YOUREFIT: EMBODIED REFERENCE UNDERSTANDING WITH LANGUAGE AND GESTURE

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Tao Gao
Yixin Zhu, Song-Chun Zhu, Siyuan Huang
University of California, Los Angeles
{ethanchen,liqing,deqiankong,allen29,yixin.zhu,huangsiyuan}@ucla.edu
{tao.gao,sczhu}@stat.ucla.edu

ABSTRACT

We study the machine's understanding of **embodied reference**: One agent uses both language and gesture to refer to an object to another agent in a shared physical environment. Of note, this new visual task requires understanding **multimodal** information with **visual perspective-taking** to identify which object is being referred to. To tackle this problem, we introduce **YouRefIt**, a new crowd-sourced, large-scale real-world dataset of embodied reference; the dataset contains 4,195 unique reference clips in 432 indoor scenes. To the best of our knowledge, this is the first embodied reference dataset that affords us to study referring expressions in real-world scenes for understanding referential behavior, human communications, and human-robot interaction. We further devise two benchmarks for image-based and video-based embodied reference understanding. Our results provide overwhelming evidence that gestural information is as critical as language information in understanding the embodied reference, indicating the significance of incorporating gestures for visual scene understanding.

1 INTRODUCTION

Human communication (Tomasello, 2010) relies heavily on establishing common ground by referring to objects in a shared environment. This process usually takes place in two forms: language (abstract symbolic code) and gesture (unconventionalized and uncoded). In the computer vision community, efforts of understanding reference have been primarily devoted in the first form through an artificial task, Referring Expression Comprehension (REF) (Yu et al., 2016; Hu et al., 2017; Yu et al., 2018b; Liu et al., 2019b; Ye et al., 2019; Yang et al., 2019a; 2020a), but the second form, gesture, has been left almost untouched.

Fundamentally, all existing works deviate from the natural setting of reference understanding in daily scenes, which is **embodied**: An agent refers an object to another in a *shared* physical space, as exemplified by Fig. 1. Embodied reference possesses two distinctive characteristics compared to REF. First, it is **multimodal**. People often use both natural language and gestures when referring to an object. Second, recognizing embodied reference requires **visual perspective-taking** (Krauss & Fussell, 1991; Batson et al., 1997), the awareness that others see things from different viewpoints and the ability to imagine what others see from their perspectives. To address the deficiencies in prior works and study reference understanding at a full spectrum, we introduce a large-scale real-world and crowd-sourced dataset, ***YouRefIt***, for embodied reference understanding. For each reference clip, we annotate the reference target (object) with a bounding box. We also identify **canonical frames** in a clip: They are the "keyframes" of the clip and contain sufficient information of the scene, human gestures, and referenced objects that can truthfully represent the reference instance.

To measure the machine's ability in Embodied Reference Understanding (ERU), we devise two tasks based on the proposed *YouRefIt* dataset. (i) **Image ERU** takes a canonical frame and the transcribed sentence of the reference instance within as the inputs, and predicts the bounding box of the referenced object. (ii) **Video ERU** takes the video clip and the sentence as the input, and identifies the canonical frames and locates the reference target within the clip. Incorporating both gestural and language cues, we formulate a new multimodal framework to tackle the ERU tasks. In experiments, we provide multiple baselines and ablations. Our results reveal that models with

Figure 1: A daily deictic-interaction scenario that illustrates the significance of **multimodal** communication and **visual perspective-taking** in **embodied reference**.

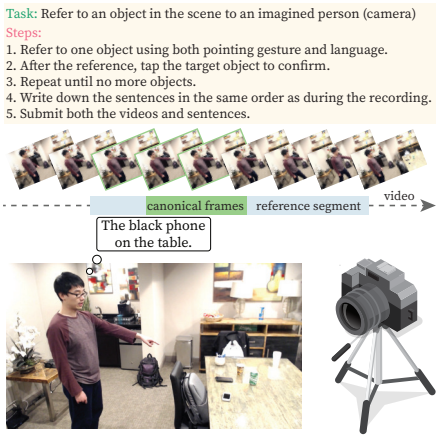


Figure 2: Dataset collection procedure. Participants were asked to film the a series of reference tasks following the instructions.

explicit gestural information yield better performance, validating our hypothesis that gesture is as critical as language information in resolving ambiguities in the embodied reference and facilitating successful communication with cooperation. We further verify that temporal information is essential in canonical frame detection, necessitating the understanding of embodied reference in dynamic and natural sequences.

2 THE *YouRefIt* DATASET

To study the embodied reference understanding, we introduce a new dataset *YouRefIt*, a large-scale video collection of people referring to objects with both language and gesture in indoor scenes.

2.1 DATA COLLECTION

Our dataset was collected via Amazon Mechanic Turk (AMT); see the illustration of the data collection process in Fig. 2. Workers were asked to record a video containing actions of referring to objects in the scene to an imagined person (camera) using both sentences and pointing gestures. Most videos were collected in indoor scenes, such as offices, kitchens, and living rooms.

2.2 DATA ANNOTATION

The annotation process takes two stages: (i) annotation of temporal segments, canonical frames, and referent bounding boxes, and (ii) annotation of sentence parsing.

Since each collected video consists of multiple reference actions, we first **segment** the video into clips; each contains an exact one reference action. A segment is defined from the start of gesture movement or utterance to the end of the reference, which typically includes the raise of hand and arm, pointing action, and reset process, synchronized with language description. In each segment, the annotators were asked to further annotate the **canonical frames**, which contain the "keyframes" that the referrer holds the steady pose to clearly indicate what is being referred. Combined with natural language, it is sufficient to use any canonical frame to localize the referred target. The participants were instructed to tap the referred objects after each reference action. Using this information, **bounding box** of the referred object were annotated using Vatic (Vondrick et al., 2013), and the tapping actions were discarded. The object color and material were also annotated if identifiable. The taxonomy of object color and material is adopted from Visual Genome dataset (Krishna et al., 2017). Given the sentence provided by the participants who performed reference actions, AMT annotators were asked to refine the sentence further and ensure it matches the raw audio collected from the video. We further provided more fine-grained **sentence parsing** results for natural language understanding. AMT annotators annotated target, target-attribute, spatial-relation, and comparative-relation.

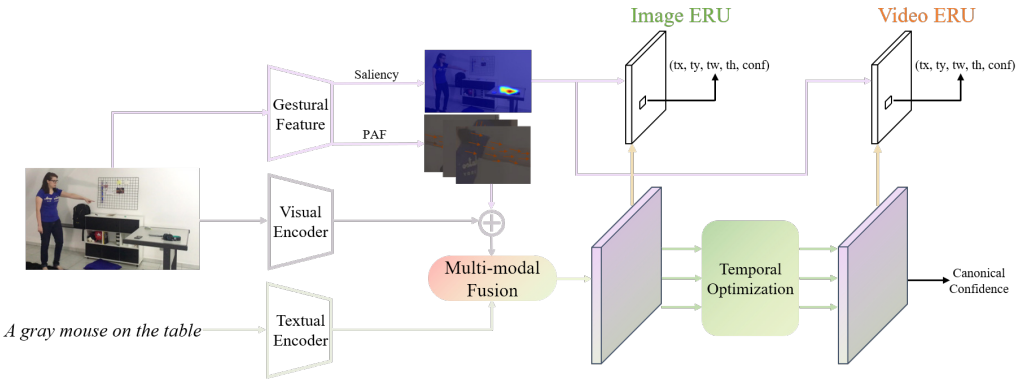


Figure 3: The proposed multimodal framework for the ERU task that incorporates both language and gestural information.

2.3 Dataset Statistics

In total, *YouRefIt* includes 432 recorded videos and 4,195 localized reference clips for 395 object categories. We retrieved 8.83 hours of video during the post-processing and annotated 497,348 frames. The total duration of all the reference actions is 3.35 hours, with an average duration of 2.81 seconds per reference. Each reference process was annotated with segments, canonical frames, bounding boxes of the referred objects, and the sentences with semantic parsing. All videos were collected with synchronized audio.

3 Embodied Reference Understanding (ERU)

3.1 Image ERU

Given the canonical frame and sentence from an embodied reference instance, Image ERU aims at locating the referred object in the image through the human gesture and language information. We use accuracy similar to Mao et al. (2016) as the evaluation metric. Following object detection benchmark (Geiger et al., 2012), we report the results under three Intersection over Union (IoU)s: 0.25, 0.5, and 0.75 with various object sizes, *i.e.*, *all small*, *medium* and *large*.

Methods We devise a novel multimodal framework for Image ERU that leverages both the language and gestural information; see Fig. 3. At a high-level, our framework includes both the visual and language encoder, similar to prior REF models (Yang et al., 2019b; 2020b; Luo et al., 2020). We also explicitly incorporate two types of gesture features: (i) the Part Affinity Field (PAF) (Cao et al., 2019) heatmap, and (ii) the pointing saliency heatmap following Kroner et al. (2020). We utilize the features from three modalities to effectively predict the target bounding box.

Results and Discussion Table 1 tabulates the quantitative results of the Image ERU. We categorize the models based on their information sources: *Language-only*, *Gesture-only*, and *Language + Gesture*. Below, we summarize some key findings.

First, gestural information is essential for embodied reference understanding. From Table 1, we can see that FAOA and ReSC models show significant performance improvement when trained on the original *YouRefIt* dataset compared with trained on the inpainted version, where humans are masked by He et al. (2017) and inpainted by DeepFill (Yu et al., 2018a).

Second, language information eases ambiguities that cannot be fully resolved by the gesture. As shown by the *Gesture-only* models, RPN+heatmap models suffer from the ambiguities of gestural information; pointing gestures are used to suppress descriptions of target location and focus attention on a spatial region but not object-centric. Performance of Ours$_{no_lang}$ also deteriorates compared to Ours$_{Full}$ if no referring expression is provided.

Third, explicit gestural features are beneficial for understanding embodied reference. Ours$_{PAF_only}$, which incorporates PAF features outperforms the origin FAOA and ReSC models. By further adding

the saliency heatmap, our full model Ours$_{Full}$ achieves the best performance in all baselines and ablation. Taken together, these results indicate that the fusion of the extracted information from language and gesture could be the crucial ingredient.

Table 1: Comparisons of Image ERU performances on *YouRefIt*.

Model	IoU=0.25				IoU=0.5				IoU=0.75			
	all	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*
Language-only												
MAttNet$_{pretrain}$	14.2	2.3	4.1	34.7	12.2	2.4	3.8	29.2	9.1	1.0	2.2	23.1
FAOA$_{pretrain}$	15.9	2.1	9.5	34.4	11.7	1.0	5.4	27.3	5.1	0.0	0.0	14.1
FAOA$_{inpaint}$	23.4	14.2	23.6	32.1	16.4	9.0	17.9	22.5	4.1	1.4	4.7	6.2
ReSC$_{pretrain}$	20.8	3.5	17.5	40.0	16.3	0.5	14.8	36.7	7.6	0.0	4.3	17.5
ReSC$_{inpaint}$	34.3	20.3	38.9	44.0	25.7	8.1	32.4	36.5	9.1	1.1	10.1	16.0
Gesture-only												
RPN+Pointing$_{15}$	15.3	10.5	16.9	18.3	10.2	7.2	12.4	11.0	6.5	3.8	9.1	6.6
RPN+Pointing$_{30}$	14.7	10.8	17.0	16.4	9.8	7.4	12.4	9.8	6.5	3.8	8.9	6.8
RPN+SaliencyKroner et al. (2020)	27.9	29.4	34.7	20.3	20.1	21.1	26.8	13.2	12.2	**10.3**	17.9	8.6
Ours$_{noLang}$	41.4	29.9	48.3	46.3	30.6	17.4	37.0	37.4	10.8	1.7	13.9	16.6
Language + Gesture												
FAOAYang et al. (2019b)	44.5	30.6	48.6	54.1	30.4	15.8	36.2	39.3	8.5	1.4	9.6	14.4
ReSCYang et al. (2020b)	50.4	35.4	59.0	56.8	37.3	17.0	48.4	46.8	12.6	1.7	16.4	18.7
Ours$_{PAF_only}$	53.3	39.3	**61.1**	61.9	40.1	21.8	**50.4**	52.1	13.4	1.9	19.1	21.0
Ours$_{Full}$	**55.1**	**42.7**	60.8	**62.5**	**42.1**	**23.9**	50.3	**54.0**	**14.4**	2.6	**19.3**	**23.4**
Human	94.2±0.2	93.7±0.0	92.3±1.3	96.3±1.7	85.8±1.4	81.0±2.2	86.7±1.9	89.4±1.7	53.3±4.9	33.9±7.1	55.9±6.4	68.1±3.0

3.2 VIDEO ERU

Compared with Image ERU, Video ERU is a more natural and practical setting in human-robot interaction. Given a referring expression and a video clip that captures the whole dynamics of a reference action with consecutive body movement, Video ERU aims at recognizing the canonical frames and estimate the referred target. For each reference instance, we sample image frames with 5 FPS from the original video clip. Average precision, recall, and F1-score are reported for the canonical frame detection. For referred bounding box prediction, we report the averaged accuracy in all canonical frames.

Table 2: Video ERU performance comparisons on *YouRefIt*.

Model	IoU=0.25				IoU=0.5				IoU=0.75			
	all	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*
Frame-based	**55.2**	42.3	**58.9**	**64.8**	**41.7**	**22.7**	53.4	**48.8**	16.9	1.6	21.8	**27.0**
Transformer	52.3	40.2	55.6	58.3	38.8	21.2	54.1	47.1	13.9	1.5	20.8	22.7
ConvLSTM	54.8	**43.1**	57.5	60.0	39.3	22.5	**54.8**	46.7	**17.3**	**1.8**	**24.3**	25.5
Ours$_{Full}$	55.1	42.7	60.8	62.5	42.1	23.9	50.3	54.0	14.4	2.6	19.3	23.4

Results and Discussion Table 2 shows quantitative results of predicting reference targets with the ground-truth canonical frames of the video. On the one hand, we observe that the frame-based method and the temporal optimization methods reach similar performance, comparable to the model that only trained on selected canonical frames (*i.e.*, Ours$_{Full}$). It shows the canonical frames can provide sufficient gestural and language information for clear reference, and the temporal models may be distracted from non-canonical frames. On the other hand, as shown in Table 3, temporal information can greatly improve performance on canonical frame detection since both the *ConvLSTM* and the *Transformer* model outperform the *Frame-based* method by a large margin. These results indicate the importance of distinguishing different stages of reference behaviors, *e.g.*, initiation, canonical moment and ending, for better efficacy in embodied reference understanding.

Table 3: Canonical frame detection performance.

Method	Avg. Prec	Avg. Rec	Avg. F1
Frame-based	31.9	37.7	34.5
Transformer	35.1	**44.2**	39.1
ConvLSTM	**57.0**	37.9	**45.4**

4 CONCLUSION AND FUTURE WORK

In this work, we study the reference understanding in an embodied manner, which we argue is a more natural way for understanding human communication with both language and gesture. To explore this problem, we crowd-source a large-scale, real-world video dataset *YouRefIt* and devise two benchmarks at both the image and video levels. We also propose a multimodal learning framework and conduct extensive experiments on *YouRefIt*. The experimental results provide strong empirical evidence that language and gesture coordination is critical for embodied reference understanding and human communication.

REFERENCES

C Daniel Batson, Shannon Early, and Giovanni Salvarani. Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and social psychology bulletin*, 23 (7):751–758, 1997.

Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Adam D Galinsky, William W Maddux, Debra Gilin, and Judith B White. Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological Science*, 19(4):378–384, 2008.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Yasuhiko Hato, Satoru Satake, Takayuki Kanda, Michita Imai, and Norihiro Hagita. Pointing to space: modeling of deictic interaction referring to regions. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

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

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Alfred Kobsa, Jurgen Allgayer, Carola Reddig, Norbert Reithinger, Dagmar Schmauks, Karin Harbusch, and Wolfgang Wahlster. Combining deictic gestures and natural language for referent identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 1986.

Robert M Krauss and Susan R Fussell. Perspective-taking in communication: Representations of others' knowledge in reference. *Social cognition*, 9(1):2–24, 1991.

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

Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020.

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

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b.

Andy Lücking, Thies Pfeiffer, and Hannes Rieser. Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79, 2015.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Karola Pitsch and Sebastian Wrede. When a robot orients visitors to an exhibit. referential practices and interactional dynamics in real world hri. In *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2014.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

Boris Schauerte and Gernot A Fink. Focusing computational visual attention in multi-modal human-robot interaction. In *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, 2010.

Boris Schauerte, Jan Richarz, and Gernot A Fink. Saliency-based identification and recognition of pointed-at objects. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2010.

Dadhichi Shukla, Ozgur Erkent, and Justus Piater. Probabilistic detection of pointing directions for human-robot interaction. In *International Conference on Digital Image Computing: Techniques and Applications*, 2015.

Dadhichi Shukla, Özgür Erkent, and Justus Piater. A multi-view hand gesture rgb-d dataset for human-robot interaction scenarios. In *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016.

Michael Tomasello. *Origins of human communication*. MIT press, 2010.

Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision (IJCV)*, 101(1):184–204, 2013.

Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019b.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. *arXiv preprint arXiv:2008.01059*, 2020b.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018a.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

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

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

A Related Work

Our work is related to two topics: (i) Referring Expression Comprehension (REF) studied in the context of Vision and Language, and (ii) reference recognition in the field of Human-Robot Interaction. Below, we compare our work with prior arts with a focus on these two topics.

A.1 Referring Expression Comprehension (REF)

REF is a visual grounding task. Given a natural language expression, it requires an algorithm to locate a particular object in a scene. Several datasets (Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016; Plummer et al., 2015; De Vries et al., 2017; Chen et al., 2020) have been constructed by asking annotators to provide expressions describing regions of images. Recently, Liu et al. Liu et al. (2019a) build a synthetic REF dataset by synthesizing both images and complex queries. To solve REF, researchers have attempted various approaches (Ye et al., 2019; Liu et al., 2019b; Yang et al., 2019a; 2020a). Representative methods include (i) localizing a region by reconstructing the sentence using an attention mechanism (Rohrbach et al., 2016), (ii) incorporating contextual information to grounding referring expressions (Zhang et al., 2018; Yu et al., 2016), (iii) using neural modular networks to better capture the structured semantics in sentences (Hu et al., 2017; Yu et al., 2018b), and (iv) devising a one-stage approach (Yang et al., 2019b; 2020b).

Our work fundamentally differs from REF at two levels.

Task-level REF primarily focuses on building correspondence between visual signals and natural language. In comparison, the proposed ERU task mimics the minimal human communication process in an embodied manner, which requires a mutual understanding of both verbal and nonverbal messages signaled by the sender. Recognizing reference in an embodied setting also introduces new challenges, such as visual perspective-taking (Galinsky et al., 2008): The referrers need to consider the perception from the counterpart's perspective for effective verbal and nonverbal communication, requiring a more holistic visual scene understanding both geometrically and semantically. In this paper, to study the reference understanding that echoes the above characteristics, we collect a new dataset containing natural reference scenarios with both language and gestures.

Model-level Since previous REF approaches only capable of comprehending communicative messages in natural language and completely ignore the gestural information, it is insufficient in the ERU setting or to apply on our newly collected dataset. To tackle this deficiency, we design a new principled framework to combine natural language and gestures by a multimodal fusion module. The proposed framework outperforms prior methods by a large margin, verifying the significant role of the gestural cue in addition to the language cue in embodied reference understanding.

A.2 Reference in Human-Robot Interaction

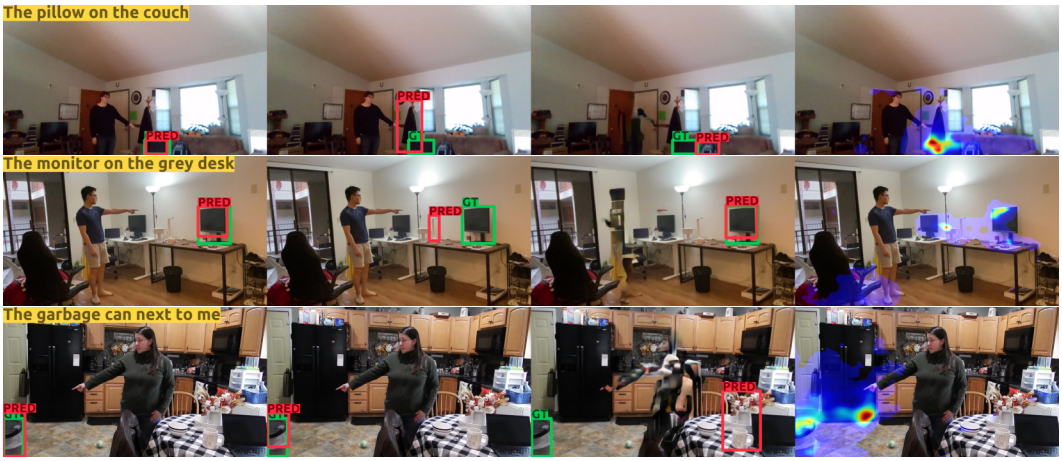
The combination of visual and verbal communication for reference is one of the central topics in Human-Robot Interaction. Compared with REF, this line of work focuses on more natural settings but with limited and specialized scenarios. Some works emphasize pointing direction and thus are not object-centric while missing language reference: The Innsbruck Pointing at Objects dataset (Shukla et al., 2015) investigates two types of pointing gestures with index finger and tool, and the Innsbruck Multi-View Hand Gesture Dataset (Shukla et al., 2016) records hand gestures in the context of human-robot interaction in close proximity. The most relevant works are ReferAt (Schauerte & Fink, 2010) and PointAt (Schauerte et al., 2010), where participants are tasked to point at various objects with or without linguistic expressions. Some other notable works include (i) a robotics system that allows users to combine natural language and pointing gestures to refer to objects on a display (Kobsa et al., 1986), (ii) experiments that investigate the semantics and pragmatics of co-verbal pointing through computer simulation (Lücking et al., 2015), (iii) deictic interaction with a robot when referring to a region using pointing and spatial deixis (Hato et al., 2010), and (iv) effects of various referential strategies, including talk-gesture-coordination and handshape, for robots interacting with humans when guiding attentions in museums (Pitsch & Wrede, 2014).

Although related, the above literature is constrained in lab settings with limited sizes, scenarios, and expressions, thus insufficient for solving the real-world reference understanding with both vision and language. In comparison, crowd-sourced by AMT, our dataset is much more diverse in environment setting, scene appearance, and language usage. Our dataset also collects videos instead of static images commonly used in prior datasets, opening new venues to study dynamic and evolutionary patterns that occurred during human communications.

B Method Details

We devise a novel multimodal framework for Image ERU that leverages both the language and gestural information; see Fig. 3. At a high-level, our framework includes both the visual and language encoder, similar to prior REF models (Yang et al., 2019b; 2020b; Luo et al., 2020), as well as explicitly extracted gesture features. We utilize the features from three modalities to effectively predict the target bounding box.

Specifically, we use Darknet-53 (Redmon & Farhadi, 2018) pre-trained on COCO object detection Lin et al. (2014) as the visual encoder. The textual encoder is the uncased base version of BERT (Devlin et al., 2018) followed by two fully connected layers. We incorporate two types of gesture features: (i) the PAF (Cao et al., 2019) heatmap, and (ii) the pointing saliency heatmap. Inspired by visual saliency prediction, we train MSI-Net (Kroner et al., 2020) on the *YouRefIt* dataset to predict the salient regions by considering both the latent scene structure and the gestural information, generating more accurate guidance compared with commonly used Region of Interests (RoIs). Fig. 4 shows some examples of predicted salient regions. We aggregate the visual feature and PAF heatmaps by max-pooling and concatenation, later further fused with textual features by a sub-query module (Yang et al., 2020b). The saliency map is directly used to refine the anchor box confidence score; we use the same classification and regression loss as described in Yang et al. (2019b) for anchor-based bounding box prediction.



(a) Ours$_{Full}$ (b) Ours$_{no_lang}$ (c) ReSC$_{inpaint}$ (d) Saliency Map

Figure 4: **Qualitative results in Image ERU of representative models with various information sources and pointing saliency map.** Green/red boxes are the predicted/grounth-truth reference targets. Sentences used during the references are shown at the top-left corner.

Figure 5: **Qualitative results in Video ERU of the ConvLSTM model.** Each row represents four selected frames from one reference clip. Green / red boxes indicate the predicted / grounth-truth reference targets. 0 means non-canonical frame and 1 means canonical frame.