *Proceedings of Robotics: Science & Systems (RSS)*, 2018.

[116] Tianmin Shu, MS Ryoo, and Song-Chun Zhu. Learning social affordance for human-robot interaction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3454–3461, 2016.

[117] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. abs/1409.1556, 2014.

[118] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2014.

[119] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[120] Mark Steedman and Jason Baldridge. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and explicit models of grammar*, pages 181–224, 2011.

[121] Luc Steels. Fluid construction grammar. In *The Oxford handbook of construction grammar*. 2013.

[122] Luc Steels, Joachim De Beule, and Pieter Wellens. Fluid construction grammar on real robots. In *Language grounding in robots*, pages 195–213. Springer, 2012.

[123] Yu Sun, Shaogang Ren, and Yun Lin. Object-object interaction affordance learning. *Robotics and Autonomous Systems*, 62(4):487–496, 2014.

[124] Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. Towards human-friendly referring expression generation. *arXiv preprint arXiv:1811.12104*, 2018.

[125] Spyridon Thermos, Georgios Th Papadopoulos, Petros Daras, and Gerasimos Potamianos. Deep affordance-grounded sensorimotor object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6167–6175, 2017.

[126] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. Improving grounded natural language understanding through human-robot dialog. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6934–6941, 2019.

[127] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning multi-modal grounded linguistic semantics by

playing "i spy". In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3477–3483, 2016.

[128] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[129] Gyanendra K Verma and Uma Shanker Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *Neuro Image*, 102:162–172, 2014.

[130] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[131] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[132] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Learning compact hash codes for multimodal representations using orthogonal deep structure. *IEEE Transactions on Multimedia*, 17(9):1404–1416, 2015.

[133] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1960–1968, 2019.

[134] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipu-

lation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018.

[135] Yeyi Wang, Li Deng, and Alex Acero. Semantic frame-based spoken language understanding. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 41–91, 2011.

[136] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going beyond literal command-based instructions: Extending robotic natural language interaction capabilities. In *Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[137] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1583–1597, 2016.

[138] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5945–5954, 2017.

[139] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015.

[140] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2193–2202, 2017.

[141] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6580–6588, 2017.

[142] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4894–4902, 2017.

[143] Raymond A Yeh, Minh N Do, and Alexander G Schwing. Unsupervised textual grounding: Linking words to image concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6125–6134, 2018.

[144] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6241–6250, 2019.

[145] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, 2016.

[146] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[147] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1307–1315, 2018.

[148] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, pages 69–85, 2016.

[149] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7282–7290, 2017.

[150] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1974–1982, 2017.

[151] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1114–1120, 2018.

[152] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision (ECCV)*, pages 818–833, 2014.

[153] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[154] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5907–5915, 2017.

[155] Hang Zhang, Jia Xue, and Kristin Dana. Deep ten: Texture encoding network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2896–2905, 2017.

[156] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4158–4166, 2018.

[157] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 349–357. IEEE, 2019.

[158] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European Conference on Computer Vision (ECCV)*, pages 408–424. Springer, 2014.

[159] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4252–4261, 2018.

Declaration of Oath

Eidesstattliche Versicherung

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, February 4th, 2020	Jinpeng Mi
City, date	Signature
Ort, Datum	Unterschrift