Object Referring in Videos with Language and Human Gaze

Arun Balajee Vasudevan[1], Dengxin Dai[1], Luc Van Gool[1,2]
ETH Zurich[1] KU Leuven [2]
{arunv,dai,vangool}@vision.ee.ethz.ch

Abstract

We investigate the problem of object referring (OR) i.e. *to localize a target object in a visual scene coming with a language description. Humans perceive the world more as continued video snippets than as static images, and describe objects not only by their appearance, but also by their spatio-temporal context and motion features. Humans also gaze at the object when they issue a referring expression. Existing works for OR mostly focus on static images only, which fall short in providing many such cues. This paper addresses OR in videos with language and human gaze. To that end, we present a new video dataset for OR, with* $30,000$ *objects over* $5,000$ *stereo video sequences annotated for their descriptions and gaze. We further propose a novel network model for OR in videos, by integrating appearance, motion, gaze, and spatio-temporal context into one network. Experimental results show that our method effectively utilizes motion cues, human gaze, and spatio-temporal context. Our method outperforms previous OR methods. For dataset and code, please refer* https://people.ee.ethz.ch/~arunv/ORGaze.html.

1. Introduction

In their daily communication, humans refer to objects all the time. The speaker issues a referring expression and the co-observers identify the object referred to. In reality, co-observer also verifies by watching the gaze of the speaker. Upcoming AI machines, such as cognitive robots and autonomous cars, are expected to have the same capacities, in order to interact with their users in a human-like manner. This paper investigates the task of object referring (OR) in videos with language and human gaze.

OR has received increasing attention in the last years. Notable examples are interpreting referring expressions [57, 34], phrase localization [39, 55], and grounding of textual phrases [44]. Thanks to these excellent works, OR could be pushed to large-scale datasets [27, 34, 57] with sophisticated learning approaches [34, 23, 36, 39, 55]. However, previous OR methods are still limited to static images, whereas humans are well aware of the world's dynamic aspects. We describe objects not only by their appearance, but also their spatial-temporal contexts and motion features, such as *'the car in front of us turning left'*; *'the boy running fast under the tree there'*. Static images fall short in providing many of such cues. Thus, there is a strong need to push s-o-a OR to videos.

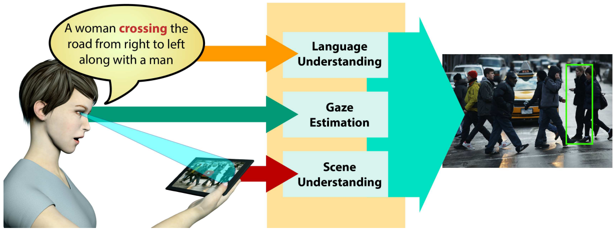

A woman crossing the road from right to left along with a man
Language Understanding
Gaze Estimation
Scene Understanding

Figure 1: A human issuing a referring expression while gazing at the object in the scene. The system combines multiple modalities such as appearance, motion and stereo depth from the video, along with the expression and the gaze to localize the object.

Another important cue that co-observers use to identify the objects is Gaze of the speaker. While describing the object, speakers gaze at the object to come out with an unique expression. Gaze is another important cue for object localization from the point of view of co-observer, along with the language expression. For example, suppose a car occupant instructs his/her autonomous car with expression *'Park under the yellow tree on the right'*, it is highly likely that he/she is gazing at that *tree*, or did so in the brief past. This gaze cue can be a promising aid to the car to localize the *tree*. In this work, we also investigate how gaze can be useful in assisting the OR task. As shown in Fig. 1, we use text language, gaze estimates, visual appearance, motion features and depth features to localize the object being referred.

As shown several times in computer vision, large-scale datasets can play a crucial role in advancing research, for instance by enabling the deployment of more complex learning approaches and by benchmarking progress. This paper presents a video dataset for OR, the first of its kind, with $30,000$ objects in $5,000$ stereo video sequences. The dataset is annotated with the guidance of Gricean Maxims [16] for cooperative conversations between people. That is, the descriptions need to be truthful, informative, relevant, and brief for co-observers to find the target objects easily and unambiguously. Later, human gazes are recorded as videos while they look at the annotated objects.

We further propose a novel Temporal-Spatial Context Recurrent ConvNet model, by integrating appearance, motion, gaze, and spatial-temporal context into one network. See Fig. 2 for a diagram of our model. The model learns the interactions between language expressions and object characteristics in the 'real' 3D world, providing human users the freedom to interact by speaking and gazing. Experimental results show that our method effectively uses motion cues, temporal-spatial context information, and human gazes.

Our main contributions are: 1) presenting a new video dataset for object referring, featuring bounding-boxes, language descriptions and human gazes; 2) developing a novel OR approach to detect objects in videos by learning from appearance, motion, gaze, and temporal-spatial context.

2. Related Work

Our work is relevant to the joint understanding of language and visual data. It is especially relevant to referring expression generation and language-based object detection.

The connection between language and visual data has been extensively studied in the last three years. The main topics include image captioning [25, 56, 13], visual question answering (VQA) [42, 9, 8] and referring expressions [20, 23]. Although the goals are different, these tasks share many fundamental techniques. Two of the workhorses are Multimodal Embedding [14, 25, 15] and Conditional LSTM [24, 44, 57, 34]. Multimodal Embedding projects textual data and visual data both to a common space, in which similarity scores or ranking functions are learned. Multimodal Embedding was initially explored for the task of image captioning [14, 12, 25] and later reinforced in VQA [42, 33, 15]. It is common practice to represent visual data with CNNs pre-trained for image recognition and to represent textual data with word embeddings pre-trained on large text corpora [38]. A Conditional LSTM is a generative model conditioned on visual input, and it is usually trained to maximize the generative probability of language descriptions [23, 34] or answers to questions [9, 42]. Our model conditions LSTMs not only on images but also on motion, depth and gaze.

Language-based Object Referring. Language-based object referring (OR) has been tackled under different names. Notable ones are referring expressions [57, 34], phrase localization [39, 55], grounding of textual phrases [44, 40, 10], language-based object retrieval [23] and segmentation [22]. Recent research foci of language based OR can be put into 2 groups: 1) learning embedding functions [15, 25, 54] for effective interaction between vision and language; 2) modeling contextual information to better understand a speaker's intent, be it global context [34, 23], or local among 'similar' objects [36, 57, 34]. Our work extends [23] from static images to stereo videos to exploit richer, more realistic temporal-spatial contextual information along with gaze cues for the task of OR.

Object Referring Datasets. This section discusses relevant OR datasets: Google Refexp [34], UNC Refexp [57], ReferIt [27]. The Google Refexp dataset, which was collected by Mao *et al.* [34], contains $104,560$ referring expressions annotated for $54,822$ objects from $26,711$ images from the MSCOCO dataset [32]. UNC Refexp was collected in the same spirit as GoogleRef, but applying the ReferIt game [27] on MSCOCO. While these datasets are large-scale and of high quality, they contain only static images. This excludes useful information about the visual scenes such as motion cues and 3D spatial configuration, and also limits the descriptions to mere appearances and 2D spatial information. We build on the success of these datasets and present a new object referring dataset for stereo videos. Annotators were encouraged to use descriptions about 3D configuration and motion cues when necessary.

Gaze Estimation. Gaze or eye tracking has been used in computer vision tasks like object detection [26, 58] and tracking [46], image captioning [50], image/video annotation [46, 37, 49] and others. We focus on object referring. There are some works [1, 3] which support that speakers gaze reliably precedes his reference to an object and this speaker's referential gaze helps in listeners' comprehension. Common sense also tells us that we have to gaze at the object before we refer them. Listeners use this gaze information for a better object localization [3, 7, 2]. Misu *et al.* [35] uses gaze and speech recorded inside a car to locate real world landmark points. Vadivel *et al.* [46] uses eye tracking over videos to extract salient objects as tracklets. The task of [37] matches closely with us to detect objects using gaze. Nonetheless, they use gaze to fasten the annotation for object detection. [41] proposes to follow the gaze of people in the scene to localize the place where they look. Krafka *et al.* [29] build an mobile application to enable large scale eye tracking annotation by crowdsourcing. Inspired from [29], we create a web interface to record gaze via Amazon Mechanical Turk (AMT) for object localization in videos.

3. Approach

Object referring (OR) is widely used in our daily communication. Here, we follow the literature [23, 34] to formulate the problem as an object detection task. Given a video sequence of visual scene $\mathbf{I} = (I^1, I^2, ..., I^t)$ and a video sequence of speaker's gaze $\mathbf{G} = (G^1, G^2, ..., G^t)$,

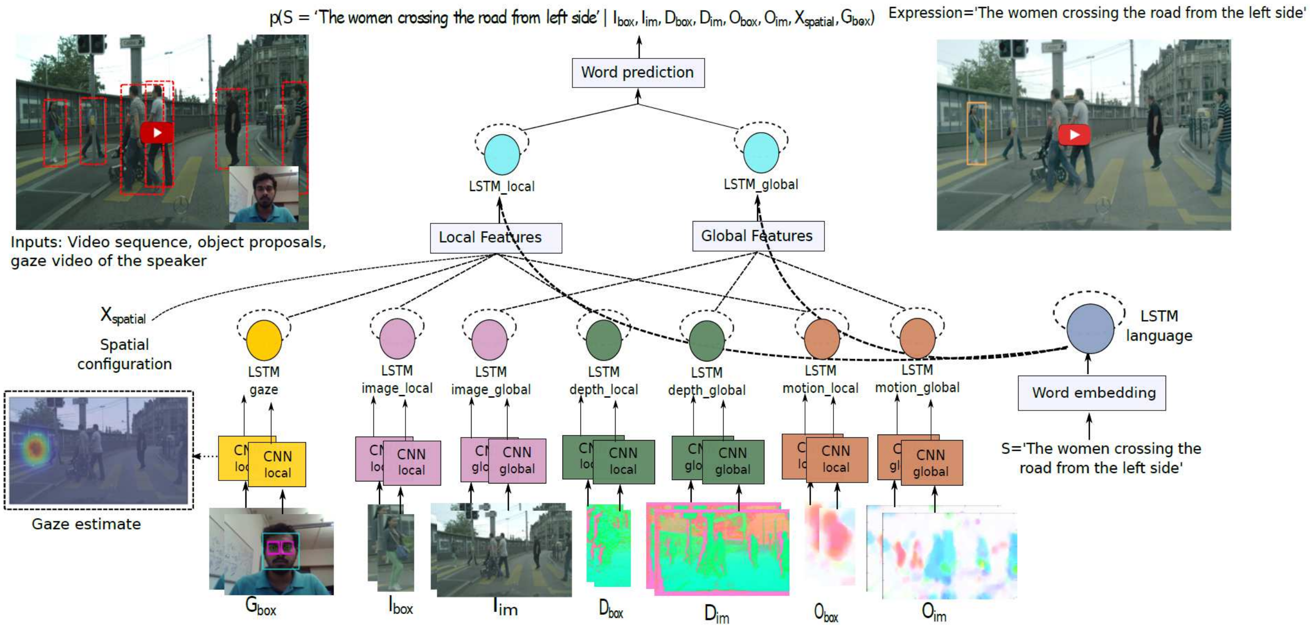


Figure 2: The illustrative diagram of our model for object referring in stereo videos with language expression and human gaze. Given a referring expression S, our model scores all the M bounding box candidates by jointly considering local appearance (I^t_{box}), local motion (O^t_{box}), local depth (D^t_{box}), local human gaze (G^t_{box}), spatial configuration X_{spatial}, and the global temporal-spatial contextual information (I^t, D^t and O^t).

where t is the *current* frame at which the referring expression S is issued, our goal is to identify the referred object $\hat{b}^t$ out of all object proposals $\{b^t_m\}^M_{m=1}$ at frame t. M is the total number of object proposals considered. Note that we assume that t is known a priori to simplify the task. In real application, the exact t needs to be inferred from speaker's speech and the visual scene. The performance of our method is also evaluated at frame t.

3.1. Network Architecture

Following [23] and the work on image captioning [24], we choose to maximize the generative probability of the expression for the target object. Our model is based on the Spatial Context Recurrent ConvNet model developed in [23] for OR in static images. The model in [23] unifies three LSTMs [21] to integrate information from language expressions, global visual context and local object content. It has gained success in OR for static images. This work extends it so that information from stereo videos and human gazes can be incorporated, resulting in our model architecture as shown in Fig. 2.

Let us denote the seven **visual** LSTM models by $\text{LSTM}_{\text{gaze}}$, $\text{LSTM}_{\text{image\_local}}$, $\text{LSTM}_{\text{image\_global}}$, $\text{LSTM}_{\text{depth\_local}}$, $\text{LSTM}_{\text{depth\_global}}$, $\text{LSTM}_{\text{motion\_local}}$ and $\text{LSTM}_{\text{motion\_global}}$, and their hidden states by $\mathbf{h}^{\text{gaze}}$, $\mathbf{h}^{\text{image}}_{\text{local}}$, $\mathbf{h}^{\text{image}}_{\text{global}}$, ..., $\mathbf{h}^{\text{motion}}_{\text{global}}$, respectively and denote the **language** LSTM model by $\text{LSTM}_{\text{language}}$ with hidden state $\mathbf{h}^{\text{language}}$. We concatenate the local and global features separately as shown in Fig. 2. Successively, we have **visual-language** LSTM models namely $\text{LSTM}_{\text{local}}$ and $\text{LSTM}_{\text{global}}$ which take concatenated local and global features respectively along with $\mathbf{h}^{\text{language}}$ as inputs. Let us denote their hidden states as $\mathbf{h}_{\text{local}}$ and $\mathbf{h}_{\text{global}}$ respectively. A word prediction layer is used on top of these two visual-language LSTMs to predict the words in the expression S. Practically, our model is trained to predict the conditional probability of the next word w_{n+1} in S, given the local content of the objects: G^t_{box}, I^t_{box}, D^t_{box} and O^t_{box}, the corresponding spatio-temporal contexts: I^t, D^t and O^t as detailed in Sec. 3.3, and all the n previous words. The problem can be formulated as:

$$\begin{aligned}&p(w_{n+1}|w_n, ..., w_1, I^t, I^t_{\text{box}}, D^t, D^t_{\text{box}}, O^t, O^t_{\text{box}}, G^t_{\text{box}})\\&= \text{SoftMax}(W_{\text{local}}\mathbf{h}_{\text{local}}(n) + W_{\text{global}}\mathbf{h}_{\text{global}}(n) + \mathbf{r}) \qquad (1)\end{aligned}$$

where W_{local} and W_{global} are the weight matrices for word prediction from $\text{LSTM}_{\text{local}}$ and $\text{LSTM}_{\text{global}}$, and $\mathbf{r}$ is a bias vector.

At training time, the method maximizes the probability of generating all the annotated expressions over the whole dataset. Following [23], all the seven LSTM models have 1000 hidden states. At test time, given a video sequence $\mathbf{I}$, a gaze sequence $\mathbf{G}$ and M candidate bounding boxes $\{b^t_m\}^M_{m=1}$ at frame t considered by the method proposed in Sec. 3.2, our model computes the OR score for b_m by computing the generative probability of S on b_m^t(box):

$$\begin{aligned} s_i &= p(S|I^t, I^t_{\text{box}}, D^t, D^t_{\text{box}}, O^t, O^t_{\text{box}}, G^t_{\text{box}}) \\ &= \prod_{w_n \in S} p(w_n|w_{n-1}, ..., w_1, I^t, I^t_{\text{box}}, D^t, D^t_{\text{box}}, O^t, O^t_{\text{box}}, G^t_{\text{box}}). \end{aligned} \tag{2}$$

The candidate with the highest score is taken as the predicted target object. Below, we describe our object proposal and feature encoding.

3.2. Object Proposals

In the spirit of object detection, we adopt the strategy of proposing candidates efficiently and then verifying the candidates with a more complex model for the OR task. This strategy has been used widely in the literature. For instance, [23] uses EdgeBox [59] for the object proposals; [34] and [25] use the faster RCNN (FRCNN) object detector [43], Mask-RCNN [18] and Language based Object Proposals (LOP) [52] and others to propose the candidates. [52] shows that LOP performs significantly better than other techniques when we propose expression-aware object candidates. For the same reason, we use LOP [52] for the object proposals.

3.3. Feature Encoding

In order to better use the spatio-temporal information provided by a stereo video, we augment I^t with the corresponding depth map D^t and optical flow map O^t. In addition to these global contexts, for a bounding box b^t, its local features are used as well: I^t_{box} for its appearance, D^t_{box} for its depth characteristics, O^t_{box} for its motion cues and G^t_{box} for the gaze. CNNs are used to encode the local and global information from the three information sources. I^t_{box}, D^t_{box} and O^t_{box} can be computed on frame t alone or together with multiple previous frames for long-range temporal interaction. The same is applicable for I^t, D^t and O^t also. The detailed evaluation can be found in Sec. 5.

Appearance. We use the *fc7* feature of VGG-16 net [48] and ResNet [19] pre-trained on ILSVRC-2015 [45] to represent I^t_{box} and I^t, which are passed through $\text{LSTM}_{\text{image\_local}}$ and $\text{LSTM}_{\text{image\_global}}$ respectively to yield features $\mathbf{f}^{image}_{local}$ and $\mathbf{f}^{image}_{global}$, respectively.

Depth. For depth, we convert depth maps to HHA images [17] and extract the CNN features with the RGB-D network of [17] before passing to $\text{LSTM}_{\text{depth\_local}}$ and $\text{LSTM}_{\text{depth\_global}}$. This leads to depth features $\mathbf{f}^{depth}_{local}$ and $\mathbf{f}^{depth}_{global}$ for D^t_{box} and D^t, respectively.

Optical Flow. Similarly, we employ the pre-trained two-stream network [47] trained for video action recognition to extract convolutional flow features. Again, the *fc7* features are used leading to 4096-dimensional features which are given to $\text{LSTM}_{\text{motion\_local}}$ and $\text{LSTM}_{\text{motion\_global}}$ to get the motion features $\mathbf{f}^{motion}_{local}$ and $\mathbf{f}^{motion}_{global}$ for O^t_{box} and O^t.

Language. The words in the expression S are represented as one-hot vectors and embedded by word2vec [38] first and later, the expression S is embedded by an LSTM model [21] $\text{LSTM}_{\text{language}}$, leading to a language feature vector $\mathbf{h}^{\text{language}}$ for S.

Human Gaze. We synchronize the video of human gaze and the Cityscapes's video, which was displayed on a laptop for gaze recording. On the extracted frames, we perform face detection to crop out the face image and then conduct facial landmark point detection using the work of [28]. Successively, we detect the left eye and the right eye, and extract them as well. An example is shown in Fig. 2. Then, we use GazeCapture model [29] which takes input of left and right eye images along with the face image and outputs the gaze estimates relative to the camera location on the device. We convert the estimate from camera coordinate system to image coordinate system by applying a linear mapping function. The mapping function is device-specific and defined by the relative position of the camera and the screen of the laptop. More details of the mapping function can be found in the supplementary material.

Finally, we plot a 2D Gaussian map around the estimated gaze coordinates on the image to accommodate the errors in gaze estimation by the GazeCapture model [29] as shown in Fig. 4. Later, we compute gaze feature for an object in a frame by region pooling (averaging) over the bounding box region inside the Gaussian map. We concatenate these features over all the frames of the video to yield a resultant gaze feature $\mathbf{f}^{gaze}$ for an object *i.e.* G^t_{box}.

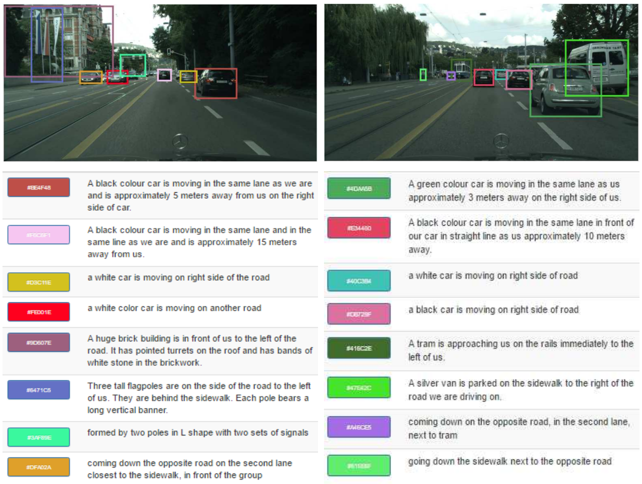
Feature Concatenation. Similar to [23], we concatenate the language feature with each of the two concatenated local and global features to obtain two meta-features: $\{[\mathbf{h}^{\text{language}}, \mathbf{f}^{image}_{local}, \mathbf{f}^{depth}_{local}, \mathbf{f}^{motion}_{local}, \mathbf{f}^{gaze}, \mathbf{f}_{spatial}]$, $[\mathbf{h}^{\text{language}}, \mathbf{f}^{image}_{global}, \mathbf{f}^{motion}_{global}, \mathbf{f}^{motion}_{global}]\}$. The meta-features are then given as the input to two LSTM models $\text{LSTM}_{\text{local}}$ and $\text{LSTM}_{\text{global}}$ respectively, to learn the interaction between language and all the 'visual' domains, *i.e.* appearance, motion, depth, gaze and their global contexts. $\mathbf{f}_{spatial}$ denotes the spatial configuration of the bounding box with respect to the 2D video frame. Following [34, 23], this $8-$dimensional feature is used:

$$\mathbf{f}_{\text{spatial}} = [x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}}, x_{\text{center}}, y_{\text{center}}, w_{\text{box}}, h_{\text{box}}] \tag{3}$$

where w_{box} and h_{box} are the width and height of the box. See Fig. 2 for all the interactions. Our method can be trained in a full end-to-end fashion if enough data is provided.

4. Dataset Annotation

As discussed in Sec. 1 and Sec. 2, previous datasets do not cater to the learning and evaluation of temporal, spatial context, and gaze information. Thus, we collected a new dataset. A video OR dataset should contain diverse



#BE4F48	A black colour car is moving in the same lane as we are and is approximately 5 meters away from us on the right side of car.
#F6C6F1	A black colour car is moving in the same lane and in the same line as we are and is approximately 15 meters away from us.
#D3C11E	a white car is moving on right side of the road
#FE001E	a white color car is moving on another road
#9D607E	A huge brick building is in front of us to the left of the road. It has pointed turrets on the roof and has bands of white stone in the brickwork.
#6471C5	Three tall flagpoles are on the side of the road to the left of us. They are behind the sidewalk. Each pole bears a long vertical banner.
#3AF89E	formed by two poles in L shape with two sets of signals
#DFA02A	coming down the opposite road on the second lane closest to the sidewalk, in front of the group

#4DA65B	A green colour car is moving in the same lane as us approximately 3 meters away on the right side of us.
#E34460	A black colour car is moving in the same lane in front of our car in straight line as us approximately 10 meters away.
#40C3B4	a white car is moving on right side of road
#DB729F	a black car is moving on right side of road
#416C2E	A tram is approaching us on the rails immediately to the left of us.
#47E42C	A silver van is parked on the sidewalk to the right of the road we are driving on.
#A46CE5	coming down on the opposite road, in the second lane, next to tram
#61E88F	going down the sidewalk next to the opposite road

Figure 3: Top: sample images from the Cityscape dataset with objects marked in differently colored bounding boxes. Bottom: corresponding referring expression annotations.

visual scenes and their objects should be annotated at the frame (time) when the expression is issued. We acknowledge that all modalities should be recorded/annotated at the same time, ideally in the real human-to-robot communication scenarios. That, however, renders data collection very labor-intensive and infeasible to crowd source.

In this work, we choose to use the existing stereo videos from Cityscapes dataset [11], and annotate language expressions, object bounding boxes, and gaze recordings via crowd sourcing. Cityscapes consists of $5,000$ high-quality video sequences in total, captured with a car mounted stereo camera system in 50 different European cities. The videos contain diverse sets of traffic scenes such as *car approaching a signal stop*, *pedestrians crossing the road*, *trams running through the street*, and *bicycles are overtaken*, and *kids crossing road lanes*, *etc*. See Fig. 3 for some examples.

Crowdsourcing. We crowdsourced the annotation task of OR in videos via AMT. Each Human Intelligence Task (HIT) contains one video. The videos in the Cityscapes dataset are all 2 seconds long, comprising 30 frames. An AMT worker was asked to annotate bounding boxes for objects on the last frame of the video (*i.e*. the 30th frame). The 30th frame is chosen mainly to make sure that annotated objects come with sufficient temporal context. In the annotation, workers are 'forced' to watch the video at least once in order to annotate an object. Replaying the video is highly encouraged if something is unclear.

Quality Control. To generate high quality annotations, we ran the first round of HIT as a qualification task. We qualified 20 workers based on their annotation of bounding boxes and natural language descriptions who further annotated the entire dataset. Following the work of Li *et al*. [31], we employed various quality control mechanisms for the syntactic validation to ensure the high quality of sentences. Some of the used validation checks are: number of words in the description must be at least 5, words must contain only ASCII characters, copy/paste operation is not allowed in the field where workers typed the descriptions and finally, we check for grammatical and spelling errors using the HTML5 spellcheck attribute. We payed 0.075 US dollar for each annotation of one bounding box, the name of the object class, and a referring expression. In total, we have collected $30,000$ annotated objects in $5,000$ videos.

Dataset Statistics. The average length of referring expressions of the objects is 15.59 words compared to 8.43 in Google Refexp and 3.61 in the UNC Refexp dataset, which are popular referring expression datasets. There are 20 classes of objects in Cityscape. The average number of referring expressions on objects annotated per image is 4.2 compared to 3.91 in Google Refexp. The distribution of annotations is 53.99% referring expressions for 'car', 22.97% for 'person', 4.9% for 'vegetation', 3.9% for 'bicycle', 3.46% for 'building', 2.95% for 'rider' and the rest for the remaining categories.

Gaze Recording. As a separate annotation task, we record human gaze for the objects which have been annotated already with referring expressions and bounding boxes. This is inspired from Krafke *et al*. [29] where eye tracking dataset is created via crowdsourcing by asking workers to gaze at particular points on the device screen with their face being recorded. Here, we collect the gaze recording on objects annotated in Cityscapes dataset as mentioned beforehand in this section. We create a web interface where we show the videos and asked the turk workers to gaze at the object one after the other. We record the faces of workers using the frontal camera of their laptops while they gaze at the shown objects.

In our annotation interface, we instruct the workers to adjust the window size such that the canvas where videos are displayed, occupies a major amount of screen space for a higher resolution of gaze estimation. With the start of the annotation, workers are asked to watch the complete video at first to put them into context. Once this is done, we show the objects (in bounding boxes) with annotated bounding boxes on the canvas. The workers are asked to click inside the box which activates the running of the video. We direct the workers to gaze at the same clicked object during the stage of video streaming while we keep recording the gaze of the worker throughout this period. Successively, we show the next objects and record the gazes correspondingly. At the end of the annotation of each video, we collect videos for the gaze recording of every annotated object.

We ensured the quality of recording by allowing only qualified workers to participate in this task. We perform the qualification same as in the earlier task except the criteria being checked is *gazing* here. Workers perform the task under different lighting conditions and at times, their visibility of their face goes down with its consequence being that the face detection fails in those cases. Hence, we re-recorded the gazes for all the videos where face detection on the workers failed. Finally, we recorded gaze for all the annotated objects of the Cityscapes.

5. Experiments

Given a video sequence of visual scene, gaze recording sequence of the speaker and a referring expression, our task is to yield bounding box location of the object. Our model scores and ranks the object proposals (which we generate using LOP [52]) based on the textual description, the spatial and temporal contextual information from stereo videos and gaze recording of the speaker. We first evaluate the performance of multiple modalities namely, RGB image, depth information and object motion. It is aimed to show the usefulness of depth and motion provided by stereo videos for the task of Object Referring (OR). Later, we show how gaze aids our model to improve the OR accuracy further.

5.1. Implementation Details

Our model is designed to incorporate gaze of the speaker, temporal and depth cues of the objects as well as the contextual information. The main advantage of Cityscapes referring expression annotations over other referring expressions datasets like GoogleRef, UNC Refexp and ReferIt is that the Cityscapes consists of short video snippets and the corresponding depth maps, suited for our task. For RGB images, we extract features from VGG16 [48] and ResNet [19] as mentioned in Sec. 3. For depth, we generate HHA images from disparity maps following the work of Gupta *et al.* [17]. Furthermore, we extract HHA features using the RCNN network used by [17]. For motion, we compute optical flow for all the frames using Fast Optical Flow by Kroeger *et al.* [30]. We extract optical flow features using the flow network of the two stream Convolutional network implemented by Simonyan *et al.* [47] for action recognition in videos. To compute object level features in all frames of videos, we compute tracks of the objects using the annotated bounding box on the last frame of the videos(30th frame in Cityscapes). We compute tracks for each object using Correlation filter based tracking [51].

As to Gaze, we sample frames from the gaze video at a frame rate greater than that of the Cityscapes video to ensure one-to-one correspondence between the sequences. Then, we extract the face from each frame with [53]. For these face images, we use Deep Alignment Network [28] to extract facial landmark points. Using the left and right eye landmark points, we extract the left and right eye image which we later give to GazeCapture model [29] along with face information. GazeCapture model outputs the gaze prediction in camera coordinates. We convert these camera coordinates to image coordinates using the linear mapping function as mentioned in Sec. 3. Then, we plot 2D Gaussian plot around the gaze estimate in image coordinates with $\sigma = 100$ pixels which is 10% of image dimension. This helps in accommodating the prediction error of gaze coordinates from the GazeCapture model as mentioned in [29]. See Fig. 4 for gaze prediction error distribution. Successively, we perform region pooling over the bounding box location of Gaussian map to obtain the object feature. We concatenate the features from each frame to get the gaze feature for the entire recording. Finally, we train our model by using the extracted features from RGB, HHA, Optical Flow images and gaze features as shown in Fig. 2.

	Methods	Edgebox	FRCNN	LOP
VGG	SimModel [39]	4.5	18.431	35.556
	MNLM [4]	-	23.954	32.418
	VSEM [6]	-	24.833	32.961
	MCB [15]	-	26.445	33.366
	NLOR [23](Ours(I))	4.1	27.150	36.895
ResNet	NLOR-ResNet (Ours(I))		29.333	38.645
	Ours (I,D)	-	38.833	41.388
	Ours (I,O)	-	39.166	42.500
	Ours (I,D,O)	-	41.205	43.750
	Ours (I,D,O,G)	-	**47.256**	**47.012**

Table 1: Numbers denote Acc@1. The # of candidate proposals M is 30. All evaluations are on Cityscapes. Abbreviations: I:RGB, D:Depth map, O:Optical Flow, G:Gaze. Since Edgebox performs poorly in baseline:Ours(I), we avoid further experiments.

	Track length(in frames)		
Methods	1	2	8
Ours (I)	38.645	-	-
Ours (I,O)	42.500	42.418	42.320
Ours (I,D,O)	**43.750**	**42.875**	**42.875**

Table 2: Comparison of methods when longer term motion is considered. Numbers denote Acc@1. Track length represents the number of past frames used for flow information. The different methods are evaluated on Cityscape.

5.2. Evaluation

Out of the $30,000$ annotated objects in our Cityscape dataset [11], we use 80% of videos for the training and 20% for the evaluation of our model on the task of OR in videos.

Evaluation Metric We evaluate the performance of our language based OR based on Accuracy@1 ($Acc@K$), follow-

Methods	Ours (I)	Ours (I,O)	Ours (I,D,O)
w/o Gaze	38.645	42.500	43.750
w/ Gaze +AvgPool	41.242	41.895	43.791
w/ Gaze	41.535	43.888	45.816
w/ Gaze +MaxPool	**42.418**	**44.248**	**47.012**

Table 3: Comparison of approaches w/ and w/o Gaze. Numbers denote Acc@1. # of candidate proposals M is 30. The different methods are evaluated on Cityscape. 1st row has overlap with Tab. 1

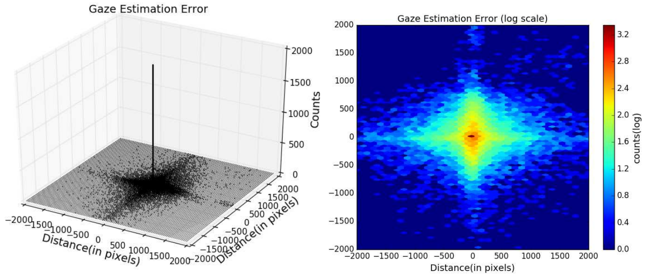


Figure 4: Gaze Estimation error distribution. We compute the distance between the gaze estimation coordinates with the groundtruth bounding box along X and Y axis(Centre denotes zero error). Left side figure represents the error in real valued scale and right side in log scale. We choose 2000 pixel distance to match with Cityscapes image dimensions.

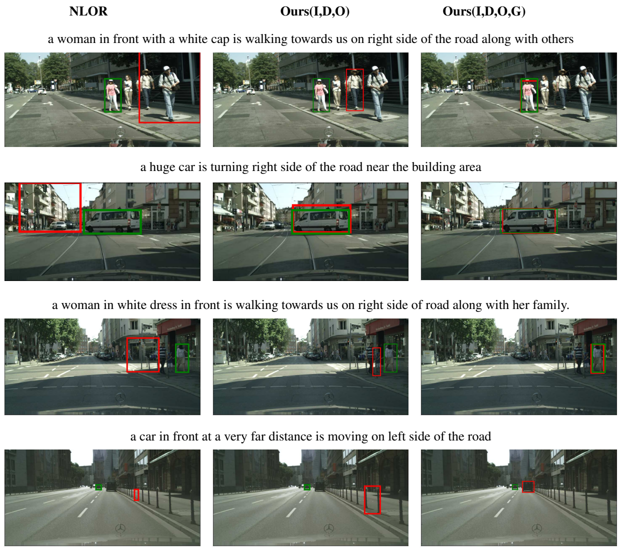

NLOR
Ours(I,D,O)
Ours(I,D,O,G)
a woman in front with a white cap is walking towards us on right side of the road along with others
a huge car is turning right side of the road near the building area
a woman in white dress in front is walking towards us on right side of road along with her family.
a car in front at a very far distance is moving on left side of the road

Figure 5: Some qualitative results from: NLOR (left column), Ours(I,D,O) (middle) and Ours(I,D,O,G) (right column). These results are obtained on the Cityscapes. Green: ground truth box and Red: predicted box.

ing [23, 34]. $Acc@1$ refers to the percentage of top scoring candidates being a true detection. A candidate is regarded as a true detection if the Intersection over Union (IoU) computed between the predicted bounding box and ground truth box is more than 0.5. In all our tables, we compute mean of the $Acc@1$ metric over all the videos in the evaluation set.

Object Proposal We use object proposal methods, e.g. Edgebox [59], FRCNN [43] and LOP [52] and compare them in Tab. 1. Since LOP performs consistently better for OR, we use the same for all later experiments.

Images vs. Stereo Videos. The performance of our model using different modalities and their combinations is reported in Tab. 1. We compare with better performing discriminative approaches (compared to CCA [5] as in [4, 6]) like MNLM [4], VSEM [6], MCB [15] as in Tab. 1. Surprisingly, the above approaches perform worse than NLOR [23], which uses a simple generative model. This could be because discriminative methods are more dataset-dependent and harder to generalize to new datasets/tasks. For instance, they expect 'carefully-engineered' *negatives* selection and sometimes more structured language expressions [5]. We choose [23] as baseline due to its simplicity and it only requiring *positives*. The table shows that our model can effectively utilize additional modalities such as the depth and motion provided in stereo videos. For instance, the $Acc@1$ is improved by 5.105% by using depth and motion; that is the improvement of Ours (I,D,O) over NLOR under LOP for 30 object proposals (I:RGB, D:Depth map, O:Optical Flow, G:Gaze). We observe a similar improvement for FRCNN as object proposals. Since LOP is consistent over all methods, we choose LOP as proposal technique for Tab. 2 and Tab. 3. We observe that both depth and motion can boost OR performance, on top of RGB appearances alone. For example, depth improves the performance by 2.743% and motion improves it by 3.855%. The combination of all four (appearance, depth, motion and gaze) yields the best results improving by 8.367%, indicating its usefulness for the OR task. Please see Fig. 5 for some visual results and how the three additional modalities (D,O,G) improve the detections over the original work [23].

Also according to the table, motion features are more useful than depth in our case. This can be ascribed to the fact that the convolutional network to extract motion was trained with a larger dataset than for the depth network. A better representation of depth than HHA images is a good topic for further research. In Tab. 2, we included flow information from the past by experimenting with different track lengths. We see that longer tracks do not bring significant improvement to OR accuracy. This may be because a) the referring expressions are annotated for the last frame (referring expressions are likely to be valid only for a short time as both camera and objects may be moving. *e.g. the women **entering** the door*) and b) the length of Cityscapes videos is

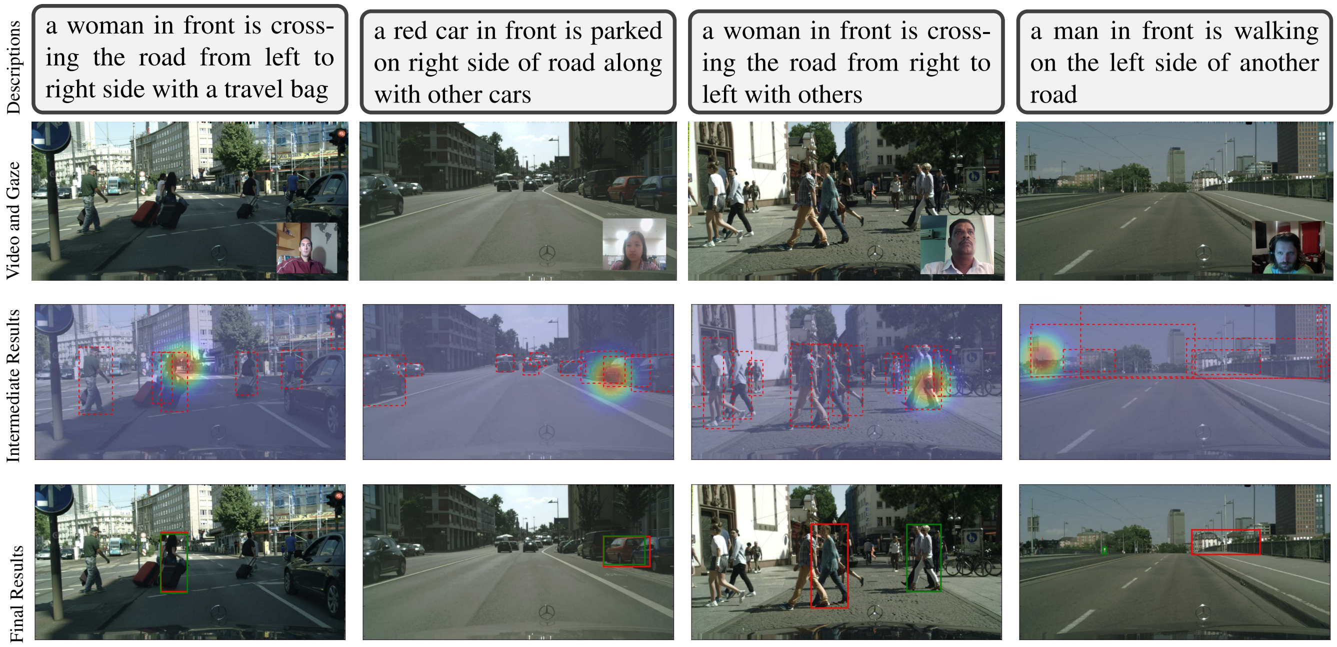

Descriptions
Video and Gaze
Intermediate Results
Final Results
a woman in front is crossing the road from left to right side with a travel bag
a red car in front is parked on right side of road along with other cars
a woman in front is crossing the road from right to left with others
a man in front is walking on the left side of another road

Figure 6: Overall results on Cityscapes. Input descriptions on the first row and input videoframe and gaze in the second row. Middle row represents intermediate results where Gaze estimation is embedded along with object proposals while bottom row represents the final OR results. Green: ground truth box and Red: proposals and the predicted boxes.

just 2 sec., too short to contain significant motion changes.

Gaze vs. w/o Gaze. Gaze features are given to our model as additional local features along with depth and flow features. Comparing our model with and without Gaze, Gaze improves its performance significantly for all its variants (Tab. 3). For instance, Gaze features under Max pooling improve the performance by 3.773% for the image-only case (Ours (I)); by 1.748% when image and motion are used (Ours (I,O)); and by 3.262% when image, motion and depth are used (Ours (I,D,O)). Human gaze consistently improves OR performance, because humans do gaze at objects when issuing referring expressions.

Given sampling rate differences between a Cityscapes video and its gaze video, we experimented with gaze feature extraction in 2 cases: a) timestamp matching between the videos, b) # gaze frames > # frames in the Cityscapes video, where we tried average and max pooling of object features to ensure one-to-one correspondence between frames. Tab. 3 shows that max pooling of object features performs as good or better than other cases such as averaging pooling because errors due to quickly changing gaze (outliers) can be avoided by max-pooling the object features.

Qualitative Results. We provide qualitative results in Fig. 6. The top row represents the inputs, incl. a Cityscapes video, a gaze recording video and the referring expression. Having overlaid the gaze estimate over object proposals (middle row), we can also observe the proximity of the gaze estimate to the referred objects. We show the predicted and groundtruth boxes in the bottom row. We add some failure cases in Fig. 6. From the above experiments, we infer that using multi-modal inputs - depth, motion, and gaze, along with RGB image features - improves OR performance.

6. Conclusions

In this work, we have proposed a solution for object referring (OR) in videos using language and speaker's gaze. The main contributions are: 1) a new video OR dataset with 30, 000 objects annotated across 5, 000 different video sequences; 2) a novel approach Temporal-Spatial Context Recurrent ConvNet for OR in videos, which integrates appearance, motion, depth, human gaze and spatio-temporal context that can be trained in an end-to-end fashion; and 3) gaze recordings for all annotated objects and demonstration of their effectiveness for OR. Experiments show that our model can effectively utilize motion cues, gaze cues and spatio-temporal context provided by stereo videos, outperforming image-based OR methods consistently. Training and evaluating our method, especially the contribution of the multiple modalities, in a real human-to-robot communication system are future works.

Acknowledgement: The work is supported by Toyota via project TRACE-Zurich. We also thank Vaishakh Patil and Prashanth Chandran for helpful comments.

References

[1] Peripheral speaker gaze facilitates spoken language comprehension: syntactic structuring and thematic role assignment in german. In *ECCS*, 2011. 2

[2] Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? *Frontiers in Psychology*, 2012. 2

[3] Influence of speaker gaze on listener comprehension: Contrasting visual vs intentional accounts. *Cognition*, 2014. 2

[4] Unifying visual-semantic embeddings with multimodal neural language models. *arXiv*, 2014. 6, 7

[5] Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *ICCV*, 2015. 7

[6] Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, 2015. 6, 7

[7] Referential gaze makes a difference in spoken language comprehension:human speaker vs. virtual agent listener gaze. 2017. 2

[8] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016. 2

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[10] K. Chen, R. Kovvuri, and R. Nevatia. Query-guided regression network with context policy for phrase grounding. 2

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6

[12] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[13] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015. 2

[14] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 2

[15] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2, 6, 7

[16] H. Grice. Syntax and semantics: Speech acts. In P. Cole and J. Morgan, editors, *Syntax and Semantics: Speech Actsn*, volume 3, page 4158. Academic Press, 1975. 2

[17] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, 2014. 4, 6

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017. 4

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[20] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. *ICCV*, 2017. 2

[21] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3, 4

[22] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 2

[23] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 1, 2, 3, 4, 6, 7

[24] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 2, 3

[25] 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. 2, 4

[26] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Ecksteinz, and B. Manjunath. From where and how to what we see. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 625–632, 2013. 2

[27] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. 1, 2

[28] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. *arXiv preprint arXiv:1706.01789*, 2017. 4, 6

[29] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2176–2184, 2016. 2, 4, 5, 6

[30] T. Kroeger, R. Timofte, D. Dai, and L. Van Gool. Fast optical flow using dense inverse search. In *European Conference on Computer Vision*, pages 471–488. Springer, 2016. 6

[31] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 5

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2

[33] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9, 2015. 2

[34] 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. 1, 2, 4, 7

[35] T. Misu, A. Raux, I. Lane, J. Devassy, and R. Gupta. Situated multi-modal dialog system in vehicles. In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*, pages 25–28. ACM, 2013. 2

[36] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 1, 2

[37] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *European Conference on Computer Vision*, pages 361–376. Springer, 2014. 2

[38] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 2, 4

[39] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik. Phrase localization and visual relationship detection with comprehensive linguistic cues. *CoRR*, abs/1611.06641, 2016. 1, 2, 6

[40] W. H. M. D. R Yeh, J Xiong and A. Schwing. Interpretable and globally optimal prediction for textual grounding using image concepts. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2

[41] A. Recasens*, A. Khosla*, C. Vondrick, and A. Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. * indicates equal contribution. 2

[42] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2953–2961, 2015. 2

[43] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 4, 7

[44] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 1, 2

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4

[46] K. Shanmuga Vadivel, T. Ngo, M. Eckstein, and B. Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[47] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, 2014. 4, 6

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4, 6

[49] M. Soliman, H. R. Tavakoli, and J. Laaksonen. Towards gaze-based video annotation. In *Image Processing Theory Tools and Applications (IPTA), 2016 6th International Conference on*, pages 1–5. IEEE, 2016. 2

[50] Y. Sugano and A. Bulling. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*, 2016. 2

[51] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. Torr. End-to-end representation learning for correlation filter based tracking. *arXiv preprint arXiv:1704.06036*, 2017. 6

[52] A. B. Vasudevan, D. Dai, and L. Van Gool. Object referring in visual scene with spoken language. *arXiv preprint arXiv:1711.03800*, 2017. 4, 6, 7

[53] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001. 6

[54] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2

[55] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng. Structured matching for phrase localization. In *European Conference on Computer Vision*, pages 696–711. Springer, 2016. 1, 2

[56] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015. 2

[57] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 1, 2

[58] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746, 2013. 2

[59] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 4, 7