Comprehensive Multi-Modal Interactions for Referring Image Segmentation

Kanishk Jain and Vineet Gandhi
Center for Visual Information Technology, KCIS, IIIT Hyderabad
kanishk5991@gmail.com, vgandhi@iiit.ac.in

Abstract

We investigate Referring Image Segmentation (RIS), which outputs a segmentation map corresponding to the given natural language description. To solve RIS efficiently, we need to understand each word's relationship with other words, each region in the image to other regions, and cross-modal alignment between linguistic and visual domains. Recent methods model these three types of interactions sequentially. We argue that such a modular approach limits these methods' performance, and joint simultaneous reasoning can help resolve ambiguities. To this end, we propose a Joint Reasoning (JRM) module and a novel Cross-Modal Multi-Level Fusion (CMMLF) module for tackling this task. JRM effectively models the referent's multi-modal context by jointly reasoning over visual and linguistic modalities (performing word-word, image region-region, word-region interactions in a single module). CMMLF module further refines the segmentation masks by exchanging contextual information across visual hierarchy through linguistic features acting as a bridge. We present thorough ablation studies and validate our approach's performance on four benchmark datasets, and show that the proposed method outperforms the existing state-of-the-art methods on all four datasets by significant margins.

1. Introduction

Fundamental computer vision tasks related to localization, like detection and segmentation, aim to grant computers' visual abilities comparable to humans. Traditionally, these tasks have dealt with a pre-defined set of categories, making them difficult to scale and limit their practical use. Substituting the pre-defined categories with natural language expressions is a logical extension to counteract the above problems. Indeed, this is how humans interact with objects in their environment by referring them with linguistic queries. 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 natural language expression is known as Visual Grounding. Existing works approach the grounding problem either by predicting a bounding box around the referred object or by predicting a segmentation mask corresponding to the referred object. In this paper, 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).

Figure 1. Comparison of the proposed approach with CMPC [8]. In both the examples CMPC fails at the first stage itself, where it completely misses the actual referred entity. Our approach performs the exhaustive forms of interactions in a single step and identifies the correct referred entity. Best viewed in color and under zoom.

RIS task requires understanding both visual and linguistic modalities at an individual level, specifically word-word and region-region interactions. Additionally, a joint understanding of both modalities is required to identify the referred object from the linguistic expression and localizing it in the image. For instance, to ground a sentence "whatever is on the truck", it is necessary to understand the relation-

ship between words as grounding just the individual words will not work. Similarly, region to region interactions in visual modality help group semantically similar regions, ex: 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 state-of-the-art methods [8, 9, 7] take a modular approach to the RIS task, where these interactions happen in parts, sequentially.

Different methods differ in how they model these interactions. Huang *et al*. [8] first perform a region-word alignment (cross modal interaction). The second stage takes these region word alignments as input and selects the final relevant regions by reasoning over the entire linguistic expression. The reasoning step exploits the relationship and attributes corresponding to the referent in the textual expression. For example, for the sentence "The man holding a white Frisbee", the first stage will localize all instances of "man" and "Frisbee," and the second stage would select the correct instance of the "man" associated with a "white Frisbee". They use a Graph Convolutional Network for relational reasoning. Hui *et al*. [9] uses the dependency tree structure of the referring expression for the reasoning stage instead. Hu *et al*. [7] take a slightly different approach; instead of selecting the relevant region for each word, they select a relevant combination of words for each region. The second stage selects the relevant regions corresponding to referent based on the affinities with other regions. The problem with these approaches is that they model different forms of interactions in different stages. As a result, errors in the first stage of interaction limit the performance of subsequent ones (bottom row of Figure 1). The sequential interactions are also limited by design, as some RIS instances ideally require to model these interactions simultaneously (top row of Figure 1).

In this paper, we propose to perform all three forms of interactions simultaneously. We propose a Joint Reasoning Module (JRM) which jointly models inter-modal interactions and intra-modal interactions between the visual and linguistic modalities. Inter-modal interactions handle the cases for identifying the semantically similar words and regions in both modalities. Intra-modal interactions are used to transfer the contextual information between modalities to identify the referential context. Additionally, we propose a novel CMMLF module to exchange contextual information for referent across modalities and visual hierarchies and refine the referred object's segmentation mask.

We motivate the benefits of simultaneous interaction over the sequential interactions in Figure 1. We use the best performing CMPC [8] which first perceives all entities from the expression and individually aligns them in the visual domain. A reasoning step then follows. In both examples, CMPC fails at the first stage of entity perception. For the prediction in the top row (Figure 1), the sentence is to be understood as *a whole*, since *referred entity* is not explicitly mentioned in the expression. CMPC identifies "people" as the only present entity and ends up giving a wrong prediction. Similarly, in the second row (Figure 1), the expression is "store on left, next to hats with blanket draped in front". Both the scene and expression used are complex, as a lot of closely cluttered objects are there in the scene, and the language used to describe the referent uses complex relations between linguistic words. The first stage of CMPC predicts "hats" and "blankets" as entities and completely misses the actual referred object "store". In both cases, our approach with exhaustive interactions is able to understand the essence of the textual expression and reason about the referred object in the visual modality. Overall, our work makes the following contributions:-

1. We propose a Joint Reasoning Module (JRM) to jointly reason over regions; words, and region-word features. Joint Reasoning allows each modality to focus on semantic information common to both modalities to identify the referred object.
2. We propose Cross-Modal Multi-level Fusion (CMMLF) module, which allows contextual information to be exchanged across visual hierarchies through linguistic features, enabling common semantic information for referent to be aggregated from different visual hierarchies and result in a refined segmentation mask.
3. We present thorough quantitative and qualitative experiments to demonstrate the efficacy of our approach and show notable performance gains against current state-of-the-art methods on four RIS benchmarks.

2. Related Work

2.1. Semantic Segmentation

In semantic segmentation, the goal is to predict a label for each pixel in the image. Introduction of Fully Convolution Networks [14] 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. DeepLab and its variants [2, 3] introduce atrous Convolution to enlarge the receptive field of convolutional filters and aggregate multi-scale context using atrous spatial pyramid pooling. PSPNet [29] performs region-based context aggregation through pyramid pooling to extract multi-scale context. DANet [4] utilizes channel and position attention to adaptively integrate local features with their global dependencies. Recent works like ResNeSt [28] and HRNet-OCR [23] use attention-based approaches to combine information across feature map groups

and to combine multi-scale predictions, respectively. The task of RIS is a more generalized and natural variant of semantic segmentation where natural language referring expressions replace the predefined set of object categories.

2.2. Referring Expression Comprehension

Referring Expression Comprehension (REC) aims to localize the entities in the image referred to by the referring expression. In the REC task, the localization is performed using bounding box proposals. Existing approaches in REC can be categorized into two groups based on the model pipeline, (1) two-stage methods and (2) one-stage methods. In the two-stage methods, the first stage utilizes a pre-trained object detector to generate candidate bounding boxes for the given image, and the second stage selects the bounding box relevant to the object referred by the natural language expression. All the existing two-stage methods differ in their approaches for selecting the relevant bounding box proposal in the second stage. Earlier works like [6] used a scoring function on candidate boxes based on text query, and [20] use an attention mechanism for selecting the bounding box. Recent Works like [24] use cross-modal attention to model relations between language and vision modalities, followed by Graph Convolutional Network to perform relational reasoning to select the correct bounding box. In contrast to two-stage methods, one-stage methods combine the proposal generation network with the proposal selection network to create an end-to-end trainable network. [25] performs single-stage localization by augmenting the object detector with textual features. ZSGNet [21] combines the detector network and the grounding network and predicts classification scores and regression parameters for the candidate bounding boxes.

2.3. Referring Image Segmentation

Bounding Box based methods in REC are limited in their capabilities to capture the inherent shape of the referred object and are known to struggle with multi-scale objects. Referring Image Segmentation (RIS) task was proposed to alleviate the problems associated with REC tasks. RIS task was first introduced in [5], where they generate the referent's segmentation mask by directly concatenating visual features from CNN with tiled language features from LSTM. Later works like [13], perform sequential reasoning over individual words and visual regions through a convolutional multi-modal LSTM. [11] proposed Recurrent Refinement Networks (RRN) to generate refined segmentation masks by incorporate multi-scale semantic information from the image. Since each word in expression makes a different contribution to identify the desired object, [22] model visual context for each word separately using query attention. [26] uses a self-attention mechanism to capture long-range correlations between visual and textual modalities. Recent works [7, 8, 9] utilize cross-modal attention to model multi-modal context, [9] use dependency tree structure and [8] use coarse labelling for each word in the expression for selective context modelling. Most of the existing works capture only a subset of multi-modal interactions to model the context for referent. In this work, we concurrently and comprehensively model the intra-modal and inter-modal interactions across visual and linguistic modalities.

3. Method

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 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 Joint Reasoning Module (JRM) simultaneously aligns visual regions with textual words and jointly reasons about both modalities to identify the multi-modal context relevant to the referent. JRM is applied to hierarchical visual features extracted from CNN backbone since hierarchical features are better suited for segmentation tasks [26, 1, 7]. A novel Cross-Modal Multi-Level Fusion (CMMLF) is applied to effectively fuse JRM'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 JRM and CMMLF modules are described in the subsequent sections.

3.1. 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. All hierarchical visual features are transformed to the same spatial resolution and channel dimension through pooling and convolution operations. 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. We first initialize each word with a pre-trained word-embedding for the linguistic expression, which are then passed as input to the LSTM encoder. 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.

3.2. Joint Reasoning Module

In this section, we describe the Joint Reasoning Module (JRM). To successfully segment the referent, we need

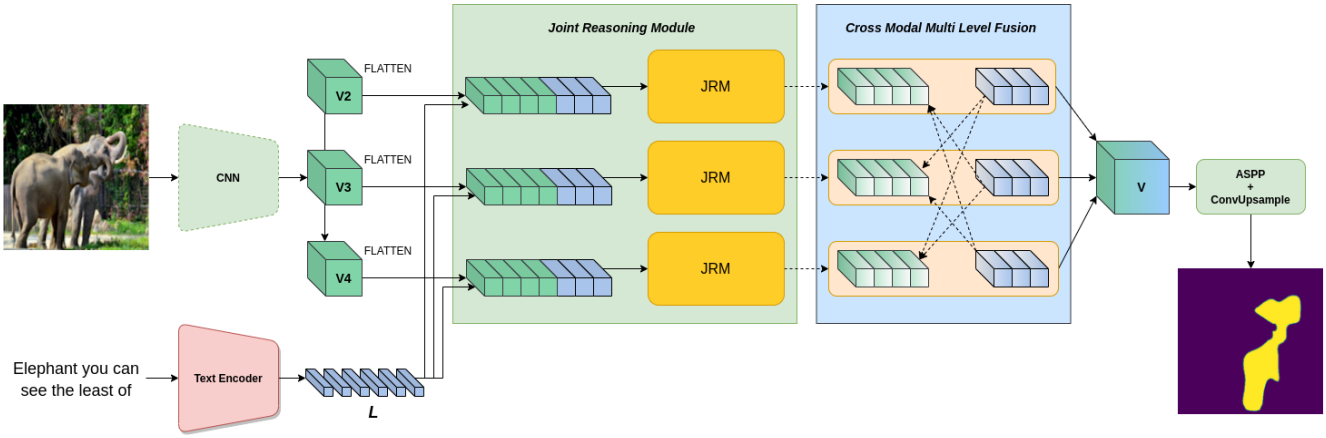


Figure 2. The proposed network architecture.

to identify the semantic information relevant to it in both the visual and linguistic modalities. This requires identifying region-region, word-word, and region-word pairs with similar contextual information. We model JRM as a multi-modal transformer encoder to capture the inter-modal and intra-modal interactions between visual and linguistic modalities. JRM is illustrated in Figure 3.

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 JRM, with $C_v = C_l = C$. We add separate positional embeddings to visual and linguistic features. For the visual features, we add spatially aware positional embedding S_V of shape $\mathbb{R}^{C \times H \times W}$, and for linguistic features, we add length aware positional embeddings S_l of shape $\mathbb{R}^{C \times T}$.

$$V^p = V + S_v \tag{1}$$

$$L^p = L + S_l \tag{2}$$

Here, V^p and L^p are the same shape as V and L, respectively. Following this, we flatten the spatial dimensions of visual features V^p and perform a length-wise concatenation with the linguistic features L^p to get a multi-modal feature M of shape $\mathbb{R}^{C \times (HW+T)}$. M is passed as input to the multi-modal transformer encoder. The self-attention mechanism in the encoder captures region-region and word-word interactions to identify similarly related regions and similarly related words. Further, region-word interactions help in reasoning about the referent by selecting regions and words with similar semantic context relevant to the referent. The output of JRM is a multi-modal feature X with cross-modal contextual information for the referent. X is the same shape as M. We compute X for all hierarchical visual features $\{V_2, V_3, V_4\}$, resulting in hierarchical cross-modal output $\{X_2, X_3, X_4\}$.

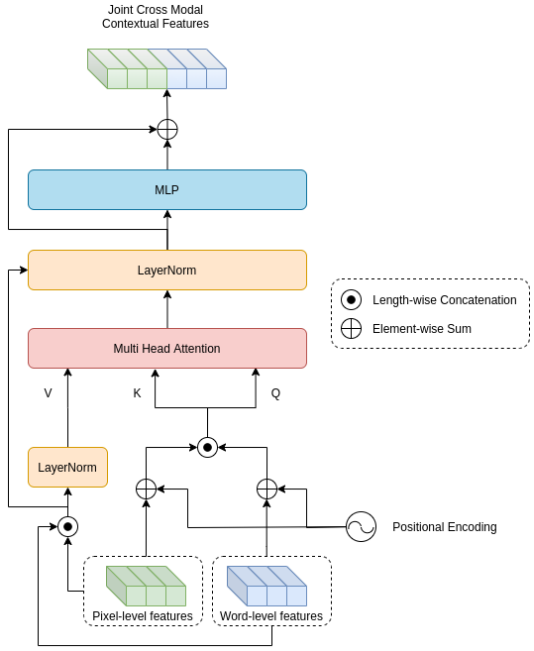


Figure 3. Joint Reasoning Module

3.3. Cross Modal Multi-level Fusion

Since features from different hierarchies in the CNN capture different aspects of the image, the input to JRM will differ in the visual information, as a result visual contextual information captured in X_i's will be different. In order to predict a refined segmentation mask for the referent, we need to aggregate the relevant contextual information from all hierarchies effectively. We propose a novel cross-modal multi-level fusion (CMMLF) module to address this.

The input to CMMLF module are the multi-modal features X_is from JRM. Since each X_i has shape of $\mathbb{R}^{C \times (HW+T)}$, they contain contextual information from

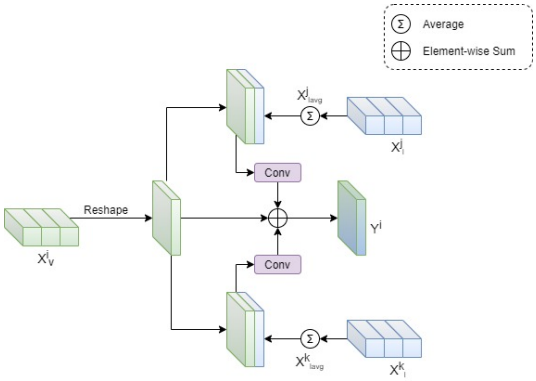


Figure 4. Cross Modal Multi Level Fusion Module

both modalities. First, we separate the visual and linguistic context from X_is to get visual features with linguistic context $X_i^v \in \mathbb{R}^{C\times HW}$, and linguistic features with visual context $X_i^l \in \mathbb{R}^{C\times T}$. X_i^l is averaged along the length dimension to result in a global visually attended linguistic feature $X_{l_{avg}}^i$.

Because of the hierarchical visual features, the visual context captured by each $X_{l_{avg}}^i$ is different. We utilize this aspect to use these linguistic features as a bridge to exchange visual information with other hierarchies. We take visual features from one hierarchy and concatenate it with linguistic part attended at the other hierarchies. More specifically, we take i^{th} layer output's visual part X_i^v and concatenate it with tiled textual part $X_{l_{avg}}^j$ of a different j^{th} layer along channel dimension. The concatenation is done separately with each of the remaining two layers. The full procedure is described in Figure 4.The visual contextual information is aggregated in the following way:-

$$\Lambda_{ij} = \sigma(Conv([X_v^i; X_{l_{avg}}^j])) \tag{3}$$

$$Y^i = X_v^i + \sum_{j\in\{2,3,4\}/\{i\}} \Lambda_{ij} \odot X_v^i \tag{4}$$

Here $\Lambda_{ij} \in \mathbb{R}^{C\times H\times W}$ are similarity weights between the i^{th} and j^{th} level hierarchies, and $Y^i \in \mathbb{R}^{C\times H\times W}$ is a refined multi-modal feature with visual context from other hierarchies. Finally, Y^i's are fused by stacking them along new dimension, resulting in $\mathbb{R}^{3\times C\times H\times W}$ dimensional vector, which is passed through 3-D Convolution to aggregate visual information from multiple levels to result in final refined multi-modal feature Y.

3.4. Mask Generation

Finally, Y is passed through Atrous Spatial Pyramid Pooling (ASPP) decoder 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 G to train the entire network end-to-end.

4. Experiments

4.1. Experimental Setup

We conduct experiments on four Referring Image Segmentation datasets: UNC [27], UNC+ [27], G-Ref [15] and Referit [10]. We describe each dataset separately.

UNC: The UNC dataset contains 19,994 images taken from MS-COCO [12] with 142,209 referring expressions corresponding to 50,000 objects. Referring Expressions for this dataset contain words indicating the location of the object. Two or more objects of the same object category appear in each image.

UNC+: THE UNC+ dataset is also based on images from MS-COCO. It contains 19,992 images, with 141,564 referring expressions corresponding to 50,000 objects. Unlike UNC, this dataset does not contain words that indicate the object's location, and the expression describes the object based on their appearance and context within the scene.

G-Ref: Like UNC and UNC+, G-Ref is also curated using images from MS-COCO. It contains 26,711 images, with 104,560 referring expressions for 50,000 objects. Each image contains 2 to 4 objects of the same category. G-Ref contains longer sentences with an average length of 8.4 words; compared to G-Ref, other datasets have an average sentence length of less than 4 words.

Referit: Referit dataset comprises of 19,894 images collected from IAPR TC-12 dataset. It includes 130,525 expressions for 96,654 objects. The expressions are shorter compared to other datasets. The foreground regions consist of objects and stuff (e.g., sky, mountains, and ground).

4.2. Implementation details

We adopt DeepLabv3+ [3] with Resnet-101 as a backbone for image feature extraction. Like previous works [26, 1, 7], our CNN backbone is pre-trained on Pascal VOC, and its parameters are fixed during training. For multi-level features, we extract features from layers 2, 3 and 4 of the CNN backbone. We conduct experiments with images at spatial resolutions of 448×448 and 576×576. At 448×448 resolution, $H = W = 14$ and at 576×576 resolution, $H = W = 18$. We use GLoVe embeddings [17] 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 Adam optimizer with weight decay (AdamW) with batch size set to 50; the initial learning rate is set to $2.5e^{-4}$ and weight decay of $5e^{-4}$ is used. The initial learning rate is gradually decreased using polynomial decay with a power of 0.5.

Evaluation Metrics: Following previous works [26, 1, 7], we evaluate the performance of our model using overall Intersection-over-Union (overall IoU) and Precision@X

Method	UNC			UNC+			G-Ref	Referit
	val	testA	testB	val	testA	testB	val	test
LSTM-CNN [5]	-	-	-	-	-	-	28.14	48.03
KWAN [22]	-	-	-	-	-	-	36.92	59.09
DMN [16]	49.78	54.83	45.13	38.88	44.22	32.29	36.76	52.81
ASGN [19]	50.46	51.20	49.27	38.41	39.79	35.97	41.36	60.31
RRN [11]	55.33	57.26	53.95	39.75	42.15	36.11	36.45	63.63
CMSA [26]	58.32	60.61	55.09	43.76	47.60	37.89	39.98	63.80
STEP [1]	60.04	63.46	57.97	48.19	52.33	40.41	46.40	64.13
BRIN [7]	61.35	63.37	59.57	48.57	52.87	42.13	48.04	63.46
LSCM [9]	61.47	64.99	59.55	49.34	53.12	43.50	48.05	66.57
CMPC [8]	61.36	64.53	59.64	49.56	53.44	43.23	49.05	65.53
JRNet* 448×448	64.31	68.13	60.48	52.00	56.44	43.96	48.72	68.03
JRNet* 576×576	**65.76**	**69.33**	**60.93**	**53.97**	**60.06**	**45.49**	**49.49**	**68.58**

Table 1. Comparison with State-Of-the-Arts on *Overall IoU* metric, $*$ indicates results without using DenseCRF post processing

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 $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

4.3. Comparison with State of the Art

We evaluate our method's performance on four benchmark datasets and present the results in Table 1. Like previous works, we use the Overall IoU metric to compare the performance against other state-of-the-art methods. At 576×576 input resolution, we outperform the existing methods by significant margins and achieve state-of-the-art numbers on all four datasets. Our method also achieves superior performance on three of the dataset at the 448×448 resolution. Most previous methods present results after post-processing the segmentation maps through a Dense Conditional Random Field (Dense CRF). In contrast, the presented results of our approach are without any such post-processing.

The expressions in UNC+ avoid using positional words while referring to objects; instead, they are more descriptive about the object's attributes and relationships. Substantial 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 1.46-1.88% over the next best performing method LSCM [9] 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 is able to resolve among multiple instances of the same type of objects and effectively locate the referred one.

The performance gains on the G-Ref dataset are marginal, achieving an improvement of 0.33% over CMPC. G-Ref is a relatively complex dataset with longer and verbose referring expressions (the average sentence contains more than 8 words). The results suggest scope for better modeling of the longer sentences. We experimented with contextual embeddings like ELMo [18] instead of the GLoVe; however, that did not improve the performance.

Our approach also achieves the highest gains on Precision@X metric on all datasets, specifically for $X = 0.9$. Our best performing model (with two encoder layers in JRM at resolution 576×576) gives 18.31% score in Precision@0.9 metric, compared to 12.89 of CMPC, achieving an improvement of 5.41%. More comprehensive evaluation results on Precision@X metric are presented in the supplementary material.

4.4. Ablation Studies

We perform ablation studies on the UNC dataset's validation split to validate the effectiveness of different modules in the proposed architecture. All methods are evaluated on Precision@X and Overall IoU metrics and the results are illustrated in Table 2 and Table 3. All ablations are performed at an input resolution of 448×448. 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.

The baseline model involves direct concatenation of visual features with the tiled textual feature to result in multi-modal feature of shape $\mathbb{R}^{(C_v+C_t)\times H\times W}$. This multi-modal feature is passed as input to ASPP + ConvUpsample decoder. The baseline model achieves a better Overall IOU score than some of the older methods like DMN [16] and ASGN [19].

CMMLF without JRM: "Only CMMLF" network differs with baseline method only on the fusion process of hierarchical multi-modal features. Introducing the CMMLF module over baseline results in 4.83 % improvement on the

Method	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	Overall IoU
Baseline	61.47	54.01	43.74	27.47	7.21	54.70
Only CMMLF	68.44	61.58	52.10	35.63	9.71	59.53
Only JRM	72.56	66.58	57.91	40.73	12.82	62.16
JRM+ConvLSTM	75.27	69.49	60.87	42.95	13.35	63.30
JRM+Conv3D	74.07	68.74	60.50	43.14	13.58	63.16
JRNet w/o Glove	74.23	68.42	59.77	42.47	13.66	62.19
JRNet w/o P.E.	74.18	68.36	59.71	43.15	13.36	63.07
JRNet	**76.52**	**71.66**	**63.43**	**45.70**	**15.69**	**64.31**

Table 2. Ablation Studies on Validation set of UNC, JRNet is the full architecture with both JRM and CMMLF modules. We use single encoder layer for all experiments using JRM. The input image resolution is 448×448 in each case.

Figure 5. Qualitative results comparing baseline against JRNet.

Split	JRM layers			
	n=1	n=2	n=3	n=4
val	64.25	**64.31**	63.59	63.36
testA	67.45	**68.13**	67.33	66.93
testB	60.07	**60.48**	59.73	59.35

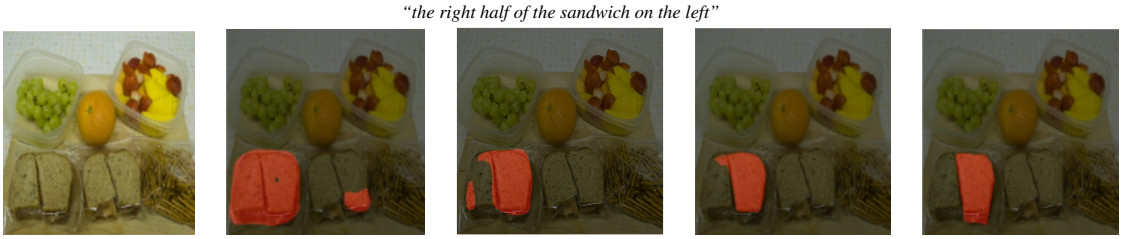
Table 3. Results on Overall IoU metric by varying the number of encoder layers in JRM on the UNC dataset.

Overall IoU metric and an improvement of 2.5 % on the $prec@0.9$ metric (illustrated in Table 2), indicating that the CMMLF module results in refined segmentation masks.

JRM without CMMLF: Similarly, the "Only JRM" network differs from the baseline method in the way different types of visual-linguistic interactions are captured. We observe significant performance gains of 7.46 % over the baseline, validating our claim that joint reasoning helps identify the referent.

JRM + X: We replace CMMLF module with other multi-level fusion techniques like ConvLSTM and Conv3D. Comparing the performance of JRM+ConvLSTM with JRNet (JRM+CMMLF), we observe that CMMLF is indeed effective at fusing hierarchical multi-modal features (Table 2). For JRM+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. JRM+Conv3D achieves a similar performance as JRM+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 CMMLF also uses Conv3D at the end, and the additional gains of JRNet over JRM+Conv3D suggest the benefits of hierarchical information exchange in CMMLF.

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), with the drop



(a) Original Image (b) Only CMMLF module (c) Only JRM module (d) JRNet (e) Ground Truth

Figure 6. We present qualitative results corresponding to combinations of proposed modules. In (b) we show results when only CMMLF module is used, (c) result with only JRM module being used, (d) output mask when both JRM and CMMLF modules are used

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

Figure 7. Output predictions for anchored image with varying linguistic expressions.

being more significant in the variant without Glove embeddings. These ablations suggest the importance of capturing word-level semantics and positional-aware features.

# of encoder layers in JRM: In Table 3, we present ablations on the UNC dataset by varying the number of encoder layers in JRM. We use the full model (JRNet) and find that a two-layer encoder gives the best performance on all UNC dataset splits. Increasing the number of layers deteriorates the performance gradually.

4.5. Qualitative Results

Figure 5 presents qualitative results comparing the baseline model against JRNet. JRNet is able to localize heavily occluded objects (Figure 5 (a) and (b)) and reason on the overall essence of the highly ambiguous sentences (e.g. "person you cannot see", "right photo not left photo") and ground them. It is able to distinguish among multiple instances of the same type of object based on attributes and appearance cues (Figure 5 (b), (c), and (e)). In contrast, the baseline model struggles to segment the correct instance and confuses it with other similar objects (e.g., fails to distinguish among different animals, bowls, and the two refrigerators). Figure 5 (d) and (f) illustrate the ability of JRNet to localize unstructured non-explicit objects like "dark area" and "blue thing". The potential of JRNet to perform relative positional reasoning is highlighted in Figure 5 (b), (e), and (f).

To further highlight the contribution of both JRM and CMMLF modules, we present qualitative results with networks trained using "Only CMMLF", "Only JRM" and JRNet in Figure 6. "Only CMMLF" network does not involve any reasoning; however, it manages to predict the left sandwich with refined boundaries. "Only JRM" network is able to understand 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 also seen. The full model benefits from the reasoning in "JRM," and when combined with CMMLF, it further facilitates information exchange across hierarchies and predicts a correct and refined mask as output.

In Figure 7, we anchor an image and make predictions by varying the natural language expression. Our approach is able to correctly segment all of the instances, clearly highlighting the flexibility and adaptability of the proposed JRNet model.

5. Conclusion

In this work, we tackled the task of Referring Image Segmentation. We proposed to solve this problem by comprehensively capturing interactions between different words in the linguistic expression, different regions of the image, and cross-modal interactions between words and image regions, in a single step. Furthermore, we introduced a novel fusion module, CMMLF, that fuses hierarchical multi-modal features by effectively exchanging and aggregating the contextual information relevant to the referent. We present thorough quantitative and qualitative experiments to demonstrate the efficacy of our method. The proposed JRNet achieves substantial gains over the state-of-the-art on all the four commonly used RIS benchmarks.

References

[1] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for re-

ferring image segmentation. In *ICCV*, 2019.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan 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.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[4] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[5] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016.

[6] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. *CVPR*, 2016.

[7] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, 2020.

[8] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, 2020.

[9] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, 2020.

[10] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[11] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018.

[12] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[13] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

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

[16] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *ECCV*, pages 630–645, 2018.

[17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[19] Shuang Qiu, Yao Zhao, Jianbo Jiao, Yunchao Wei, and Shikui Wei. Referring image segmentation by generative adversarial learning. *IEEE Transactions on Multimedia*, 22(5):1333–1344, 2019.

[20] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016.

[21] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *ICCV*, 2019.

[22] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, 2018.

[23] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[24] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. *CVPR*, 2019.

[25] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019.

[26] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019.

[27] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[28] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[29] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

Supplementary Material

Kanishk Jain and Vineet Gandhi
Center for Visual Information Technology, KCIS, IIIT Hyderabad
kanishk5991@gmail.com, vgandhi@iiit.ac.in

1. Detailed analysis on Precision@X

In this section, we present comprehensive results on Precision@X metric. We first compare against the existing approaches on Precision@X metric on UNC dataset and then present results of JRNet on Precision@X across all datasets.

In Table 1, we compare the performance of recent state-of-the-art methods on Precision@X metric. Other methods provide Precision@X metric results only on UNC dataset's validation split in their ablation studies and results on other datasets are not available. Hence the comparisons are limited to UNC's validation split. For other papers, we directly pick the results presented in their ablation section. As illustrated in Table 1, our approach achieves significantly higher Precision@X score for all values of X at both resolutions. At Precision@0.8 the performance improvement is 11.57% (30% relative improvement over the best performing CMPC). At Precision@0.9 our method achieves 6.16% improvement over CMPC (52% relative improvement, increasing to 19.05 from 12.89). The relative improvement over other methods increase with higher values of X in Precision@X, clearly illustrates the ability of our network to provide more refined segmentation maps, compared to the previous state-of-the-art methods.

In table 2, we present JRNet's performance on all four datasets on the Precision@X metric. High numbers at prec@0.5 metric indicate that our approach is able to localize the referent on a large number of cases (e.g. the correct referent is localized in more than 73% of cases across all splits of UNC dataset).

2. Comparison at different Resolution

We understand that our input image resolution is higher than existing methods. For a fair comparison, we train the current best performing method CMPC [8] at higher resolutions. In Table 3 and Table 4, we compare our approach against CMPC trained at different image resolutions of 448×448 and 576×576, respectively. Our method consistently outperforms CMPC on both resolutions across all metrics by significant margins. Interestingly, our network trained at 448×448 resolution beats CMPC trained at 576×576 resolution at almost all metrics by good margins. This indicates our approach's capability to effectively utilize additional visual semantic information at higher resolution. When comparing on Precision@X metric, we observe that the performance gap between CMPC and JRNet increases with increase in X, while comparing the models trained at the same resolution.

We would like to point out that, despite higher resolution of input images, the feature map resolution of visual features in our approach is very low compared to other methods. Our approach utilizes feature maps that are down-sampled by a factor of 32 from original image resolution, compared to other methods that down-sample the visual features only by factor of 8. We observed that using higher resolution feature map for the same image resolution results in increased training time with insignificant improvement in performance.

There is a small typo in Table1 of results section of our main paper. Lower values are reported for our method in the Overall IoU results for UNC's testB split. They change from 60.48% to 60.98% for JRNet at 448×448 resolution and from 60.93% to 61.93% for JRNet at 576×576 resolution.

3. Qualitative Results

In this section, we present additional qualitative results for JRNet model on variety of image-expression pairs.

In Figure S1, we present results where JRNet successfully grounded the referring expression in the image. JRNet 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), JRNet is 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, JRNet localizes objects

Method	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	Overall IoU
CMSA [26]	66.44	59.70	50.77	35.52	10.96	58.32
STEP [1]	70.15	63.37	53.15	36.53	10.45	60.04
BRINet [7]	71.83	65.05	55.64	39.36	11.21	61.35
LSCM [9]	70.84	63.82	53.67	38.69	12.06	61.47
CMPC [8]	71.27	64.44	55.03	39.28	12.89	61.36
JRNet (448 x 448)	76.52	71.66	63.43	45.70	15.69	64.31
JRNet (576 x 576)	**77.72**	**72.21**	**65.07**	**50.85**	**19.05**	**65.37**

Table 1. Comparison with other methods at precision@X metric.

Dataset	Split	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	Overall IoU
UNC	val	77.82	72.99	65.34	50.73	20.03	65.76
	testA	81.74	77.49	70.19	54.97	18.82	69.33
	testB	72.66	66.59	58.64	46.32	21.92	62.12
UNC+	val	64.68	59.90	53.21	40.31	14.38	53.97
	testA	71.72	67.32	60.06	46.99	15.01	60.06
	testB	53.65	48.76	41.68	31.15	12.80	45.49
G-ref	val	56.29	48.60	38.76	25.72	7.46	49.49
Referit	test	66.07	58.84	49.15	35.44	16.68	68.58

Table 2. Evaluation Results on Precision@X metric for JRNet at 576×576 resolution.

Method	Split	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	Overall IoU
CMPC	val	74.24	67.97	58.52	40.76	11.53	62.84
	testA	78.84	72.91	62.61	44.98	11.24	65.78
	testB	70.06	62.02	52.26	37.68	13.48	60.14
JRNet (ours)	val	76.51	71.66	63.42	45.69	15.69	64.31
	testA	81.19	76.45	68.14	49.35	14.44	68.13
	testB	71.42	65.96	57.17	43.33	17.48	60.98

Table 3. Comparison with CMPC at 448×448 resolution on UNC dataset

Method	Split	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	Overall IoU
CMPC	val	74.40	67.63	58.63	43.27	14.92	63.13
	testA	78.25	72.65	63.77	47.32	14.56	66.00
	testB	69.53	62.10	52.28	39.25	14.77	60.33
JRNet (ours)	val	77.72	72.21	65.07	50.85	19.05	65.37
	testA	81.72	77.14	69.82	55.33	18.58	68.88
	testB	73.11	66.42	58.50	45.98	21.15	61.93

Table 4. Comparison with CMPC at 576×576 resolution on UNC dataset

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 S1, the referred objects occupy extremely small region in the image space and JRNet is able to accurately locate them.

In Figure S2, 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, JRNet predicts a well segmented and refined output and the class predictions are also correct (an umbrella, a chair, a bottle, a keyboard etc.).

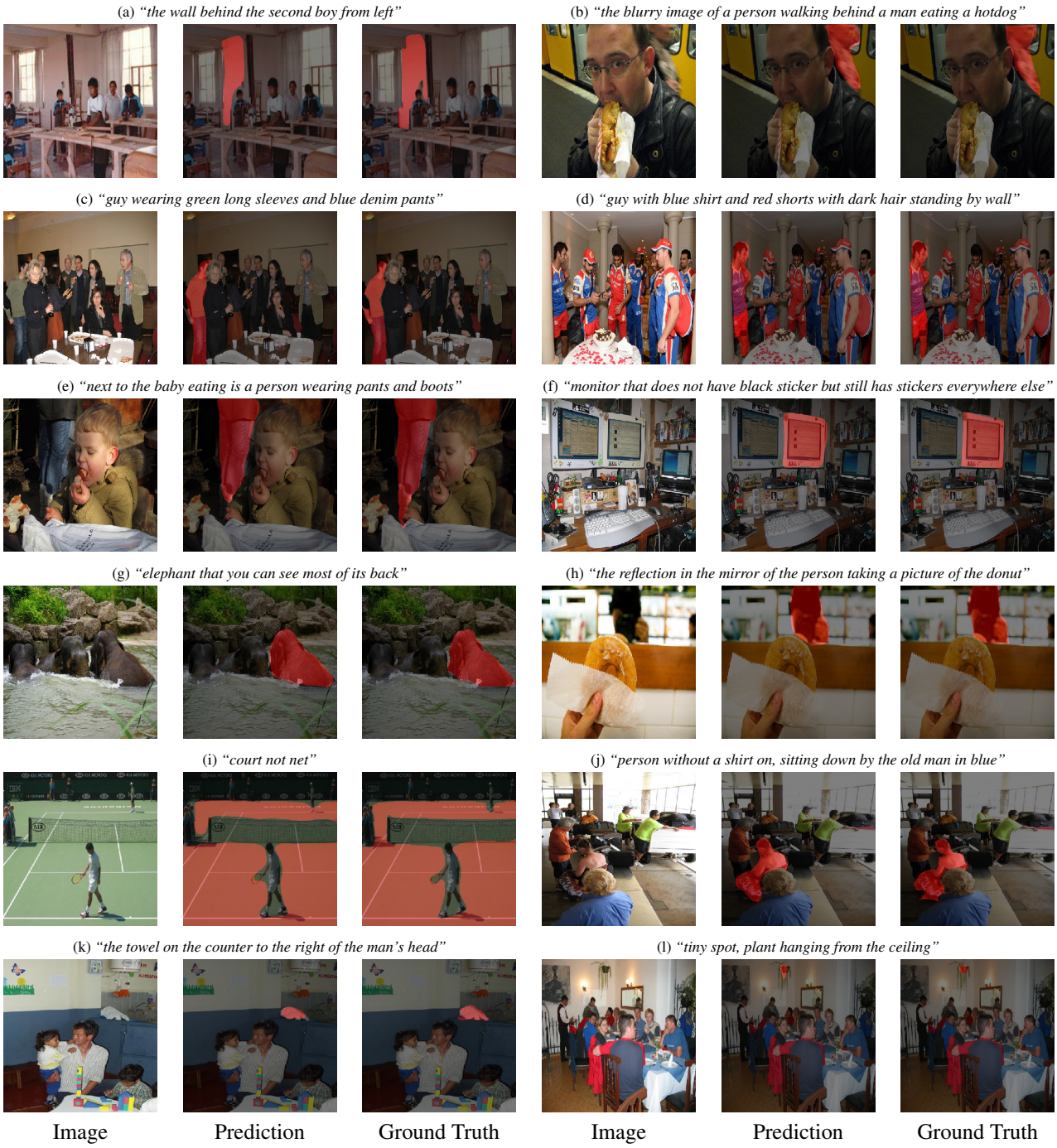


Figure S1. Qualitative examples where JRNet successfully localized the referred object.

Figure S2. Qualitative examples where JRNet failed to localize the referred object.