Visual Grounding for Multi-modal Applications

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Kanishk Jain
2021701023
kanishk.j@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
November 2022

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled " **Visual Grounding for Multi-modal Applications**" by **Kanishk Jain**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Vineet Gandhi

This thesis is dedicated to my family, advisor and friends for their endless support and encouragement.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor Dr. Vineet Gandhi for believing in me and giving me the opportunity to work under his guidance. This thesis would not be possible without his continuous guidance and support. His dedication and passion for research have been a great source of motivation and have helped me develop a scientific rigour to move forward in my research journey. Apart from research, I admire his outlook on different aspects of life and strive to achieve them myself.

I thank Dr. K. Madhava Krishna for his trust and confidence in me. His meticulous teaching methods have been crucial to developing my understanding of robotics. I thoroughly enjoyed the thought-provoking discussions with him, ranging from academics to the very nature of consciousness. I would also like to thank Dr. Kavita Vemuri, as I got my first taste of research by working under her guidance as an honours student.

This acknowledgement wouldn't feel complete without the special mention of my friends Anurag Deshmukh, Aditya Sreekar, Ujjwal Tiwari and Kaveri Anuranjana, who were constantly by my side at every step of the way and for making my college life an exciting ride.

Finally and most importantly, my thesis and journey would not be possible without my loving parents' support, encouragement and guiding hand.

Abstract

The task of Visual Grounding is at the intersection of computer vision and natural language processing tasks. The Visual Grounding (VG) task requires spatially localizing an entity in a visual scene based on its linguistic description. The capability to ground language in the visual domain is of significant importance for many real-world applications, especially for human-machine interaction. One such application is language-guided navigation, where the navigation of autonomous vehicles is modulated using a linguistic command. The VG task is intimately linked with the task of vision-language navigation (VLN), as both the tasks require reasoning about the linguistic command and the visual scene simultaneously. Existing approaches to VG can be divided into two categories based on the type of localization performed: (1) bounding-box/proposal-based localization and (2) pixel-level localization. This work focuses on pixel-level localization, where the segmentation mask corresponding to the entity/region referred to by the linguistic expression is predicted. The research in this thesis focuses on a novel modeling strategy for visual and linguistic modalities for the VG task, followed by the first-ever visual grounding based approach to the VLN task.

We first present a novel architecture for the task of pixel-level localization, also known as Referring Image Segmentation (RIS). The architecture is based on the hypothesis that both intra-modal (word-word and pixel-pixel) and inter-modal (word-pixel) interactions are required to identify the referred entity successfully. Existing methods are limited because they either compute different forms of interactions sequentially (leading to error propagation) or ignore intra-modal interactions. We address this limitation by performing all three interactions synchronously in a single step. We validate our hypothesis empirically against existing methods and achieve State-Of-the-Art results on RIS benchmarks.

Finally, we propose the novel task of Referring Navigable Regions (RNR), i.e., grounding regions of interest for navigation based on the linguistic command. RNR is different from RIS, which focuses on grounding an object referred to by the natural language expression instead of grounding a navigable region. We additionally introduce a new dataset, Talk2Car-RegSeg, which extends the existing Talk2car dataset with segmentation masks for the regions described by the linguistic commands. We present extensive ablations and show superior performance over baselines on multiple evaluation metrics. A downstream path planner generating trajectories based on RNR outputs confirms the efficacy of the proposed framework.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Vision and language form the basis of human interaction with their environment. We are constantly capturing and processing vast amounts of visual and linguistic data to make sense of our surroundings and communicate ideas with other people. Humans are blessed with the innate capability of instinctively processing multi-modal data from visual and linguistic modalities and then combining information from these modalities to perform a multitude of daily tasks. For example, describing a visual scene using linguistic expression (Image Captioning), matching a visual scene with its linguistic description (Image-Text Retrieval), localizing an object in the visual world based on its linguistic description (Visual Grounding), and navigating in an environment based on a linguistic instruction (Vision-Language Navigation).

With the advancements in Deep Learning and the availability of large-scale datasets over the last decade, there has been an increasing interest in designing learning-based algorithms for various vision-language tasks. [29, 69] were one of the first deep learning-based attempts towards the task of Image Captioning, which requires generating a linguistic description for a given image. They trained their deep neural network (DNN) on a large amount of image-caption pairs by generating individual words of the caption in an auto-regressive manner for a given input image. The task of Image-Text retrieval is concerned with matching an image with the corresponding text and vice-versa. This is achieved by learning an alignment between the visual and linguistic modalities in a common semantic space [63, 19] in which the distance between matching image-text pairs is smaller than that between non-matching pairs. Similarly, every vision-language task differs in how the two modalities interact with each other and the role of deep learning-based methods is to identify / model the relation between the two modalities by leveraging the data statistics.

Since the introduction of AlexNet in 2012, neural-network-based architectures have become the go-to approaches for various computer vision tasks like image classification, object detection, and semantic segmentation. Image classification is a fundamental task that tests the ability to classify an image by assigning a label from a list of pre-determined categories. Similarly, the tasks of object detection and semantic segmentation involve classification based on pre-defined categories, but they additionally require spatial localisation of the objects of interest within the image. However, using a limited set

of categories hampers the scalability and practicality of the derivative approaches in the real world. Consider the simple example of an image classifier trained on the famous imagenet dataset consisting of 1000 class labels. Now, whenever we encounter a new class label not present in the 1000 category list, the classifier needs to be re-trained on the updated category list to incorporate the new labels for classification. Furthermore, a single label cannot be used to discriminate between instances of the same object. For example, an object detector for pedestrian detection will detect all the pedestrians in the scene, but it does not have the capabilities to distinguish between different pedestrians.

Moreover, learning with fixed category labels is in stark contrast with how humans learn. Instead, humans utilise natural language descriptions to interact with objects in our surroundings. In order to distinguish between instances belonging to the same class, natural language can provide a specific description of the relevant object based on its relationship with the surrounding environment. The task of visual grounding aims to design learning algorithms which imitate such human-like learning. Specifically, it requires learning correlations between entities in the visual scene and their linguistic counterparts.

In this thesis, we first tackle the Visual Grounding task by presenting a novel vision-language grounding network. Then, we propose a novel visual-grounding-based solution for the task of Vision-Language Navigation. Both the tasks require a joint understanding of the visual and linguistic modalities for successful completion. We describe each task in detail in the subsequent sections.

1.1 Contributions

More formally, we make the following contributions:

1. We propose a novel architecture for the task of Referring Image Segmentation (RIS) which models the word-word and region-region (intra-modal) interactions and word-region (inter-modal) interactions to ground the referred entity in the image.

2. We propose the novel task of Referring Navigable Regions (RNR) for Vision-Language Navigation which grounds the navigable regions on the road corresponding to the natural language command.

3. We present a new dataset, Talk2Car-RegSeg for the proposed RNR task. The dataset aids our visual-grounding based approach to vision language navigation.

1.2 Visual Grounding

Humans have the exceptional capability of associating the natural language description with the entities in the visual world. The goal of Visual Grounding (VG) is to design intelligent systems with similar capabilities. At its core, the task requires learning an alignment between the entities of visual

Figure 1.1 Referring Expression Comprehension (REC) and Referring Image Segmentation (RIS) are two tasks for visual grounding which differ based on the type of localization used. REC uses bounding-box-based localization while RIS uses pixel-based localization.

modality (objects, regions of an image) with the referred entities of linguistic modalities (word, sub-phrases of textual description), followed by spatial localization of the referred entity in the visual scene.

Earlier works [30, 23, 53] first generate candidate bounding-boxes corresponding to the objects in the image using a pre-trained object detector as the bounding-box proposal network. Then they rank each candidate bounding box based on its similarity with the linguistic expression in a common semantic space for visual and linguistic modalities. These approaches are trained using contrastive or margin ranking loss to ensure that the proposal features for the correct referred object are closer to the linguistic features of the natural language expression than that corresponding to the incorrect object. Moreover, [53] add an additional supervision by reconstructing the natural language expression based on the language conditioned ranking of the candidate bounding boxes. However, the performance of these approaches is highly reliant on the performance of the pre-trained proposal network, i.e., if the proposal network fails to predict a bounding-box corresponding to the correct referred object, then the whole approach will fail.

Parallel to these works, [22] proposed an approach to ground the natural language expression in an image through binary segmentation masks. They used an encoder-decoder architecture to first encode the input image and the linguistic expression, followed by the fusion between the two modalities. Finally, the fused multi-modal feature is passed as an input to the decoder, which predicts a segmentation mask as the output. Furthermore, their approach was end-end trainable as instead of using pre-trained

object detectors, they directly utilized the spatial feature map from the last layer of the Convolutional Neural Network (CNN) backbone.

Depending on the type of localization used, the visual grounding task can be divided into two categories. When a bounding box/proposal is used to localize the referred object, the task is formally known as Referring Expression Comprehension (REC). In comparison, a pixel-based localization is formally known as Referring Image Segmentation (RIS). The difference between the two tasks is illustrated in Figure 1.1. While there is a clear distinction between the task based on the type of localization used, the approaches for each task have certain advantages and disadvantages over the other. Approaches for REC usually have fewer trainable parameters and, as a result, low computation costs compared to approaches for RIS. However, they are more prone to errors because of their modular nature, where the first stage corresponds to generating proposals, and the next stage corresponds to ranking these proposals conditioned on the natural language query. In REC approaches, the proposal network serves as a bottleneck since only the generated proposals are ranked irrespective of whether any generated proposals contain the correct referred object.

On the other hand, approaches for RIS usually have high computation costs as they have a higher number of trainable parameters. In terms of localization, a segmentation mask is better suited to capture the exact shape and orientation of the referred object than a bounding box. Furthermore, in a highly occluded scene, the bounding box may capture other overlapping objects that do not match the linguistic expression. In Figure 1.1, we show examples where a bounding-box-based localization contains multiple objects which do not satisfy the natural language description.

In this work, we tackle the task of Referring Image Segmentation by proposing a novel architecture. The proposed architecture achieves state-of-the-art performance on several RIS benchmarks, and extensive quantitative and qualitative ablations validate the effectiveness of our approach.

1.3 Vision Language Navigation

The task of Vision Language Navigation (VLN) requires navigating in an environment based on natural language commands. For example, consider the example in Figure 1.2, the linguistic command is ”park next to the first white car,” the VLN task requires understanding the semantics in both the linguistic and visual modalities and navigating to the target location to execute the linguistic command successfully. For humans, scenarios similar to that depicted in Figure 1.2 are commonplace. We can understand the contextual relations between the linguistic and visual modalities and execute a navigational manoeuvre to the desired location. The VLN task aims to progress research towards designing autonomous agents capable of adapting to human interventions.

Existing works model the task as a sequence-to-sequence prediction problem [57, 82, 71] or as a reinforcement learning problem [20, 44, 1]. For outdoor environments, recently proposed [57, 82, 71], predict a sequence of actions from a fixed list of directional movements, namely, forward, right, left and stop. These approaches utilize the TOUCHDOWN [8] dataset, which is composed of images from

Figure 1.2 We solve the task of VLN as a visual grounding problem. Given a linguistic command, we identify the navigable regions on the road corresponding to the manoeuvre associated with the linguistic command.

google street view maps, and the directional actions are used to navigate between the images. Similarly, for indoor environments, [44, 1] utilize the Matterport3D dataset proposed in [6]. Matterport dataset consists of panoramic images from scenes inside buildings like houses, apartments, hotels, offices, and churches. The viewpoints for panoramic images are separated by an average distance of 2.25 meters, i.e., the floor is divided into a discrete set of navigable regions. However, discretizing the navigable regions and the action space severely limits the type of actionable navigational maneuvers. For example, consider the linguistic command, "park between the red and black cars," a discrete action space comprising forward, left, and right directions are insufficient to execute the navigational maneuver corresponding to the linguistic command successfully. Similarly, an environment consisting of discrete navigable regions limits the fine control of the car required to execute these navigational maneuvers successfully. Furthermore, the predictions of these sequence-to-sequence approaches are non-trivial to interpret as they lack human-understandable feedback.

We approach the task of VLN as a visual grounding problem by predicting a segmentation mask corresponding to navigable regions on the road for a given linguistic command. Our novel approach is called Referring Navigable Regions (RNR). The predicted navigable region is then used by an external motion planner to navigate to the desired location. This approach allows counteracting the issues associated with discrete environments and discrete action spaces, as any region on the road can be predicted as a navigable region candidate. Furthermore, the proposed approach permits grounding linguistic commands requiring fine control of the vehicle, such as "park between the red and white car", as the precise location corresponding to the command can be directly predicted. Additionally, the proposed approach is interpretable as the predicted segmentation masks also function as visual feedback for humans to understand the predictions.

1.4 Related Work

1.4.1 Semantic Segmentation

The task of semantic segmentation requires identifying the segments of different objects within an image. It is formulated as a pixel-wise classification problem, where each pixel is assigned a semantic label from a fixed set of object categories corresponding to the object it belongs. Introduction of Fully Convolution Networks (FCN) [38] led to a significant breakthrough in Semantic Segmentation. FCN replaces the fully connected layer in classification networks with convolutional layers and introduces skip connection for generating dense predictions for pixel-wise labels. However, the major limitation of FCN was that it did not take into account the global contextual information from the visual scene because of a low receptive field and the predicted segmentation mask was of low resolution. Parsenet [36] introduced global contextual information to FCNs by utilizing the average feature for a CNN layer to augment the features at each spatial location. Later works [45, 3] employed an encoder-decoder architecture, where the feature map representation from the penultimate CNN layer is passed through deconvolution / transposed convolution layers to upsample the spatial resolution of feature maps and output a segmentation mask at higher resolution. Similarly, U-Net [55] introduced skip-connections from the encoder to the decoder to utilize fewer training samples for the semantic segmentation task effectively. Feature Pyramid Network (FPN) [33] was introduced to effectively utilize the multi-scale hierarchical information captured by the deep CNNs; they take a single-scale image as input, and output feature maps at multiple scales, in a fully convolutional fashion. DeepLab and its variants [9, 10, 11] introduce atrous convolution to enlarge the receptive field of convolutional filters and aggregate multi-scale context using atrous spatial pyramid pooling. Atrous convolution introduces another parameter to convolutional layers, the dilation rate. A 3×3 kernel with a dilation rate of 2 will have the same size receptive field as a 5×5 kernel while using only 9 parameters, thus enlarging the receptive field with no increase in computational cost. PSPNet [81] performs region-based context aggregation through pyramid pooling to extract multi-scale context. DANet [21] utilizes channel and position attention to integrate local features with their global dependencies adaptively. Recent works like ResNeSt [80] and HRNet-OCR [67] use attention-based approaches to combine information across feature map groups and to combine multi-scale predictions, respectively. In this work, we tackle a more generalized and natural variant of semantic segmentation where natural language referring expressions replace the predefined set of object categories. We utilize the DeepLabv3+ architecture (Figure 1.3) from [11], pre-trained on semantic segmentation task on PASCAL-VOC dataset [17] as backbone for extracting visual features for the tasks of referring image segmentation (RIS) and referring navigable regions (RNR).

1.4.2 Attention Mechanism

Attention Mechanism is a powerful technique in deep learning literature popularized by its widespread use in various natural language processing tasks. The first use of the attention mechanism was proposed

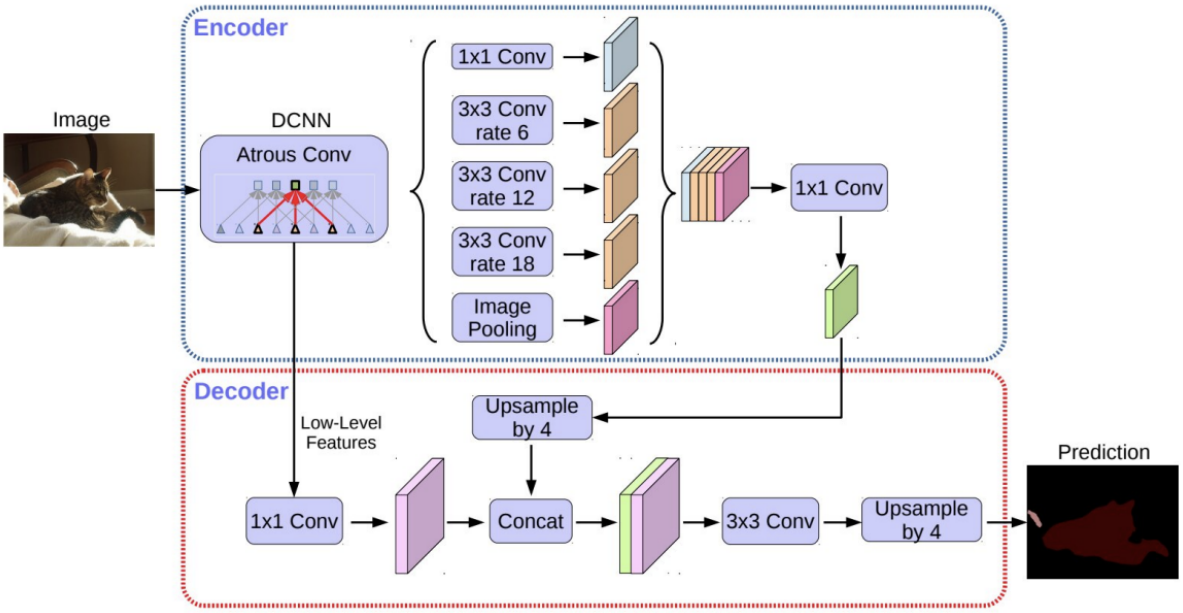


Figure 1.3 DeepLabv3+ [11] utilized an encoder-decoder architecture. The encoder module capture multi-scale contextual information by applying atrous convolution at multiple scales.

by [4] for the task of neural machine translation (NMT). They formulate the task as a sequence-to-sequence prediction task and utilize an encoder-decoder architecture. The encoder is an RNN that takes the linguistic sentence in the source language as the input, and the decoder is also an RNN that outputs the linguistic sentence in the target language. The decoder uses an attention mechanism over the words in the encoded input sentence to generate words in the target language in an auto-regressive manner. Subsequently, works like [72, 28, 77, 41, 2, 75, 83] utilized attention-based approaches for various multi-modal tasks in vision-language modalities. Specifically, [72, 28, 77] propose an attention-based approach for the task of image captioning, which utilizes attention over the spatial regions of the image while generating the individual words for the caption. Similarly, [41, 2, 75, 83] utilize attention for the task of Visual Question Answering (VQA); they employ joint attention over the spatial regions of the image and the words of the linguistic question to predict the answer. Concurrently, [53] tackle the task of grounding linguistic expression in the image by employing attention between the linguistic expression and the bounding-box proposals for the image.

Introduction of the transformer architecture [68] led to a major breakthrough in the deep learning literature. The transformer architecture utilizes the self-attention mechanism, which relates tokens of a single sequence with other tokens in the same sequence to compute a feature representation of the same sequence. Additionally, they introduce a multi-head attention mechanism, which splits the input sequence into fixed-size segments and then computes the self-attention over each segment in parallel. The original paper presented results on the machine translation task, but since its introduction, the transformer architecture has become ubiquitous in the deep learning literature. [16] proposed vision trans-

former, a pure transformer architecture without reliance on CNN for the task of image classification. They divide the image into patches and pass the patch embeddings as input to the vision transformer. ViLBERT [39] is a task-agnostic network for learning joint image and text representations. It utilizes co-attentional transformer layers to model the interactions between visual and linguistic modalities for multiple tasks like visual question answering, visual commonsense reasoning, referring expressions and caption-based image retrieval. Videobert [66] employs BERT-like [15] architecture to model bidirectional joint distribution over video and linguistic modalities for the task of video captioning. In this work, we utilize the attention mechanism to model the intra-modal and inter-modal interactions between the visual and linguistic modalities for the tasks of referring image segmentation (RIS) and referring navigable regions (RNR).

1.5 Thesis Organization

The rest of the thesis is organized as follows:

- In Chapter 2, we tackle the task of Referring Image Segmentation (RIS). We describe our novel strategy of effectively capturing intra and inter modal interactions between the visual and linguistic modalities. We effectively utilize the feature hierarchy associated with the visual features to identify the referred object and predict a refined segmentation mask corresponding to it.
- In Chapter 3, we describe our visual-grounding based approach to vision language navigation. We motivate the benefits of the proposed RNR approach and highlight its practicality in improving human-machine interaction by bringing interpretability to the VLN task.
- Finally in Chapter 4, we present our concluding thoughts.

Chapter 2

Comprehensive Multi-Modal Interactions for Referring Image Segmentation

2.1 Introduction

Traditional computer vision tasks like detection and segmentation have dealt with a pre-defined set of categories, limiting their scalability and practicality. Substituting the pre-defined categories with natural language expressions (NLE) is a logical extension to counteract the above problems. Indeed, this is how humans interact with objects in their environment; for example, the phrase "the kid running after the butterfly" requires localizing only the child running after the butterfly and not the other kids. Formally, the task of localizing objects based on NLE is known as Visual Grounding. Existing works either approach the grounding problem by predicting a bounding box around the referred object or a segmentation mask corresponding to the referred object. We focus on the latter approach, as a segmentation mask can effectively pinpoint the exact location and capture the actual shape of the referred object. The task is formally known as Referring Image Segmentation (RIS).

RIS requires understanding both visual and linguistic modalities at an individual level, specifically word-word and region-region interactions. Additionally, a mutual understanding of both modalities is required to identify the referred object from the linguistic expression and localize it in the image. For instance, to ground a sentence "whatever is on the truck", it is necessary to understand the relationship between words as grounding just the individual words will not work. Similarly, region-to-region interactions in visual modality help group semantically similar regions, e.g., all regions belonging to the truck. Finally, to identify the referent regions, we need to transfer the distinctive information about the referent from the linguistic modality to the visual modality; this is taken care of by the cross-modal word-region interactions. The current SOTA methods [74, 18, 25, 27, 24] take a modular approach, where these interactions happen in parts, sequentially.

Different methods differ in how they model these interactions. [25] first perform a region-word alignment (cross-modal interaction). The second stage takes these alignments as input to select relevant image regions corresponding to the referent. [74] and [27] use the dependency tree structure of the

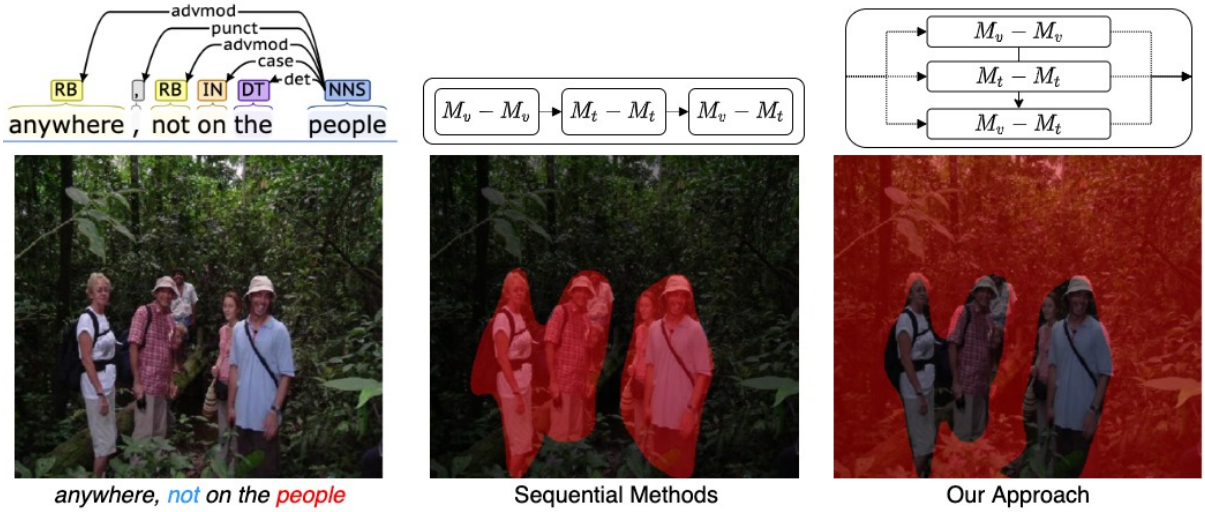


Figure 2.1 Unlike existing methods which model interactions in a sequential manner, we synchronously model the Intra-Modal and Inter-Modal interactions across visual and linguistic modalities. Here, M_v and M_t represent Visual and Linguistic Modalities, and $\{$-$\}$ represents interactions between them.

referring expression for the reasoning stage instead. [24] select a suitable combination of words for each region, followed by selecting the relevant regions corresponding to referent based on the affinities with other regions. The performance of the initial stages bounds these approaches. Furthermore, they ignore the crucial intra-modal interactions for RIS.

2.2 Contributions

In this work, we perform all three forms of interactions simultaneously. We propose a Synchronous Multi-Modal Fusion Module (SFM) which captures the inter-modal and intra-modal interactions between visual and linguistic modalities in a single step. Intra-modal interactions handle the cases for identifying the relevant set of words and semantically similar image regions. Inter-modal interactions transfer contextual information across modalities. Additionally, we propose a novel Hierarchical Cross-Modal Aggregation Module (HCAM) to exchange contextual information relevant to referent across visual hierarchies and refine the referred object's segmentation mask.

We motivate the benefits of simultaneous interactions over sequential in Figure 2.1 by presenting a failure case of the latter. For the given referring expression "anywhere, not on the people", sequential approaches fail to identify the correct word to be grounded, and the error gets propagated till the end. CMPC [25] which predicts the referent word from the expression in the first stage, identifies "people" as the referent (middle image in Figure 2.1) and completely misses "anywhere" which is the correct entity to ground. Similarly, [74], and [27], which utilize dependency tree structure to govern their reasoning process, identify the referred entity "anywhere" as an adverb from the dependency tree. However, considering the expression in context with the image, the word "anywhere" should be perceived as a

"pronoun". The proposed SFM module successfully addresses the aforementioned limitations. Overall, our work makes the following contributions:-

1. We propose SFM to reason over regions, words, and region-word features in a synchronous manner, allowing each modality to focus on relevant semantic information to identify the referred object.

2. We propose a novel HCAM module, which routes hierarchical visual information through linguistic features to produce a refined segmentation mask.

3. We present thorough quantitative and qualitative experiments to demonstrate the efficacy of our approach and show notable performance gains on four RIS benchmarks.

2.3 Related Work

Referring Expression Comprehension: Localizing a bounding box/proposals based on an NLE is a task commonly referred to as Referring Expression Comprehension (REC). The majority of methods for REC learn a joint embedding space for visual and linguistic modalities and differ in how joint space is computed and how it is used. Earlier methods, [23, 54, 49] used joint embedding space as a metric space to rank proposal features with linguistic features. Later methods like [73, 13, 37] utilized attention over the proposals to select the appropriate one. More Recent Methods like [40, 12] utilize transformer-based architecture to project multi-modal features to common semantic space. Specifically, they utilize a self-attention mechanism to align *proposal-level features* with linguistic features. In our work, we utilize *pixel-level image features* which are crucial for the task of RIS. Additionally, compared to [40], we *explicitly* capture inter-modal and intra-modal interactions between visual and linguistic modalities.

Referring Image Segmentation: Bounding Box-based methods in REC are limited in their capabilities to capture the inherent shape of the referred object, which led to the proposal of the RIS task. It was first introduced in [22], where they generate the referent's segmentation mask by directly concatenating visual features from CNN with tiled language features from LSTM. [32] generates refined segmentation masks by incorporating multi-scale semantic information from the image. Since each word in expression makes a different contribution in identifying the desired object, [59] model visual context for each word separately using query attention. [76] uses a self-attention mechanism to capture long-range correlations between visual and textual modalities. Recent works [24, 25, 27] utilize cross-modal attention to model multi-modal context, [27, 74] use dependency tree structure and [25] use coarse labelling for each word in the expression for selective context modelling. Most of the existing works capture Inter and Intra modal interactions separately to model the context for referent. In this work, we *concurrently* model the comprehensive interactions across visual and linguistic modalities.

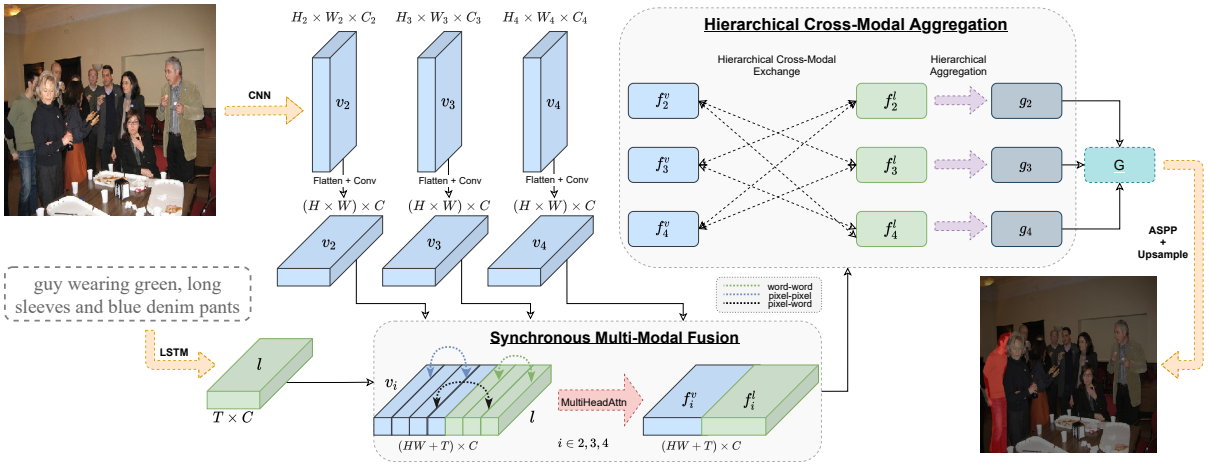


Figure 2.2 The proposed network architecture. Synchronous Multi-Modal Fusion captures pixel-pixel, word-word and pixel-word interaction. Hierarchical Cross-Modal Aggregation exchanges information across modalities and hierarchies to selectively aggregate context relevant to the referent.

2.4 Approach

Given an image and a natural language referring expression, the goal is to predict a pixel-level segmentation mask corresponding to the referred entity described by the expression. The overall architecture of the network is illustrated in Figure 3.2. Visual features for the image are extracted using a CNN backbone, and linguistic features for the referring expression are extracted using a LSTM. A Synchronous Multi-Modal Fusion Module (SFM) simultaneously aligns visual regions with textual words and jointly reasons about both modalities to identify the multi-modal context relevant to the referent. SFM is applied to hierarchical visual features extracted from CNN backbone since hierarchical features are better suited for segmentation tasks [76, 7, 24]. A novel Hierarchical Cross-Modal Aggregation module (HCAM) is applied to effectively fuse SFM's multi-level output and produce a refined segmentation mask for the referent. We describe the feature extraction process in the next section, and both SFM and HCAM modules are described in the subsequent sections.

2.4.1 Feature Extraction

Our network takes an image and a natural language expression as input. We extract hierarchical visual features for an image from a CNN backbone. Through pooling and convolution operations, all hierarchical visual features are transformed to the same spatial resolution and channel dimension. Final visual features for each level are of shape $\mathbb{R}^{C_v \times H \times W}$, with H, W and C_v being the height, width, and channel dimension of the visual features. Final visual features are denoted as $\{V_2, V_3, V_4\}$, corresponding to layers 2, 3 and 4 of the CNN backbone. For ease of readability, we denote the visual

features as V. GloVe embeddings for each word in the referring expression are then passed as input to LSTM. The hidden feature of LSTM at i^{th} time step $l_i \in \mathbb{R}^{C_l}$, is used to denote the word feature for the i^{th} word in the expression. The final linguistic feature of the expression is denoted as $L = \{l_1, l_2, ..., l_T\}$, where T is the number of words in the referring expression.

2.4.2 Synchronous Multi-Modal Fusion

In this section, we describe the Synchronous Multi-Modal Fusion Module (SFM). To successfully segment the referent, we need to identify the semantic information relevant to it in both the visual and linguistic modalities. We capture comprehensive intra-modal and inter-modal interactions explicitly in a synchronous manner, allowing us to jointly reason about visual and linguistic modalities while considering the contextual information from both.

Hierarchical visual features $V \in \mathbb{R}^{C_v \times H \times W}$ and linguistic word-level features $L \in \mathbb{R}^{C_l \times T}$ are passed as input to SFM, with $C_v = C_l = C$. We flatten the spatial dimensions of visual features and perform a lengthwise concatenation with linguistic feature, followed by layer normalization to get multi-modal feature X of shape $\mathbb{R}^{C \times (HW+T)}$. We then add separate positional embedding P_v and P_l to visual $X_v \in \mathbb{R}^{C \times HW}$ and linguistic $X_l \in \mathbb{R}^{C \times T}$ part of X to distinguish between visual and linguistic part. Finally, we apply multi-head attention over X to capture the inter-modal and intra-modal interactions between visual and linguistic modalities. Specifically, pixel-pixel, word-word and word-pixel interactions are captured. Pixel-pixel and word-word interactions help in independently identifying semantically similar pixels and words in their respective modalities, pixel-word interaction helps in identifying corresponding pixels and words with similar contextual semantics across modalities.

$$\begin{aligned} X &= \text{LayerNorm}(V \odot L) \\ X &= X + (P_v \odot P_l) \\ F &= \text{MultiHead}(X) \end{aligned} \tag{2.1}$$

Here, $\odot$ is length-wise concatenation, F is the final output of SFM module having same shape as X. We process all hierarchical visual features $\{V_2, V_3, V_4\}$ individually through SFM, resulting in hierarchical cross-modal output $\{F_2, F_3, F_4\}$.

2.4.3 Hierarchical Cross-Modal Aggregation

Hierarchical visual features of CNN capture different aspects of images. As a result, depending on the hierarchy, visual features can focus on different aspects of the linguistic expression. In order to predict a refined segmentation mask, different hierarchies should be in agreement regarding the image regions to focus on. Therefore, all visual hierarchical features should also focus on image regions corresponding to linguistic context from other hierarchies. This will ensure that all hierarchical features are focusing on common regions. We propose a novel Hierarchical Cross-Modal Aggregation (HCAM)

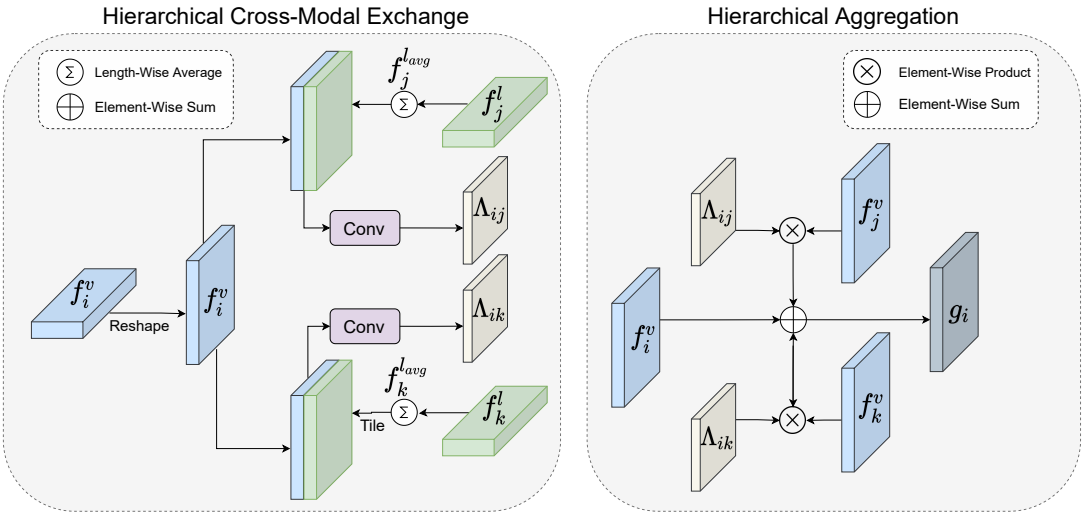


Figure 2.3 Our Novel Hierarchical Cross-Modal Aggregation Module consisting of Hierarchical Cross-Modal Exchange and Hierarchical Aggregation steps.

module for this purpose. HCAM includes two key steps: (1) **Hierarchical Cross-Modal Exchange**, and (2) **Hierarchical Aggregation**. Both steps are illustrated in Figure 2.3.

Hierarchical Cross-Modal Exchange: During the HCME step, we calculate the affinity weights Λ_{ij} between the j^{th} layer's linguistic context f_j^l and the spatial regions for i^{th} layer's visual features f_i^v, where f_i^v and f_i^l are the visual and linguistic part of i^{th} layer's output of SFM F_i.

$$\Lambda_{ij} = \sigma(Conv([f_i^v; f_j^{l_{avg}}])) \tag{2.2}$$

Here $\Lambda_{ij} \in \mathbb{R}^{C\times H\times W}$, $f_j^{l_{avg}} \in \mathbb{R}^C$ is the global linguistic context for j^{th} layer and is computed as length-wise average of linguistic features f_j^l, σ is the sigmoid function. Here, $f_j^{l_{avg}}$ act as a bridge to route linguistic context from j^{th} layer to spatial regions of i^{th} layer's visual hierarchy. Similarly, Λ_{ik} is computed with $i \neq j \neq k$, allowing for cross-modal exchange between all permutations of visual and linguistic hierarchical features.

Hierarchical Aggregation: After computing the affinity weights Λ_{ij}, we perform a layer-wise contextual aggregation. For each layer, visual context from other hierarchies is aggregated in the following way:

$$\begin{aligned} g_i &= f_i^v + \sum_{j\neq i} \Lambda_{ij} \circ f_j^v \\ G &= Conv3D([g_2; g_3; g_4]) \end{aligned} \tag{2.3}$$

Here, $\circ$ is element-wise product and $[;]$ represents stacking features along length dimension, ie:- $\mathbb{R}^{3\times C\times H\times W}$ dimensional feature. $g_i \in \mathbb{R}^{C\times H\times W}$ contains the relevant regions corresponding to the linguistic context from the other two hierarchies. Finally, we use 3D convolution to aggregate g_i's to include the common regions corresponding to the linguistic context from all visual hierarchies. G is the final multi-modal context for referent.

2.4.4 Mask Generation

Finally, G is passed through Atrous Spatial Pyramid Pooling (ASPP) decoder [11] and Up-sampling convolution to predict final segmentation mask S. Pixel-level binary cross-entropy loss is applied to predicted segmentation map S and the ground truth segmentation mask Y to train the entire network end-to-end.

2.5 Experiments

2.5.1 Experimental Setup

Datasets: We conduct experiments on four Referring Image Segmentation datasets. **UNC** [78] contains 19,994 images taken from MS-COCO [34] with 142,209 referring expressions corresponding to 50,000 objects. Referring Expressions for this dataset contain words indicating the location of the object. **UNC+** [78] is also based on images from MS-COCO. It contains 19,992 images, with 141,564 referring expressions corresponding to 50,000 objects. In UNC+, the expression describes the object based on their appearance and context within the scene without using spatial words. **G-Ref** [43] is also curated using images from MS-COCO. It contains 26,711 images, with 104,560 referring expressions for 50,000 objects. G-Ref contains longer sentences with an average length of 8.4 words; compared to other datasets which have an average sentence length of less than 4 words. **Referit** [31] comprises of 19,894 images collected from IAPR TC-12 dataset. It includes 130,525 expressions for 96,654 objects. It contains unstructured regions (e.g., sky, mountains, and ground) as ground truth segmentations.

Implementation Details:: We experiment with two backbones, DeepLabv3+ [11] and Resnet-101 for image feature extraction. Like previous works [76, 7, 24], DeepLabv3+ is pre-trained on Pascal VOC semantic segmentation task while Resnet-101 is pre-trained on Imagenet Classification task, and both backbone's parameters are fixed during training. For multi-level features, we extract features from the last three blocks of CNN backbone. We conduct experiments at two different image resolutions, 320×320 and 448×448 with $H = W = 18$. We use GLoVe embeddings [48] pre-trained on Common Crawl 840B tokens to initialize word embedding for words in the expressions. The maximum number of words in the linguistic expression is set to 25. We use LSTM for extracting textual features. The network is trained using AdamW optimizer with batch size set to 50; the initial learning rate is set to $1.2e^{-4}$ and weight decay of $9e^{-5}$ is used. The initial learning rate is gradually decreased using polynomial decay with a power of 0.7. We train our network on each dataset separately.

Evaluation Metrics: Following previous works [76, 7, 24], we evaluate the performance of our model using overall Intersection-over-Union (overall IoU) and Precision@X as metrics. Overall IoU metric calculates the ratio of the intersection and the union computed between the predicted segmentation mask and the ground truth mask over all test samples. Precision@X metric calculates the percentage of test samples having IoU greater than the threshold X, with $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

Method	UNC			UNC+			G-Ref	Referit
	val	testA	testB	val	testA	testB	val	test
RRN [32]	55.33	57.26	53.95	39.75	42.15	36.11	36.45	63.63
CMSA [76]	58.32	60.61	55.09	43.76	47.60	37.89	39.98	63.80
STEP [7]	60.04	63.46	57.97	48.19	52.33	40.41	46.40	64.13
BRIN [24]	61.35	63.37	59.57	48.57	52.87	42.13	48.04	63.46
LSCM [27]	61.47	64.99	59.55	49.34	53.12	43.50	48.05	66.57
CMPC [25]	61.36	64.53	59.64	49.56	53.44	43.23	49.05	65.53
MCN* [42]	62.44	64.20	59.71	50.62	54.99	44.69	-	-
BUSNet* [74]	62.56	65.61	60.38	50.98	56.14	43.51	**49.98**	-
EFN* [18]	62.76	65.69	59.67	51.50	55.24	43.01	<u>51.93</u>	66.70
SHNet((320×320)	*63.98*	*67.51*	*60.48*	*51.79*	*56.49*	*43.83*	48.95	*68.38*
SHNet* (448×448)	**65.32**	**68.56**	**62.04**	**52.75**	**58.46**	**44.12**	*49.90*	**69.19**

Table 2.1 Comparison with State-Of-the-Arts on *Overall IoU* metric, $*$ indicates results without using DenseCRF post processing. Best scores are shown in bold and the second best are shown in italics. Our method uses DeepLabv3+ backbone for both resolutions.

2.5.2 Comparison with State of the Art

We evaluate our method's performance on four benchmark datasets and present the results in Table 2.1. Since three of the datasets are derived from MS-COCO and have significant overlap with each other, pre-training on MS-COCO can give misleading results and should be avoided. Hence, we only compare against methods for which the backbone is pre-trained on Pascal VOC. Unless specified, all the approaches in Table 2.1 are at 320×320 resolution. Our approach, SHNet (SFM+HCAM), achieves state-of-the-art performance on three datasets without post-processing. In contrast, most previous methods present results after post-processing through a Dense Conditional Random Field (Dense CRF). The expressions in UNC+ avoid using positional words while referring to objects; instead, they are more descriptive about their attributes and relationships. Consistent performance gains on the UNC+ dataset at all splits showcases the effectiveness of utilizing comprehensive interactions simultaneously across visual and linguistic modalities. Similarly, our approach gains 1.68% over the next best performing method EFN [18] on the Referit dataset, reflecting its ability to ground unstructured regions (e.g., the sky, free space). We also achieve solid performance gains on the UNC dataset at both resolutions, indicating that our method can effectively utilize the positional words to localize the correct instance of an object from multiple ones. EFN [18] (underlined in Table 2.1) gives the best performance on G-Ref dataset; however, it is fine-tuned on the UNC pre-trained model. With similar fine-tuning, SHNet achieves 56.44% overall IoU, surpassing EFN by a large margin. However, such an experimental setup is incorrect, as there is a significant overlap between G-Ref test and UNC training set. Hence, in Table 2.1 we report performance on a model trained on G-Ref from scratch. Performance of SHNet is marginally below BusNet on the G-Ref dataset. Feature maps in SHNet have a lower resolution of 18×18 compared to 40×40 resolution used by other methods and that possibly leads to a drop in performance on G-Ref, which has extremely small target objects. We could not train SHNet on higher

	Method	*prec@0.5*	prec@0.6	prec@0.7	prec@0.8	prec@0.9	*Overall IoU*
1	Baseline	61.47	54.01	43.74	27.47	7.21	54.70
2	Only HCAM	68.44	61.58	52.10	35.63	9.71	59.53
3	Only SFM	72.56	66.58	57.91	40.73	12.82	62.16
4	SFM+ConvLSTM	74.34	68.89	60.67	42.95	13.35	63.30
5	SFM+Conv3D	74.07	68.74	60.50	43.14	13.58	63.16
6	SHNet w/o Glove	74.23	68.42	59.77	42.47	13.66	62.19
7	SHNet w/o P.E	74.0	68.36	59.71	43.15	13.36	63.07
8	SHNet	**75.18**	**69.36**	**61.21**	**46.16**	**16.23**	**63.98**

Table 2.2 Ablation Studies on Validation set of UNC, SHNet is the full architecture with both SFM and HCAM modules. The input image resolution is 320×320 in each case.

resolution feature maps due to memory limits induced by multi-head attention (on RTX 2080Ti GPU); however, training on higher resolution input improves results.

2.5.3 Ablation Studies

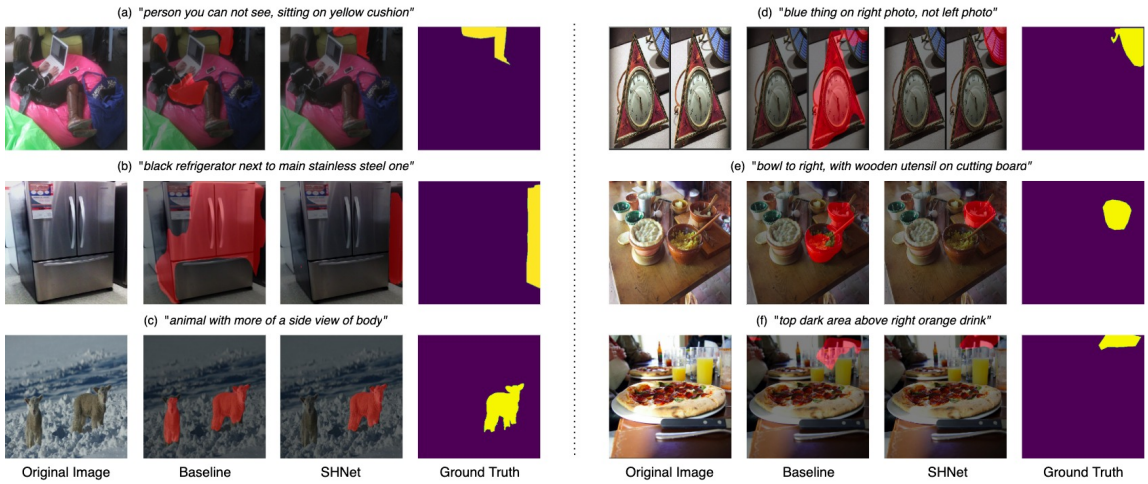


Figure 2.4 Qualitative results comparing the baseline against SHNet.

We perform ablation studies on the UNC dataset's validation split. All methods are evaluated on Precision@X and Overall IoU metrics, and the results are illustrated in Table 2.2. Unless specified, the backbone used for ablations is DeepLabv3+ trained at 320×320 resolution. The feature extraction process described in Section 3.1 is used for all ablation studies. ASPP + ConvUpsample decoder is also common to all the experiments.

Baseline: The baseline model involves direct concatenation of visual features with the tiled textual feature to result in multi-modal feature of shape $\mathbb{R}^{(C_v+C_l)\times H\times W}$. This multi-modal feature is passed as input to ASPP + ConvUpsample decoder.

HCAM without SFM: "Only HCAM" network differs with baseline method only on the fusion process of hierarchical multi-modal features. Introducing the HCAM module over baseline results in 4.83 % improvement on the Overall IoU metric and an improvement of 2.5 % on the $prec@0.9$ metric (illustrated in Table 2.2), indicating that the HCAM module results in refined segmentation masks.

SFM without HCAM: Similarly, the "Only SFM" network differs from the baseline method in how different types of visual-linguistic interactions are captured. We observe significant performance gains of 7.46 % over the baseline, indicating that simultaneous interactions help identify the referent.

SFM + X: We replace HCAM module with other multi-level fusion techniques like ConvLSTM and Conv3D. Comparing the performance of SFM+ConvLSTM with SHNet (SFM+HCAM), we observe that HCAM is indeed effective at fusing hierarchical multi-modal features (Table 2.2). For SFM+Conv3D, we stack multi-level features along a new depth dimension resulting in 3D features, and perform 3D convolution on them. The same filter is applied to different level features that result in each level feature converging on a common region in the image. SFM+Conv3D achieves a similar performance as SFM+ConvLSTM while using fewer parameters. Using Conv3D achieves higher Precision@0.8 and Precision@0.9 than ConvLSTM, suggesting that it leads to more refined maps. It is worth noting that HCAM also uses Conv3D at the end, and the additional gains of SHNet over SFM+Conv3D suggest the benefits of hierarchical information exchange in HCAM.

Glove and Positional Embeddings: We verify Glove embeddings' significance by replacing it with one hot embedding. We also validate the usefulness of Positional Embeddings (P.E.) by training a model without them. Both variants observe a drop in performance (Table 2.2), with the drop being more significant in the variant without Glove embeddings. These ablations suggest the importance of capturing word-level semantics and positional-aware features.

In Table 2.3, we present ablations with different backbones at different resolution. The results demonstrate that our approach does not heavily rely on backbone for its performance gains, as even with a vanilla Imagenet pre-trained Resnet101 backbone, not fine-tuned on segmentation task, we outperform existing methods at both resolutions. Predictably, using a backbone fine-tuned on a segmentation task gives further performance gain.

backbone	resolution	val	testA	testB
Resnet101	320 x 320	63.76	67.05	60.15
	448 x 448	64.88	68.08	60.82
DeepLabv3+	320 x 320	63.98	67.51	60.48
	448 x 448	65.29	68.56	62.04

Table 2.3 Result with different backbone at different input resolutions on UNC dataset.

We also present ablations with different aggregation modules in Table 2.4. We use the modules presented in MGATE [76], TGFE [25] and GBFM [27], for which codes were publicly available. HCAM consistently outperforms other methods by clear margins at both resolution.

Aggregation Module	Overall IOU	
	320x320	448x448
MGATE [76]	62.59	63.35
TGFE [25]	62.94	63.72
GBFM [27]	62.72	63.83
HCAM	**63.98**	**65.32**

Table 2.4 Comparing performance of recent Aggregation Modules on the UNC val dataset at different input resolutions

2.5.4 Qualitative Results

(a) Original Image (b) Only HCAM module (c) Only SFM module (d) SHNet (e) Ground Truth

Figure 2.5 Qualitative results corresponding to combinations of proposed modules. In (b) we show results when only HCAM module is used, (c) result with only SFM module being used, (d) output mask when both SFM and HCAM modules are used

"top bowl" "left plate on top" "left plate on bottom" "front bowl" "right bowl" "empty plates in center"

Figure 2.6 Output predictions of SHNet for an anchored image with varying linguistic expressions.

Figure 2.4 presents qualitative results comparing SHNet against the baseline model. SHNet localizes heavily occluded objects (Figure 2.4 (a) and (b)); reasons on the overall essence of the highly ambiguous sentences (e.g. "person you cannot see", "right photo not left photo") and; distinguishes among multiple instances of the same type of object based on attributes and appearance cues (Figure 2.4 (b), (c), and (e)). While, without any reasoning stage, the baseline model struggles to segment the correct instance and confuse it with similar objects. Figure 2.4 (d) and (f) illustrate the ability of SHNet to localize unstructured non-explicit objects like "dark area" and "blue thing". The potential of SHNet to perform relative positional reasoning is highlighted in Figure 2.4 (b), (e), and (f).

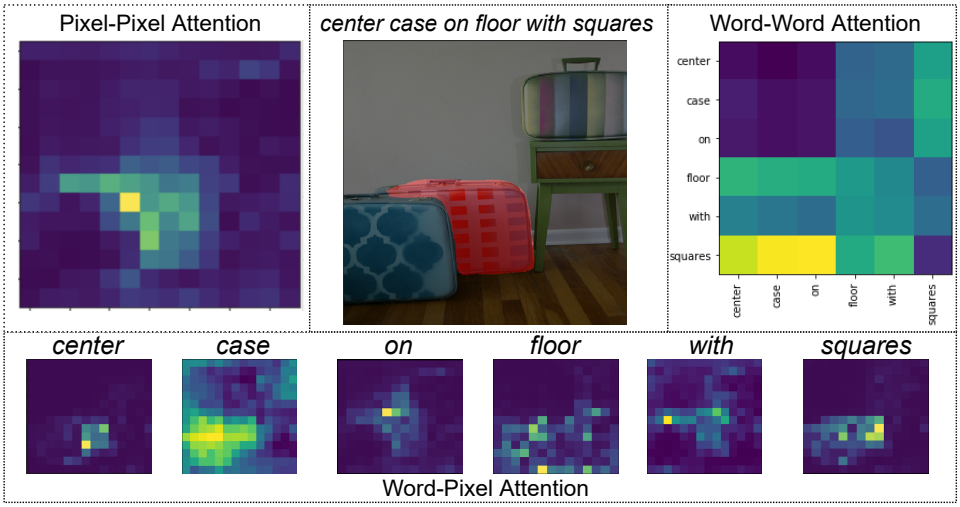


Figure 2.7 Visualization of Inter-modal and Intra-modal interactions in SFM.

We outline the contributions of both SFM and HCAM modules in Figure 2.5. "Only HCAM" network does not involve any reasoning, however, it manages to predict the left sandwich with refined boundaries. "Only SFM" network understands the concept of "the right half of the sandwich" and leads to much better output; however, the output mask bleeds around the boundaries, and an extra small noisy segment is visible. The full model benefits from the reasoning in "SFM," and when combined with HCAM facilitates information exchange across hierarchies to predict correct refined mask as output. In Figure 3.5, we anchor an image and vary the linguistic expression. SHNet is able to reason about different linguistic expressions successfully and ground them. Inter-modal and Intra-modal interactions captured by SFM are illustrated in Figure 2.7. Pixel-pixel interactions highlight image regions corresponding to the referent. For the given expression, "squares" contains the differentiating information and is assigned high importance for different words. Additionally, for each word appropriate region in the image is attended.

In Figure 2.8, we present results where our approach successfully grounded the referring expression in the image. The network is able to identify fine grained distinctive information about the referent from the referring expression, and utilize it to correctly localize the referent in complex visual scenes in (c), (d), (f) and (j). Specifically in (c), (d) and (j), we are able to identify the correct person from large group of people based on the combination of person's attribute ("dark hair"), attributes of person's clothing ("green sleeves", "no shirt" etc) and its location with respect to other objects in the image ("by the wall"). Additionally, SHNet localizes objects which are out of focus and are partially visible, ex: (b), (e), (g) and (h). We would like to point out that in these cases, rather than merely picking the most prominent objects, our network effectively incorporates the information from textual expression in visual domain to identify the less prominent correct object. In (a) and (i), the referring expressions refer

to unstructured regions in image, our network predicts these regions with refined boundaries. In (k) and (l) of Figure 2.8, the referred objects occupy extremely small region in the image space and SHNet is able to accurately locate them.

In Figure 2.9, we present some failure cases of our approach. Our approach mostly fails in cases when either the referring expression or the visual scene is ambiguous in (a), (c) and (e), the visual scene is heavily cluttered in (b) and (d), or when common sense reasoning is required like (f). For example: the expression in (a), "chair at the end of table on the left" is itself ambiguous and non-specific, as there are two chairs at the end of table on left side. Similarly, in (b) their are multiple keyboards with a mouse on top and our method predicts one of the keyboards on the left with a partial black mouse on the top. In (d), the plant branch on the left is barely visible and also a lot of clutter is present. It is noteworthy, that in each case, our approach predicts a well segmented and refined output and the class predictions are also correct (an umbrella, a chair, a bottle, a keyboard etc.).

2.6 Conclusion

In this work, we tackled the task of Referring Image Segmentation. We proposed a simple yet effective SFM to capture comprehensive interactions between modalities in a single step, allowing us to simultaneously consider the contextual information from both modalities. Furthermore, we introduced a novel HCAM module to aggregate multi-modal context across hierarchies. Our approach achieves strong performance on RIS benchmarks without any post-processing. We present thorough quantitative and qualitative experiments to demonstrate the efficacy of all the proposed components.

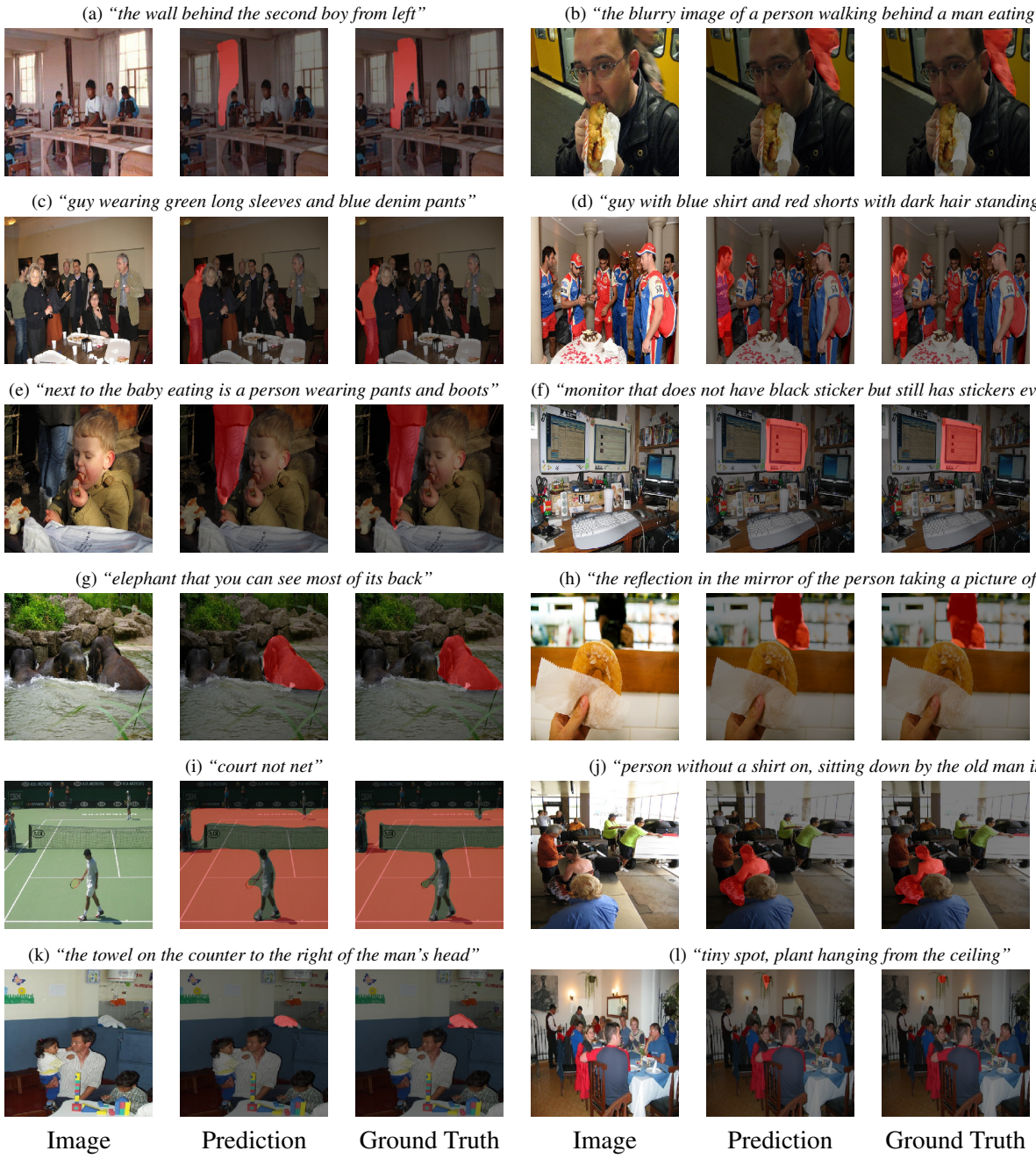


Figure 2.8 Qualitative examples where our approach successfully localized the referred object.

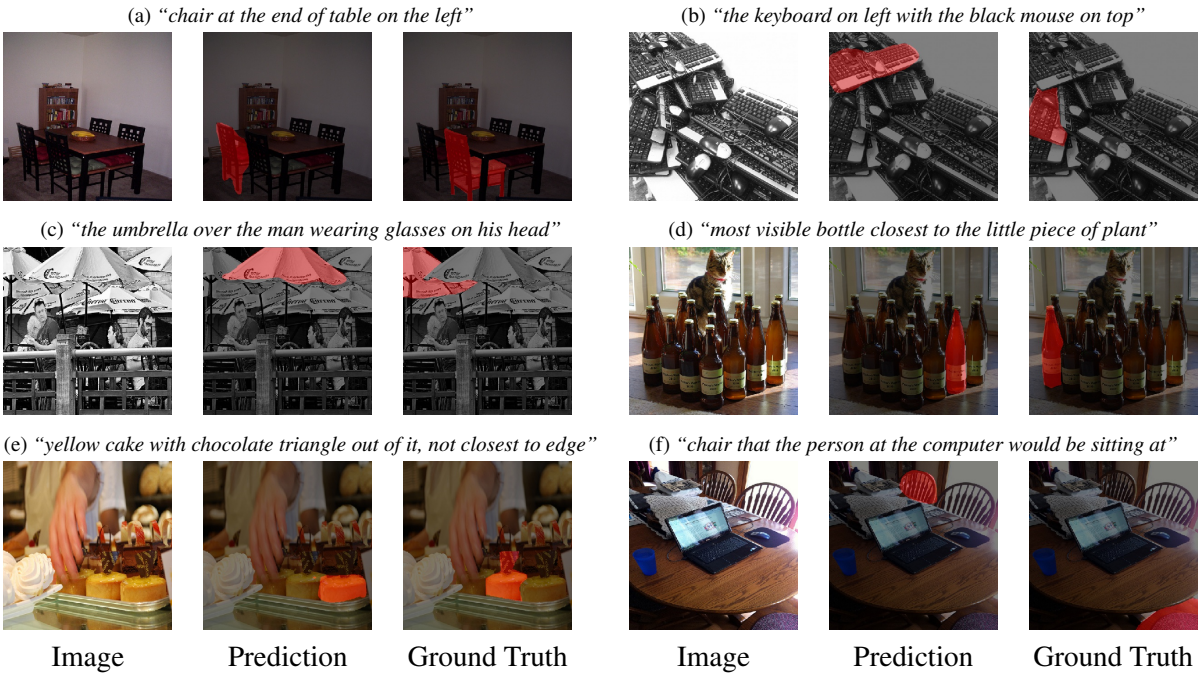


Figure 2.9 Qualitative examples where our approach failed to localize the referred object.

Chapter 3

Grounding Linguistic Commands to Navigable Regions

3.1 Introduction

Autonomous Driving (AD) is concerned with providing machines with driving capabilities without human intervention. Much of the existing work on autonomous driving has focused on modular pipelines, with each module specializing in a separate task like detection, localization, segmentation and tracking. Collectively, these tasks form the vehicle's active perception module, enabling it to perform driver-less navigation with some additional help from prior generated detailed high definition maps of the route. However, the current setup does not allow the capability to intervene and augment the vehicle's decision-making process. For example, post reaching the destination, the rider may want to give specific guidance on the place to park the car suiting his/her convenience, e.g. "park between the yellow and the red car on the left".

Similarly, sometimes the rider may wish to intervene to resolve ambiguities or to perform the desired action, e.g. "the road appears to be blocked, please move to the left lane" or "I see my friend walking on the left, please slow down and pick him up". In a chauffeur-driven car, the above scenarios are commonplace, as a human can easily understand the natural language commands and manoeuvre the car accordingly. In this work, we aim to extend similar abilities to a self-driving vehicle, i.e. the vehicle takes the natural language command and the current scene as input and predicts the region of interest where the car must navigate to execute the command. A downstream planner can take this region as input and predict the trajectory or set of manoeuvres to perform the desired navigation.

One of the fundamental tasks necessary to attain the above capabilities is comprehending the natural language command and localising it in the visual domain. The problem is formally known as visual grounding, and it has seen a surge of interest in the recent past. The interest is primarily driven by the success of deep learning models in computer vision and natural language processing. Most of the current literature in visual grounding focuses on localising an object of interest. The object of interest can be grounded either using bounding boxes (Referring Expression Comprehension) or using segmentation masks (Referring Image Segmentation). We focus on the latter type of grounding, but instead of grounding objects, we ground regions of interest on the road. Grounding regions of interest

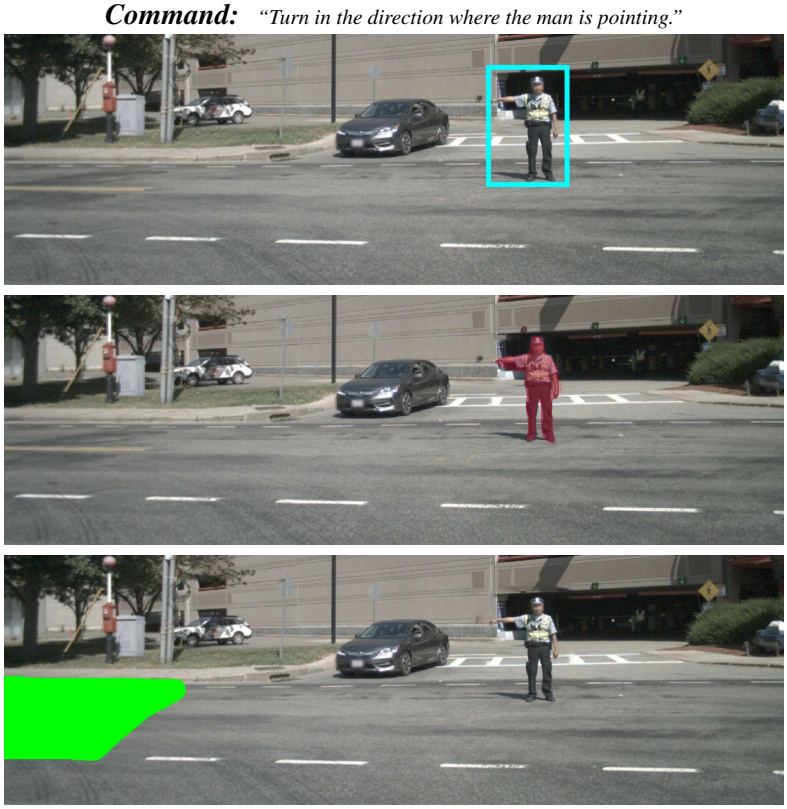


Figure 3.1 Given a natural language command, REC (top image) predicts a bounding box (cyan box) around the referred object and RIS (the middle image) predicts a segmentation map around the referred object. In the context of an AD application, such predictions are not immediately amenable to downstream tasks like planning. E.g. predicting the man in the above example does not indicate where the car should go. In contrast, our work aims to directly predict regions on the road given a natural language command (green colour annotation, bottom image).

are more natural from a navigation point of view for self-driving vehicles than the grounding of objects. Even if the referred object is correctly grounded, it leaves ambiguity on where to take the vehicle. In contrast, the task of Referring Navigable Regions proposed in our work provides feasible areas as a goal point. A motivating example is illustrated in Figure 3.1.

3.2 Contributions

To this end, we introduce a novel multi-modal task of Referring Navigable Regions (RNR), intending to ground navigable regions on the road based on natural language command in the vehicle's front camera view. Compared to RIS task, RNR task involves two-level understanding of the scene. In the first level, the referring object has to be identified and in second level the appropriate region for navigation

has to be identified based on the referred object. For instance, consider the command "park beside the white car near the tree", in addition to locating the "white car" near the "tree", the RNR task also has to predict an appropriate region where the command can be executed. Consequently, we propose a new dataset, Talk2Car-RegSeg, for the proposed task. This dataset is built on top of the existing Talk2car [14] dataset. In addition to the existing image-command pairs, we provide segmentation masks for the regions on the road where the vehicle could navigate to execute the command. We benchmark the proposed dataset with a transformer-based grounding model that can capture correlations between visual and linguistic features through the self-attention mechanism. We compare the proposed model against a set of baselines and present thorough ablation studies. We highlight the proposed task's practicability through a downstream planning module that computes a navigation trajectory to the grounded region. To summarize, the main contributions of this paper are the following:

- We introduce the novel task of RNR for applications in autonomous navigation.
- We present a new Dataset, Talk2Car-RegSeg, for this task. Here, we augment the existing Talk2car dataset with segmentation masks for navigable regions corresponding to the command.
- We benchmark the dataset using a novel transformer based model and a set of baseline approaches. We present thorough ablations and analysis studies on the proposed dataset (e.g. action type of commands, the length of commands) to assess its applicability in realistic scenarios.

3.3 Related Work

Referring Expression Comprehension: Referring Expression Comprehension (REC) predicts a rectangular bounding box in an image corresponding to the input phrase or the sentence. While object detection [52, 51] predicts bounding boxes for a pre-defined set of categories, REC does not limit on a category list. Nonetheless, the task of REC does take inspiration from the object detection pipeline. In the most commonly used framework, a set of bounding box region proposals are first generated and then evaluated against the input sentence [50, 53]. In the robotics community, significant progress has been made on using REC in Human-Computer Interactions [61, 62]. REC has also been explored on autonomous driving applications, following the introduction of the Talk2Car dataset [14]. Rufus *et al.* [56] use softmax on cosine similarity between region-phrase pairs and employ a cross-entropy loss. Ou *et al.* [46] employ multimodal attention using individual keywords and regions. Despite significant progress in REC, bounding box based localization is not accurate enough to capture the shape of the referred object and struggle with objects at a small scale. Furthermore, just predicting the bounding box is insufficient for the task of navigation (as illustrated in Figure 3.1).

Referring Image Segmentation: Referring Image Segmentation (RIS) task was introduced in [22] to alleviate the problems associated with REC by predicting a pixel-level segmentation mask for the referring object based on the referring expression. [35] propose convolutional multimodal LSTM to

encode the sequential interactions between individual words and pixel-level visual information. [60] utilize query attention and key-word-aware visual context to model relationships among different image regions, according to the corresponding query. More recent works, [27] model multimodal context by cross-modal interaction and guided through a dependency tree structure, [26] progressively exploits various types of words in the expression to segment the referent in a graph-based structure. In contrast to existing works on RIS that directly refer to objects in an image, we ground the region adjacent to the object to provide navigational guidance to a self-driving vehicle. To the best of our knowledge, our work is the first paper to explore the referring image segmentation in the context of autonomous driving and propose the task of Referring Navigable Region.

Language Based Navigation: Most of the literature on language-based navigation has focused on indoor navigation [58, 79, 70]. Typically the input to these approaches is a longer text (a paragraph), and the goal is to reach the required destination in an indoor 3D environment (long trajectory prediction). Shah *et al.*[58] utilized attention over linguistic instructions conditioned on the multi-modal sensory observations to focus on the relevant parts of the command during navigation task. [79] approach the language-based navigation task as a sequence prediction problem. They translate navigation instructions into a sequence of behaviours that a robot can execute to reach the desired destination. Wang *et al.* [70] enforces cross-modal grounding both locally and globally via reinforcement learning.

Sriram *et al.* [65] attempt language-based navigation in an autonomous driving scenario. They generate trajectory based on natural language command by predicting local waypoints. However, their work limits to eight specific behaviours like *take left, take right, not left*, etc. The only object considered in their work is a traffic signal. Our work considers much richer language instructions encompassing many objects. Furthermore, RNR predicts a segmentation map instead of a single local waypoint or a trajectory corresponding to a set of sentences. Segmentation masks unlike single waypoint encourage multiple trajectory possibilities and options to navigate into that region for a downstream planning or navigation task.

3.4 Dataset

The proposed Talk2Car-RegSeg dataset is built on top of the Talk2Car dataset, an object referral dataset containing commands written in natural language for self-driving cars. The original dataset had textual command with a specific action, referring to an object in the image, and the object of interest was referred to using a bounding box. However, for AD applications, as referring directly to objects is not amenable for downstream tasks like planning, we augmented the original dataset with segmentation masks corresponding to navigable regions. The newly created Talk2Car-RegSeg dataset has 8349 training and 1163 validation image-command pairs, similar to those used in the original dataset. We observed that the commands in Talk2Car's validation set are very complex as they are verbose, and in a significant number of cases, there were more than one actions in a single command ex: "we need to turn right instead of left, as soon as this truck pulls forward, move over to the right lane behind it."

We first present results on the full validation set; however, to evaluate the performance in a controlled setting, we also curated a novel test split (Test-RegSeg). Test-RegSeg contains 500 randomly selected images from the validation set with newly created commands. The commands in the Test-RegSeg split are simplified and straightforward. We present results, baseline comparisons, and ablations on both the complex instruction validation set and the curated simpler instruction set (Test-RegSeg). In the rest of the paper, we consider Test-RegSeg as our test set. The dataset and the code-base will be released at (rnr-t2c.github.io). In the next section, we describe the dataset creation process.

3.4.1 Dataset Curation

The authors of the paper manually annotated the navigable regions in each image based on the linguistic command. A simple Graphical User Interface (GUI) was created to make the annotation process straightforward. In the GUI, each annotator sees the image, the linguistic command, and the bounding box for the referred object in the scene. We used ground truth bounding boxes from the original Talk2Car dataset as a reference to identify the referred object in the scene to resolve ambiguities and only focus on annotating regions of interest.

To verify the quality of annotations, we hired a group of three students from the institute for the role of annotation reviewers. All the reviewers were briefed on the task and were asked to ensure that all feasible regions for navigation were annotated in the image. Depending on the reviewers' assessment, each annotation could be either accepted or sent for re-annotation. An annotation was accepted if at least two reviewers concurred on it. In the other cases, images were sent for re-annotation with reviewer comments for annotation refinement. This process was repeated iteratively until all annotations were qualitatively and logically acceptable.

Since navigation is a flexible activity in terms of different ways of performing it, we involved multiple people as annotators and reviewers to capture different perspectives and incorporate those in our dataset.

3.5 Approach

Given an image I from a front-facing camera on the autonomous vehicle and a natural language command Q, the goal is to predict the segmentation mask of the region in the image where the vehicle should navigate to fulfil the command. Here, the command Q corresponds to a navigable action in the image. Compared to the traditional task of Referring Image Segmentation, the proposed task is more involved as the ground truth masks are unstructured. To correctly identify the regions of interest, the model should be able to learn correlations between words in commands and regions in the image. We propose two models for this task, a baseline model and another transformer-based model. The feature extraction process is the same for both models. They only differ in multi-modal fusion and context modelling. We describe the feature extraction process in the next section and describe each model in the subsequent sections.

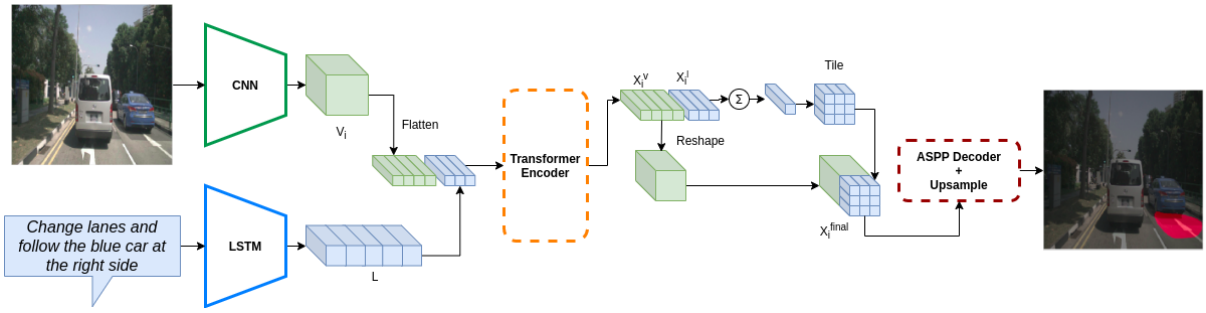


Figure 3.2 Network architecture for the Transformer Based Model (TBM).

3.5.1 Feature Extraction

We extract visual features from image using a DeepLabV3+ [9] backbone pre-trained on semantic segmentation task. Since hierarchical features are beneficial for semantic segmentation, we derive hierarchical features V_i of size $C_i \times H_i \times W_i$ with $i \in \{2, 3, 4\}$, corresponding to last 3 layers, namely $layer2$, $layer3$ and $layer4$ of CNN backbone. Here H_i, W_i and C_i correspond to height, width and channel dimension of visual features corresponding to each level. Each V_i's are transformed to same spatial resolution $H_i = H$, $W_i = W$ and channel dimension $C_i = C_v$ using 3×3 convolutional layers. We initialize each word in the linguistic command with the GloVe word embedding, which are then passed as input to LSTM encoder, to generate linguistic feature for the command. We denote the linguistic feature as $L = \{l_1, l_2, ..., l_T\}$, where T is the number of words in the command and $l_i \in \mathbb{R}^{C_l}$, $i \in \{1, 2, ..T\}$ is the linguistic feature for the i-th word. In all our experiments, $C_v = C_l = C$ and $H = W = 14$.

3.5.2 Baseline Model

Our baseline model is inspired from [22], we first compute the command feature $L_{avg} \in \mathbb{R}^{C_l}$ by averaging all the word features l_i in L. In order to fuse visual features with linguistic features, we repeat the command feature L_{avg} along each spatial location in the visual feature map and then concatenate the features from both modalities along channel dimension to get a multi-modal feature M_i of shape $\mathbb{R}^{(C_v+C_l)\times H\times W}$. Since the number of channels, $C_v + C_l$ can be large, we apply 1×1 convolution to M_i reduce the channel dimension to C, resulting in final multi-modal feature $X_i^{final} \in \mathbb{R}^{C\times H\times W}$.

3.5.3 Transformer Based Model

Our baseline model has few shortcomings: (1) the word-level information is lost when all word features are averaged to get the command feature. (2) multi-modal context is not captured effectively with a concatenation of visual and linguistic features. To address these shortcomings, we propose a

transformer-based model (TBM). We borrow from the architecture of DETR [5] for our transformer based model. Specifically, we adopt their transformer encoder, and along with image features V_i, we also pass textual feature L as input by concatenating features from both modalities along length dimension, resulting in multi-modal feature M_i of shape $\mathbb{R}^{C\times(HW+T)}$, T is the number of words in the input command. M_i is passed as input to the transformer encoder, where self-attention enables cross-modal interaction between word-level and pixel-level features, resulting in multi-modal contextual feature X_i of the same shape as M_i. Since all word features are utilized during the computation of X_i, the word -level information is preserved, and because of inter-modal and intra-modal interactions in the transformer encoder, the multi-modal context is captured effectively. To predict a segmentation mask from X_i, we need to reshape it to the same spatial resolution as V_i, i.e., $H \times W$. So, X_i is separated into attended visual features, X_i^v and attended linguistic features, X_i^l of dimensions $\mathbb{R}^{HW\times C}$ and $\mathbb{R}^{T\times C}$, respectively. X_i^l is averaged across length dimension and concatenated with X_i^v along the channel dimension and reshaped to result in a feature vector of shape $\mathbb{R}^{2C\times H\times W}$. Finally, 1×1 convolution is applied to give final multi-modal feature $X_i^{final} \in \mathbb{R}^{C\times H\times W}$.

3.5.4 Mask Generation

To generate the final segmentation mask, we stack X_i^{final} for all levels and pass them through Atrous Spatial Pyramid Pooling (ASPP) Decoder from [11]. We use 3×3 convolution kernels followed by bilinear upsampling to predict the segmentation mask at a higher resolution. Finally, sigmoid non-linearity is applied to generate pixel-wise labels for segmentation mask Y. Both baseline and transformer-based models are trained end-to-end using binary cross-entropy loss between predicted segmentation mask Y, and the ground truth segmentation mask G.

3.6 Experiments

Implementation Details: We use DeepLabV3+ [9] with ResNet-101 as backbone for visual feature extraction. Our backbone is pre-trained on the Pascal VOC-12 dataset with the semantic segmentation task. Input images are resized to 448×448 spatial resolution. We use 300d GloVe embeddings pre-trained on Common Crawl 840B tokens [47]. The maximum length of commands is set to $T = 40$ and for both visual and linguistic features, channel dimension $C = 512$. Batch size is set to 64, and our models are trained using AdamW optimizer with weight decay of $5e^{-4}$, the initial learning rate is set to $1e^{-4}$ and gradually decreased using polynomial decay by a factor 0.5.

Evaluation Metrics: In the proposed dataset, the ground truth segmentation masks incorporate all viable regions of interest for navigation, so any point inside the annotated region can be used as a target destination. Considering this aspect of our dataset, we evaluate our models' performance on three metrics, namely, Pointing Game, Recall@k and Overall IOU. Pointing Game Metric (PGM) indicates the per cent of examples where the highest activated point lies inside the ground truth mask. It is

calculated in the following way:

$$PGM\ Score = \frac{\#\ of\ hits}{total\ examples} \tag{3.1}$$

A hit occurs when the highest activated pixel lies inside the ground truth segmentation mask. It is possible that in some cases, the point with the highest activation is slightly outside the annotated ground truth region. However, the overall prediction is almost correct. Recall@k metric is used to underscore the performance of models in such scenarios. Recall@k metric is calculated as the proportion of examples where at least one of the top-k points lies inside the ground truth mask. Finally, we also show results with the Overall IOU metic. Previous works commonly use the Overall IOU metric [22, 27, 26] for RIS task, it is calculated as the ratio of total intersection and total union between the predicted and ground truth segmentation masks across all examples in the dataset.

Method	Recall @k for PGM											
	$k=5$		$k=10$		$k=50$		$k=100$		$k=500$		$k=1000$	
	Val Set	Test Set	Val Set	Test Set	Val Set	Test Set	Val Set	Test Set	Val Set	Test Set	Val Set	Test Set
Baseline	51.84	69.80	52.71	69.80	55.29	72.20	56.92	73.80	64.49	79.60	69.64	83.80
TBM	59.67	77.00	60.53	78.20	63.19	79.60	64.91	81.80	72.65	86.60	78.07	90.20

Table 3.1 Recall@k metric for the validation and test set

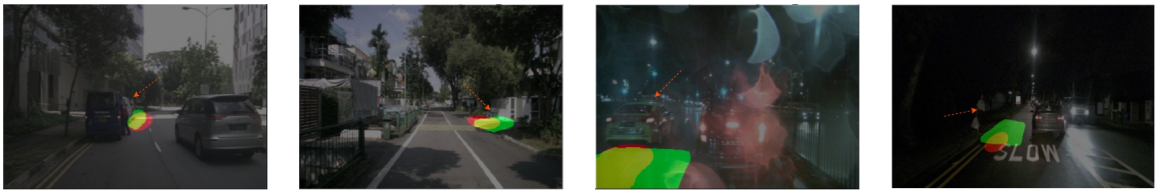
3.6.1 Experimental Results

In this section, we present the experimental results on different evaluation metrics. For all metrics, we compute the results on both validation and test split.

Pointing Game: Results on pointing game metric are presented in Table 3.6.1. First, we compare against a centre baseline to showcase the diversity of localization of annotated regions and ensuring that our dataset is free of centre-bias. In this baseline, the image's centre point is considered as the point with the highest activation for the pointing game metric. PGM score is 5.07%, and 6.61% for this baseline on validation and test splits, respectively, thus clearing our dataset for centre-bias. Next, we compare against the baseline model presented in Section 3.5.2. Our baseline model gives a PGM score of 49.78% and 66.80% on validation and test split, respectively. The test split score is high as the commands for the test split are simple and concise compared to those in the validation split. The Transformer-Based Model (TBM) gives higher numbers than both the baselines on both splits. We observe an improvement of 8% and 10% over the baseline model for validation and test split, respectively. This improvement indicates the benefits of using the proposed multi-modal attention in the transformer-based approach, which can effectively model word-region interactions.

Recall@k: Since our model mostly predicts connected and contiguous segmentation masks, Recall@k metric indicates if we can approximately locate the correct area (where the highest activation point is near the ground truth region). Results for this metric are tabulated in Table 3.6, we consider values of k=$\{5, 10, 50, 100, 500, 1000\}$. Recall@$1$ is the same as pointing game metric as in both cases, we pick the point with the highest activation. As expected, the metric performance increases with the value of

k. For transformer-based model, at $k = 1000$, metric score is 78.07% and 90.20% for validation and test splits, respectively. 1000 pixels account for $\sim 0.5\%$ of the overall pixels at the considered resolution. Hence, Recall@1000 metric suggests that we can approximately locate the correct area 90.20% of the time when using simpler and straightforward commands. This demonstrates the effectiveness of our approach and how we are able to reduce the search space for feasible regions for navigation significantly.



"Pull over by the white van" *"Park on the right by the green bin"* *"Follow the yellow car that is in the left lane"* *"pick up the woman on your left"*

Figure 3.3 Qualitative Results for Successful Groundings. Our TBM network is able to ground the appropriate regions even in cases where the referred objects are barely visible. Red arrow is used to indicate the location of these referred objects.

"Do not turn right as I said, carry on straight so I can talk to that person" *"Continue straight so I can talk to that person".* *"Wait for the moped to continue before we turn left"* *"Wait for the moped to leave"*

Figure 3.4 Differences between the network performance on the original Validation set and the newly created Test split. For each image pair, example on the left is from the Validation split and one on the right is from the Test split with simplified commands. The "person" in left pair of images is indicated using a red arrow.

Overall IOU: Since any point inside the annotated region can be considered as a target destination, computing the overall IOU metric that is normally used in segmentation literature cannot serve as an adequate performance measure and is only an indicative measure. For example. if there are three parking slots available, even if the model predicts one of them, the prediction is correct, however, the IOU might be low. The results presented in Table 3.6.1 illustrate this aspect. For the transformer-based model, the IOU metric is 22.17% for validation split and 30.61% for the test split. The numbers for test split are significantly better than those in validation split because of the simplicity of commands in test split. This metric illustrates the differences between RNR and RIS task and shows that the same metric cannot be used to judge the performance across these tasks.

Method	PGM		Overall IOU	
	Val set	Test set	Val set	Test set
Baseline	49.78	66.80	19.88	29.28
TBM	58.03	76.60	22.17	30.61

Table 3.2 PGM and Overall IOU for the validation and test set

3.6.2 Ablation Studies

In this section, we elaborate on the ablation and analysis studies performed on the proposed dataset and transformer-based model. We study various aspects of linguistic commands in the proposed Talk2car-RefSeg dataset on model performance. Specifically, we analyse the grounding performance of our model based on (1) the length of command and (2) the action specified in the command. As the commands in the test split are shorter than those in validation split, we conduct experiment (2) on the test split. Whereas based on the verbose nature of commands in the validation split, experiment (1) is conducted on the validation split. We used both baseline and transformer-based models and the pointing game metric for all the ablation studies.

Method	PGM score on the Val set		
	$T < 10$	$10 \leq T < 20$	$T \geq 20$
Baseline	52.09	48.55	44.00
TBM	60.00	57.06	52.00

Table 3.3 PGM for the validation data w.r.t. command length where $T =$ number of words in a command

Based on Command Length: We categorise the commands based on their length and present the ablation experiments in Table **??**. All commands are grouped into three buckets, {0-10, 10-20, $\geq$20} based on their length. We observed that as the command length increases, the performance on the pointing game metric decreases. The performance gap between the first two buckets is $\sim$3% , and that between the last two buckets is $\sim$5% in TBM. Since the commands in the validation split are long and complex, the network faces difficulties in grounding navigable regions for them. Some of the original talk2car dataset's commands contain unnecessary information from a grounding perspective, like addressing people using proper nouns. Because of this reason, we proposed a separate test split with concise commands. Length based grouping of commands in test split is not possible as the majority ($\sim$78%) of them are less than 10 words long.

Based on Action type: Next, we classify each command to fixed basic action/manoeuvre categories and present the results on the pointing game metric in Table **??**. For "lane change" and "turning" type of commands, our network can correctly predict the navigable region with high accuracy of 84.62% and 86.59%, respectively. For "parking" based commands, we get a pointing game score of 75.12%. Parking is a challenging action to evaluate based on the Pointing game metric. In our dataset, the annotation mask is often relatively small for these cases, especially so when referring to a far away parking slot. The highest performance is observed on "follow" type commands, where the metric is

Method	PGM score on the test set					
	Stop/ Park	Follow	Turn	Maintain Course	Go Slow/Fast	Change lanes
Baseline	62.83	72.91	74.19	50.00	76.31	68.74
TBM	75.12	88.23	86.59	83.34	77.15	84.62

Table 3.4 PGM on the test set with commands for various maneuvers

88.23%. Commands with "follow" action is easier to ground as in most cases, the navigable region is just behind the referred object (hence are less ambiguous). Results on these basic action/manoeuvre specific commands indicate the generality and practicality of our approach in realistic scenarios.

3.6.3 Qualitative Results

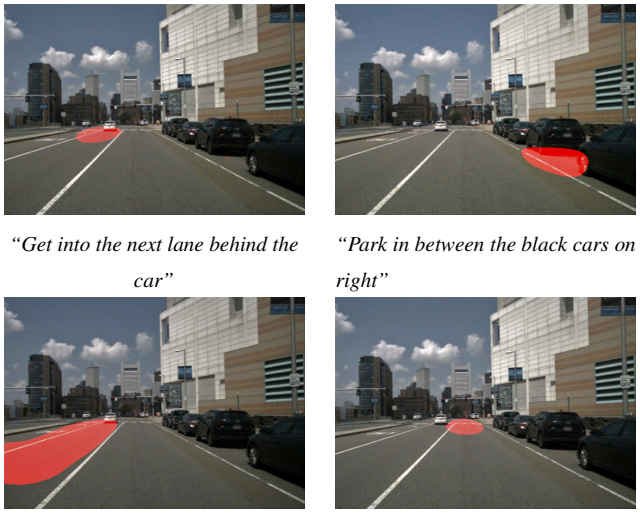
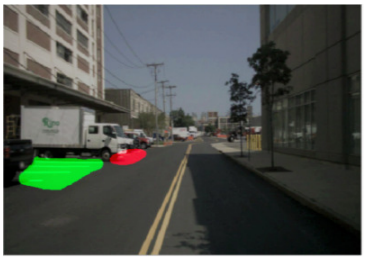


Figure 3.5 Qualitative Results for same image with different linguistic commands. Our network can successfully predict the correct navigable regions for new commands, highlighting its effectiveness in adapting to new commands flexibly.

In this section, we present the Qualitative results of our transformer-based approach. For all the example images in this section, Green, Red and Yellow signify the ground truth mask, the predicted mask and their intersection, respectively. Success cases of our approach are demonstrated in Figure 3.3. The model successfully correlates textual words with regions in the image, ex: in the leftmost image, the model can successfully ground the region beside the white van, which is barely visible. Similarly, in the second image from the left, there are two green bins, the model can successfully resolve the ambiguity. The last two images demonstrate the performance of our model during night-time. In these cases, the referred object is barely visible, but the model can still infer the correct region.

Next, we showcase the differences between the original validation split and the newly created test split in Figure 3.4. For the leftmost image, the command is a bit confusing as there is a subtle negation

"Pull over near the first trailer on the right"



"Park next to the white truck on the left side of the road"

Figure 3.6 Qualitative Results for Failure Cases. Even though the network fails to identify correct regions, it predicts a reasonable region near the referred object without knowing the parking rules.

involved. In order to resolve these issues, the model should be trained on training data with a large number of such instances. However, with the simplified command, the model predicts the correct region. Similarly, for the second image, the model gets confused when there are multiple action words in the linguistic command. In this particular case, the model correctly predicted two regions corresponding to both actions "wait" and "turn left" despite the ambiguity in the command. This underscores our network's capability in effectively modeling the word-region interactions. After simplifying the command to include only one action "wait", the model correctly predicted the corresponding navigable region.

In Figure 3.5, we further scrutinize our network by fixing an image and modifying the commands to correspond to different actions. Our network is able to incorporate the changes in command and successfully reflect them in the predicted map, highlighting the network's versatility in understanding the intent of various textual commands for the same visual scene. This result showcases the controllability aspect of our network, which is highly valuable for AD applications.

Some failure cases of our approach are shown in Figure 3.6. The results suggest that our model is able to locate the "trailer" (in the first image) and the "white truck" (in the second image). However, it fails to predict the navigable regions accurately. Looking closely, in the first case, the model is also able to understand the sub-phrases "left side of the road" and "next to white truck"; however, it predicts a place that is not appropriate for parking. The results clearly indicate the difficulty in RNR, even after correctly grounding the referred object.

3.7 Navigation and Planning

We show a downstream application wherein the navigable region output by the network is made use of by a planner to navigate to the centre of the region. While there are many potential ways of interpreting the navigable region by a downstream task, for example, one could use this as an input to a waypoint prediction network similar to [64], in this effort, we proceed with the straightforward interpretation of choosing the region centre as the goal location.

Figure 3.7 The first row corresponds to the original image and command pairs. The second row corresponds to the predicted segmentation masks (in red) overlaid onto the images. The third row shows a feasible sample trajectory to the centre point in the predicted navigable region as a goal point

First, we extract the ground plane from the LiDAR scan. Then we use LiDAR camera calibration to project the pixels corresponding to the grounded area in the image to the ground plane in the LiDAR scan. Finally, we use an RRT based sampling algorithm to construct a path to the point in 3D corresponding to the centre pixel of the region. This results in executable trajectories that appear visibly acceptable, as shown in the planned trajectories of Figure 3.7 for a few samples from our dataset. More involved integration to an AD application is a natural extension of this effort which will be tackled in future work.

3.8 Conclusion

This paper introduced the novel task of Referring Navigable Regions (RNR) based on linguistic commands to provide navigational guidance to autonomous vehicles. We proposed the Talk2Car-RegSeg dataset, which incorporates binary masks for regions on the road as navigational guidance for linguistic commands. This dataset is the first of its kind to enable control of autonomous vehicle's navigation based on linguistic commands. Furthermore, we propose a novel transformer-based model and present

thorough experiments and ablation studies to demonstrate the efficacy of our approach. Through a downstream planner, we showed how RNR task is apt for autonomous driving applications like trajectory planning compared to the RIS task. This is the first such work which has proposed RNR and showcased its direct relevance to AD applications. In this work, we focused on single frames for grounding; future work should focus on grounding at the video-level, as it is a more realistic setting for commands with temporal constraints.

Chapter 4

Conclusion

In this thesis, we tackled the problem of visual grounding and explored its utility for the task of vision language navigation. In particular, we propose a novel architecture for the Referring Image Segmentation (RIS) task, which requires predicting a segmentation mask corresponding to the object referred to by the linguistic expression. We find that both intra-modal (word-pixel) and inter-modal (word-word and pixel-pixel) interactions are needed to model the relationship between visual and linguistic modalities. Further, these multi-modal interactions are captured in a synchronous manner to avoid semantic errors encountered by existing approaches that either miss some of the interactions or capture them sequentially, resulting in error propagation. Additionally, we effectively utilise the hierarchy associated with the visual features to enable features at each hierarchy to focus on spatial regions corresponding to the referred object and predict a refined segmentation mask. We benchmark the proposed approach on multiple RIS datasets, achieving considerable performance gains over the existing state-of-the-art (SOTA) methods.

We then turn to the Vision-Language Navigation (VLN) task, which requires performing autonomous navigation based on the language commands. Existing approaches to VLN treat the task as a sequence-to-sequence prediction or a reinforcement learning problem; further, they suffer from a major limitation of interpretability as their networks are essentially a black box. Instead, we propose a paradigm shift towards visual-grounding-based solutions, which by virtue of their design, provide interpretability by localising the linguistic command in the visual scene, thus improving human-machine interaction. We propose the novel task of Referring Navigable Regions (RNR) for language-based autonomous driving, which grounds regions of interest on the road based on the language command. Further, we introduce a novel Talk2Car-RegSeg dataset, which incorporates segmented regions on the road corresponding to the linguistic command. To perform navigation, we utilised an external motion planner, which takes a point on the road sampled from the predicted navigable region as input. Finally, through extensive qualitative and quantitative ablations, we showcase the effectiveness and practicality of the proposed approach.

Related Publications

- **Kanishk Jain**, Vineet Gandhi
  Comprehensive Multi-Modal Interactions for Referring Image Segmentation accepted at Findings of the Association for Computational Linguistics: ACL 2022, May 2022, Dublin, Ireland.

- Nivedita Rufus*, **Kanishk Jain***, Unni Krishnan R Nair*, Vineet Gandhi, K Madhava Krishna
  Grounding Linguistic Commands to Navigable Regions accepted at IEEE International Conference on Intelligent Robots and System: IROS 2021, Oct 2021, Prague, Czech Republic.

Bibliography

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[6] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[7] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu. See-through-text grouping for referring image segmentation. In *ICCV*, 2019.

[8] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[10] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[12] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 104–120, Cham, 2020. Springer International Publishing.

[13] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M.-F. Moens. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*, 2019.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

[17] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[18] G. Feng, Z. Hu, L. Zhang, and H. Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15506–15515, June 2021.

[19] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[20] J. Fu, A. Korattikara, S. Levine, and S. Guadarrama. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*, 2019.

[21] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[22] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV*, 2016.

[23] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. *CVPR*, 2016.

[24] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, 2020.

[25] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, 2020.

[26] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, 2020.

[27] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, and J. Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, 2020.

[28] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*, 2015.

[29] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[30] A. Karpathy, A. Joulin, and L. F. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014.

[31] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[32] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018.

[33] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[34] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[35] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017.

[36] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[37] Y. Liu, B. Wan, X. Zhu, and X. He. Learning cross-modal context graph for visual grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11645–11652, Apr. 2020.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[39] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[40] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[41] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

[42] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[43] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[44] K. Nguyen, D. Dey, C. Brockett, and B. Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12537, 2019.

[45] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, 2015.

[46] J. Ou and X. Zhang. Attention enhanced single stage multimodal reasoner. In *ECCV Workshops*, 2020.

[47] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014.

[48] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[49] B. A. Plummer, P. Kordas, M. H. Kiapour, S. Zheng, R. Piramuthu, and S. Lazebnik. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[50] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[51] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[52] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[53] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016.

[54] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016.

[55] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[56] N. Rufus, U. K. R. Nair, K. M. Krishna, and V. Gandhi. Cosine meets softmax: A tough-to-beat baseline for visual grounding. In *ECCV Workshops*, 2020.

[57] R. Schumann and S. Riezler. Analyzing generalization of vision and language navigation to unseen outdoor areas. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7519–7532, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[58] P. Shah, M. Fiser, A. Faust, J. C. Kew, and D. Hakkani-Tur. Followne: Robot navigation by following natural language directions with deep reinforcement learning. *ICRA*, 2018.

[59] H. Shi, H. Li, F. Meng, and Q. Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, 2018.

[60] H. Shi, H. Li, F. Meng, and Q. Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, 2018.

[61] M. Shridhar and D. Hsu. Grounding spatio-semantic referring expressions for human-robot interaction. *arXiv preprint arXiv:1707.05720*, 2017.

[62] M. Shridhar, D. Mittal, and D. Hsu. Ingress: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, 39(2-3):217–232, 2020.

[63] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[64] H. Song, W. Ding, Y. Chen, S. Shen, M. Y. Wang, and Q. Chen. Pip: Planning-informed trajectory prediction for autonomous driving. In *ECCV*, 2020.

[65] N. N. Sriram, M. Tirth, K. Jayaganesh, G. Vineet, Bhowmick, Brojeshwar, and K. M. K. Talk to the vehicle: Language conditioned autonomous navigation of self driving cars. In *IROS*, 2019.

[66] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[67] A. Tao, K. Sapra, and B. Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[69] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[70] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *CVPR*, 2019.

[71] J. Xiang, X. Wang, and W. Y. Wang. Learning to stop: A simple yet effective approach to urban vision-language navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 699–707, Online, Nov. 2020. Association for Computational Linguistics.

[72] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[73] S. Yang, G. Li, and Y. Yu. Cross-modal relationship inference for grounding referring expressions. *CVPR*, 2019.

[74] S. Yang, M. Xia, G. Li, H.-Y. Zhou, and Y. Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11266–11275, June 2021.

[75] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[76] L. Ye, M. Rochan, Z. Liu, and Y. Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019.

[77] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[78] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[79] X. Zang, A. Pokle, M. Vázquez, K. Chen, J. C. Niebles, A. Soto, and S. Savarese. Translating navigation instructions in natural language to a high-level plan for behavioral robot navigation. In *EMNLP*, 2018.

[80] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Muller, R. Manmatha, M. Li, and A. Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[81] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[82] W. Zhu, X. Wang, T.-J. Fu, A. Yan, P. Narayana, K. Sone, S. Basu, and W. Y. Wang. Multimodal text style transfer for outdoor vision-and-language navigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1207–1221, Online, Apr. 2021. Association for Computational Linguistics.

[83] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.