Generating Easy-to-Understand Referring Expressions for Target Identifications

Mikihiro Tanaka[1], Takayuki Itamochi[2], Kenichi Narioka[2], Ikuro Sato[3],
Yoshitaka Ushiku[1] and Tatsuya Harada[1,4]
[1]The University of Tokyo, [2]DENSO CORPORATION, [3]DENSO IT Laboratory, Inc., [4]RIKEN

Abstract

This paper addresses the generation of referring expressions that not only refer to objects correctly but also let humans find them quickly. As a target becomes relatively less salient, identifying referred objects itself becomes more difficult. However, the existing studies regarded all sentences that refer to objects correctly as equally good, ignoring whether they are easily understood by humans. If the target is not salient, humans utilize relationships with the salient contexts around it to help listeners to comprehend it better. To derive this information from human annotations, our model is designed to extract information from the target and from the environment. Moreover, we regard that sentences that are easily understood are those that are comprehended correctly and quickly by humans. We optimized this by using the time required to locate the referred objects by humans and their accuracies. To evaluate our system, we created a new referring expression dataset whose images were acquired from Grand Theft Auto V (GTA V), limiting targets to persons. Experimental results show the effectiveness of our approach. Our code and dataset are available at https://github.com/mikittt/easy-to-understand-REG.

1. Introduction

With the popularization of intelligent agents such as robots, symbiosis with them becomes more important. Sharing what humans and agents see naturally is a particularly essential component for smooth communication in the symbiosis environment. In daily life, people often use referring expressions to indicate specific targets such as "a man wearing a red shirt." Further, communicating with agents with natural language is an intuitive method of interaction. When referring to a object with natural language, many expressions can be used that are equally *correct* from a semantic standpoint such that one can locate the target. However, they are not always equally *easy* for target identifications. As shown in Fig. 1, it is important for the expression to be comprehended easily by humans. The comprehension is divided into two processes: understanding the text and finding an referred object in an image. These can be uniformly evaluated by the comprehension time by humans. Thus, we regard that easy-to-understand referring expressions are those that are comprehended correctly and quickly by humans.

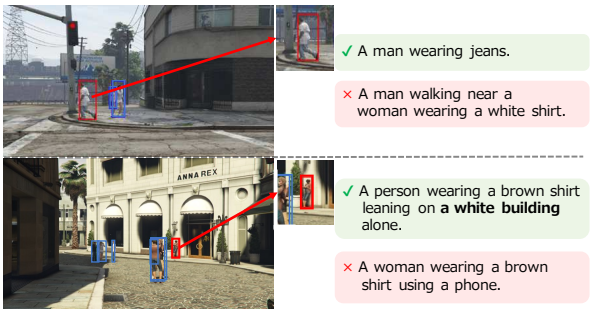


Figure 1. Examples of referring expressions to be generated in this study. In the top image, the target in the red bounding box is sufficiently salient; therefore, a brief description suffices. In the bottom image, referring to other salient objects is required to single out the target because the target itself is not sufficiently salient.

Recently, *correct* referring expression generation has demonstrated significant progress. Considering agents' views that are automatically captured such as in-vehicle images, the compositions of the images are often complex and contain more objects with low saliency than images from MSCOCO [44], which are typically used in the existing works of referring expression generation [15, 17, 24, 23, 37]. The existing studies regarded expressions that refer to objects correctly as equally good. However, if the targets become relatively less salient, identifying the referred objects can become difficult even if the sentences are *correct*.

For the agents to refer to objects in natural language, they should be described clearly for an *easier* comprehension. Expressions utilizing relationships between the targets and other salient contexts such as "a woman by the red car" would help listeners to identify the referred objects when the targets are not sufficiently salient. Thus, expressions to be generated demand the following properties:

- If the target is salient, a brief description suffices.
- If the target is less salient, utilizing relationships with salient contexts around it helps to tell its location.

If these sentences can be generated, drivers can be navigated by utilizing in-vehicle images such as, "please turn right at the intersection by which a man with a red cap stands."

We herein propose a new approach to generate referring expressions that are brief and sufficiently easy for a human to locate a target object in an image without sacrificing the semantic validity. To utilize salient contexts around the target, our model is designed to extract information from the target and from the environment. Moreover, we perform optimization for the expression generator using the time and accuracy metrics for target identifications. Although these quantities by themselves do not tell the absolute level of the goodness of the generated sentences, comparing them among candidate sentences helps to identify a preferable one. We adopt a ranking learning technique in this respect.

To evaluate our system, we constructed a new referring expression dataset with images from GTA V [1], limiting targets to humans for two reasons. (1) Targeting humans is of primary importance for the symbiosis of humans and robots as well as in designing safe mobile agents. The existence of pedestrian detection field [30, 6, 34, 26, 8, 39] also tells the importance of application. (2) Targeting humans is technically challenging because humans have various contexts as they act in various places and their appearances vary widely. We included humans' comprehension time and accuracy in the dataset for the ranking method above.

Overall, our primary contributions are as follows.

- We propose a novel task whose goal is to generate referring expressions that can be comprehended correctly and quickly by humans.
- We propose a optimization method for the task above with additional human annotations and a novel referring expression generation model which captures contexts around the targets.
- We created a new large-scale dataset for the task above based on GTA V (RefGTA), which contains images with complex compositions and more targets with low saliency than the existing referring expression datasets.
- Experimental results on RefGTA show the effectiveness of our approach, whilst the results on existing datasets show the versatility of our method on various objects with real images.

2. Related work

First, we introduce image captioning. Next, we explain referring expression generation that describes specific objects. Finally, we refer to datasets used for referring expression generation and comprehension.

2.1. Image Captioning

Following the advent in image recognition and machine translation with deep neural networks, the encoder-decoder model improved the quality of image captioning significantly, which encodes an image with a deep convolutional neural network (CNN), and subsequently decodes it by a long term-short memory (LSTM) [32]. Many recent approaches use attention models that extract local image features dynamically while generating each word of a sentence [13, 21, 40, 27, 33, 14, 43]. Lu *et al*. [13] introduced a new hidden state of the LSTM called the "visual sentinel" vector. It controls when to attend the image by holding the context of previously generated words, because words such as "the" and "of" depend on the sentence context rather than the image information. Recently, researchers have applied reinforcement learning to directly optimize automatic evaluation metrics that are non-differentiable [40, 33, 14, 43].

2.2. Referring Expression Generation

While image captioning describes a full image, referring expression generation is to generate a sentence that distinguishes a specific object from others in an image. Referring expressions have been studied for a long time as a NLP problem [42, 7]. Recently, large-scale datasets (RefCOCO, RefCOCO+ [24], RefCOCOg [17], etc.) were constructed, and both referring expression generation and comprehension have been developed in pictures acquired in the real world [17, 24, 23, 45, 15, 37, 36, 25, 10, 5]. As these problems are complementary, recent approaches of referring expression generation solve both problems simultaneously [15, 17, 24, 23, 37]. Mao *et al*. [17] introduced max-margin Maximum Mutual Information (MMI) training that solves comprehension problems with a single model to generate disambiguous sentences. Liu *et al*. [15] focused on the attributes of the targets and improved the performance. Yu *et al*. [23] proposed a method that jointly optimizes the speaker, listener, and reinforcer models, and acquired state-of-the-art performance. Their respective roles are to generate referring expressions, comprehend the referred objects, and reward the speaker for generating discriminative expressions.

2.3. Referring Expression Datasets

The initial datasets consist of simple computer graphics [12] or small natural objects [28, 31]. Subsequently, first large-scale referring expression dataset RefCLEF [38] was constructed using images from ImageClef [11]. By utilizing images from MSCOCO, other large-scale datasets such as RefCOCO, RefCOCO+ and RefCOCOg were collected. These useful datasets consist of many images captured by humans, whose compositions are simple with some subjects in the center. For images captured by robots or other intelligent agents, handling more complex images is important. Some existing studies constructed referring expression datasets with images [41] or videos [5] from Cityscapes [29]. However, this is created for comprehension and the sentence should just refer to the target correctly because the listeners are supposed to be machines. We focus on generation and the understandability of the sentence should be considered because the listeners are supposed to be humans. In this respect, we created a new dataset with images from GTA V described in Sec. 4.

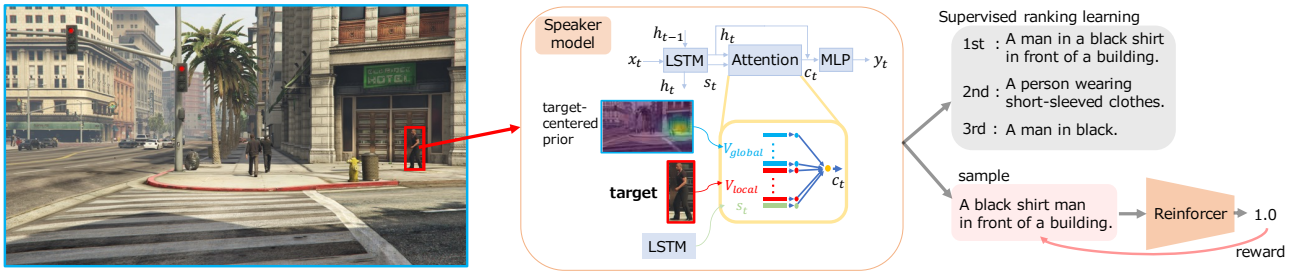


Figure 2. Our model consists of two components. The first one *speaker* is in the middle of the figure. *Speaker* is trained to generate referring expressions with supervised ranking learning and the reward from the second model *reinforcer* in the right side of the figure. *Speaker* attends to features from the target, the context around it, and the sentence context under generation s_t.

3. Model

To generate easy-to-understand referring expressions for target identifications, the model should be able to inform us of the target's location utilizing salient context around it. Similar to normal image captioning, we consider generating sentences word by word, and the context of the sentence information under generation is also utilized. We refer to this context as the sentence context. We assumed the necessary information to generate the sentences as follows.

(A) Salient features of the target
(B) Relationships between the target and salient context around it
(C) Sentence context under generation

We propose a model comprising a novel context-aware speaker and reinforcer. For the context novelty, please see our supplementary material. As reported in [23], joint optimization using both a listener and reinforcer achieves similar performance to using either one in isolation. This is mainly because both of them provide feedback to the neural network based on the same ground truth captions. Instead, we aim to generate more appropriate captions by modifying the speaker given the above assumptions (A), (B) and (C).

Moreover, expressions to be generated should help a human in locating the referred objects correctly and quickly. If the targets are sufficiently salient, brief expressions are preferable for rapid comprehension. We optimized them by comparing the time required to locate the referred objects by humans, and their accuracies among sentences annotated to the same instance.

First, we introduce a state-of-the-art method to generate referring expressions, i.e., the speaker-listener-reinforcer [23]. Next, we explain our generation model. Finally, we introduce the optimization of easy-to-understand referring expressions and describe compound loss.

3.1. Baseline Method

We explain a state-of-the-art method [23]. Three models, speaker, listener, and reinforcer were used. Herein, we explain only the speaker and reinforcer that are used in our proposed model.

Speaker: For generating referring expressions, the speaker model should extract target object features that are distinguished from other objects. Yu *et al*. [23] used the CNN to extract image features and generate sentences by LSTM. First, Yu *et al*. [23] extracted the following five features: (1) target object feature vector o_i, (2) whole image feature vector g_i, (3) the feature encoding the target's coordinate (x, y) and the size (w, h) as $l_i = [\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w_i \cdot h_i}{W \cdot H}]$, (4) difference in target object feature from others $\delta o_i = \frac{1}{n}\sum_{j \neq i} \frac{o_i - o_j}{\|o_i - o_j\|}$, (5) difference in target coordinate from others $\delta l_{ij} = [\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w_j \cdot h_j}{w_i \cdot h_i}]$. Visual feature v_i is obtained by applying one linear layer to these, $v_i = W_m[o_i, g_i, l_i, \delta o_i, \delta l_i]$. Concatenating v_i and the word embedding vector w_t, $x_t = [v_i; w_t]$ is fed into the LSTM and learned to generate sentences r_i by minimizing the negative log-likelihood with model parameters θ:

$$L_s^1(\theta) = -\sum_i \log P(r_i|v_i;\theta) \tag{1}$$

To generate discriminative sentences, they generalized the MMI [17] to enforce the model, to increase the probability of generating sentences r_i if the given positive pairs as (r_i,o_i) than if the given negative pairs as (r_j,o_i) or (r_i,o_k) where r_j and o_k are sampled randomly from other objects, and optimized by following the max-margin loss (λ_1^s, λ_2^s, M_1 and M_2 are hyper-parameters):

$$\begin{aligned} L_s^2(\theta) = \sum_i \{&\lambda_1^s \max(0, M_1 + \log P(r_i|v_k) - \log P(r_i|v_i)) \\ &+\lambda_2^s \max(0, M_2 + \log P(r_j|v_i) - \log P(r_i|v_i))\} \end{aligned} \tag{2}$$

Reinforcer: Next, we explain the reinforcer module that rewards the speaker model for generating discriminative sentences. First, the reinforcer model is pretrained by classifying whether the input image feature and sentence feature are paired by logistic regression. The reinforcer extracts image features by the CNN and sentence features by LSTM and subsequently, feed into MLP by concatenating both features to output a scalar. Next, it rewards the speaker model while fixing its parameters. Because the sampling operation of sentences $w_{1:T}$ is non-differentiable, they used policy-gradient to train the speaker to maximize the reward $F(w_{1:T}, o_i)$ by the following loss:

$$\nabla_\theta J = -E_{P(w_{1:T}|v_i)}[(F(w_{1:T}, v_i))\nabla_\theta \log P(w_{1:T}|v_i;\theta)] \tag{3}$$

3.2. Context-aware speaker model

Our speaker model (in Fig. 2) generates referring expressions that can utilize relationships between targets and salient contexts around the target. Similar to Yu [23], we encoded image features by the CNN, and decoded it into a language by LSTM. In extracting the global features from whole images whose compositions are complex, information around the target objects is more important. We replace global features g_i with g_i' that weight Gaussian distribution whose center is the center of the target (variance is a learnable parameter). We used $v_i = W_m[o_i, g_i', l_i, \delta o_i, \delta l_{ij}]$ as a target image feature to feed into the LSTM.

Next, we introduce the attention module that satisfies the requirements. We begin by defining the notations: V_{global}, V_{local} are the output features of the last convolutional layer on the CNN, containing k, l spatial features respectively ($V_{\text{global}} = [{f^g}_1, \cdots, {f^g}_k], V_{\text{local}} = [{f^l}_1, \cdots, {f^l}_l]$, $V_{\text{global}} \in \mathbb{R}^{d \times k}, V_{\text{local}} \in \mathbb{R}^{d \times l}$). To extract the required information: (A) salient features of the target, (B) relationships with salient context around it, and (C) sentence context under generation, we can use V_{local}, V_{global} for (A) and (B), respectively. As for (C), we used a sentinel vector s_t proposed by Lu *et al*. [13], which is a hidden state of the LSTM calculated as follows: (h_t: hidden state of LSTM, m_t: memory cell of LSTM):

$$s_t = \sigma(W_x x_t + W_h h_{t-1}) \odot \tanh(m_t) \tag{4}$$

For focusing more around the target, we introduce target-centered weighting G_i ($G_i \in \mathbb{R}^{1 \times k}$) with Gaussian distribution, similar as in the feature g_i'. Using four weights, $W_{\text{global}} \in \mathbb{R}^{d \times d}, W_{\text{local}} \in \mathbb{R}^{d \times d}, W_s \in \mathbb{R}^{d \times d}, w_h \in \mathbb{R}^{d \times 1}$, and defining $V_t = [V_{\text{global}}; V_{\text{local}}; s_t]$, our attention α_t is calculated as follows:

$$v_t = [W_{\text{global}} V_{\text{global}}; W_{\text{local}} V_{\text{local}_i}; W_s s_t] \tag{5}$$

$$z_t = w_h^T \tanh(v_t + W_g h_t 1^T) \tag{6}$$

$$\alpha_t = \text{Softmax}([(z_t[:, :k] + \log G_i); z_t[:, k:]]) \tag{7}$$

([;] implies concatenation, and [:,:k] implies to extract the partial matrix up to column k)

Finally, we can obtain the probability of possible words as follows:

$$c_t = \sum_{n=1}^{k+l+1} \alpha_{tn} V_{tn} \tag{8}$$

$$p(w_t | w_1, \cdots, w_{t-1}, v_i) = \text{Softmax}(W_p(c_t + h_t)) \tag{9}$$

3.3. Optimization of easy-to-understand referring expressions

In our task, sentences to be generated should be comprehended by humans (1) correctly and (2) quickly. Although (1) can be learned by the baseline method, (2) is difficult to optimize because defining an absolute indicator that can measure it is difficult. However, we can determine which sentence is better than the others by human annotations. In our task, we used the time required by humans to identify the referred objects and its accuracy for the annotations.

We now consider ranking labels as teacher information. For a target o_i, sentences $\{r_{i1}, \cdots, r_{im}\}$ are annotated. We denote a set of pairs satisfying rank(r_{ip}) $<$ rank(r_{iq}) ($p \neq q, 1 \leq p, q \leq m$) as Ω_i. In this case, the probability of generating r_{ip} should be higher than one of generating r_{iq}. We sample (r_{ip}, r_{iq}) randomly from Ω_i and perform optimization by the max margin loss as follows (λ_3^s and M_3 are hyper-parameters):

$$L_s^3(\theta) = \sum_i \{\lambda_3^s \max(0, M_3 + \log P(r_{iq}|v_i) - \log P(r_{ip}|v_i))\} \tag{10}$$

Moreover, we applied this ranking loss to the reinforcer model. We used the output before the last sigmoid activation to calculate the loss similar to the above Eqn. 10. The final loss function of the reinforcer is both the ranking loss and logistic regression. Similar to Eqn. 3, we can train the speaker to generate sentences to maximize the new reward $F'(w_{1:T}, o_i)$, which estimates how easily the generated expressions can be comprehended by humans as follows:

$$\nabla_\theta J' = -E_{P(w_{1:T}|v_i)}[(F'(w_{1:T}, v_i)) \nabla_\theta \log P(w_{1:T}|v_i; \theta)] \tag{11}$$

We also introduced sentence attention [46] into the sentence encoder of the model to capture the words that would facilitate a human's comprehension of a sentence.

Compund loss: The final loss of our speaker model L_s is a combination of Eqn. 1, Eqn. 2, Eqn. 10 and Eqn. 11 as follows (λ^r is a hyper-parameter.):

$$L_s(\theta) = L_s^1 + L_s^2 + L_s^3 - \lambda^r J' \tag{12}$$

4. Dataset Construction

In our task, the following properties are required in the dataset. (1) The composition of the images are complex. (2) Targets' appearances and locations are sufficiently diverse. However, dataset bias as for (2) tends to occur when collecting a real dataset. We acquired images which satisfy (1) from GTA V because CG can be easily controlled and can guarantee (2). Artificial datasets such as CLEVR [18] are also advantageous as they can isolate and control the qualitative difficulty of the problem and are widely applied in similar problem settings. For real world applications, synthetic data can help improve understanding as in [9] and we can also use unsupervised domain adaptation as in [16]. In this study we constructed a new referring expression dataset, RefGTA, limiting the target type to humans only.

We collected images and information such as a person's bounding boxes automatically, and subsequently annotated the referring expression by humans. (GTA V is allowed for use in non-commercial and research uses [2].)

Figure 3. Images from GTA V. Left : images with unconstrained clothing / Middle : images in which only black-clothed persons exist / Right : images in which only white-clothed persons exist

4.1. Image Collection

First, we extracted images and persons' bounding box information once every few seconds using a GTA V mod that we created (PC single-player mods are allowed [3]).

Moreover, even when multiple persons whose appearances are similar exist, the system should be able to generate expressions where humans can identify referred objects easily by utilizing the relationships between the targets and other objects *etc*. Therefore, we further collected images in which only either white-clothed or black-clothed persons exist, by setting them when the mod starts, as in Fig. 3.

Finally, we deleted similar images by the average hash. We also deleted images comprising combinations of the same characters. We set the obtained images and bounding box information as a dataset.

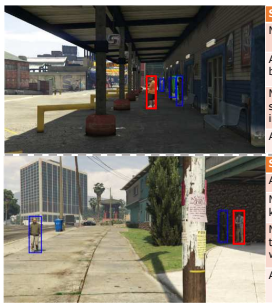
4.2. Sentence Annotation

We annotate sentences to each instance obtained in Sec. 4.1 by the following two steps. We requested the annotations of the Amazon Mechanical Turk (AMT) workers.

Annotating sentences: First, we requested the AMT workers to annotate five descriptions that are distinguished from the others for each instance. We instructed the workers to annotate a sentence that refers only to the target and is easy to distinguish from others at a glance. We also instructed the workers to use not only the target attributes but also the relative positions to other objects in the image. We instructed them not to use absolute positions inside the image and allowed the relative positions to other objects.

Validating sentences: Next, we assigned five AMT workers to localize the referred person in each description to verify whether it is an appropriate referring expression. If a referred person does not exist, we allow them to check the box, "impossible to identify." We displayed the elapsed time on the task screen and instructed the workers to obtain the referred objects as quickly as possible. We included the sentences where more than half of the workers accurately obtained the referred persons in a dataset. We also recorded the time and accuracy of five workers for each sentence.

Examples: We show the annotation examples in Fig. 4. The rightmost column is the ranking we used in Sec. 3.3. This is calculated as follows: first, all sentences are ranked by humans' comprehension accuracy; subsequently, sentences that are comprehended correctly by all workers are ranked by time. This ranking is performed as follows. When comparing the times of two sentences we take the time of three people in the middle of five people to reduce the influence of outliers. We consider sentence "A" as better than sentence "B" if the mean of "B" subtracted by the mean of "A" is greater than the sum of their standard errors. For each sentence we count the number of sentences that it is better than and rank the sentence according to this number.



Sentence	Acc	Time(s)	Rank
Man in orange shirt.	100%	2.49±0.46	1
A man in a orange short sleeve shirt and black shorts standing in the shade.	100%	4.23±1.32	3
Man in red cap short sleeved orange shirt black shorts white shoes standing in front of building entrance.	100%	5.26±0.83	3
A man in an orange shirt.	100%	3.34±0.39	1

Sentence	Acc	Time(s)	Rank
A man in khakis standing next to a wall.	100%	3.36±0.10	1
Man texting in a gray sweatshirt and khaki pants.	100%	4.28±0.33	2
Man in white cardigan standing near wall taking to another man while looking down at phone.	100%	5.71±1.16	2
A man wearing a grey jacket.	80%	4.21±0.60	4

Figure 4. Example data. Sentence: Annotated captions. Acc: Human's comprehension accuracy. Time (s): the time required by human to search. Rank: The ranking we assigned by the accuracy and time as described in this section.

	Train	Val	Test
# of images	23,950	2,395	2,405
# of created instances	65,205	6,563	6,504
# of referring expressions	177,763	17,766	17,646

Table 1. Statistics of annotations on the dataset.

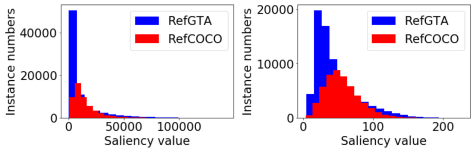


Figure 5. Targets' saliency of RefCOCO and RefGTA. Left: saliency is calculated by the sum of the saliency score inside the target bounding box. Right: saliency is normalized by dividing the square root of the area.

4.3. Statistical Information

We show the statistics of our dataset, RefGTA. The scale of RefGTA is presented in Table 1. The resolution of the image is 1920×1080. The mean length of annotated sentences is 10.06. We compared the saliency of the target using saliency model proposed by Itti *et al*. [22], which is commonly used. First, we calculated a saliency map of a whole image, and we used the value in a bounding box of a target. As in Fig. 5, RefGTA contains more targets with low saliency as compared to RefCOCO. In this case, the relationships between the targets and salient context around them becomes more important.

	RefCOCO		RefCOCO+		RefCOCOg
	Test A	Test B	Test A	Test B	val
SLR (ensemble) [23]	80.08%	81.73%	65.40%	60.73%	74.19%
re-SLR (ensemble)	78.43%	81.33%	64.57%	60.48%	70.95%
baseline: re-SLR (Listener)	**81.14%**	80.80%	**68.16%**	59.69%	72.36%
Our SR (Reinforcer)	80.44%	**81.04%**	67.81%	58.97%	**74.94%**
Our SLR (Listener)	79.05%	80.31%	65.75%	**62.18%**	73.39%
baseline: re-SLR (Speaker)	80.70%	81.71%	68.91%	60.77%	72.55%
Our SR (Speaker)	82.45%	**82.00%**	72.07%	**61.06%**	70.35%
Our SLR (Speaker)	**83.05%**	81.84%	**72.37%**	59.13%	**74.79%**

Table 2. Comprehension evaluation on RefCOCO, RefCOCO+ and RefCOCOg. () implies the model used. Ensemble implies to use both speaker and listener or reinforcer. Our speaker demonstrates comparable or better performance in most cases.

	Test
baseline: re-SLR (Listener)	**86.05%**
Our SR (Reinforcer)	85.25%
Our SLR (Listener)	84.04%
baseline: re-SLR (Speaker)	84.84%
Our SR (Speaker)	88.41%
Our SLR (Speaker)	**88.60%**
baseline: re-SLR (ensemble)	86.55%
Our SR (ensemble)	89.16%
Our SLR (ensemble)	**89.54%**

Table 3. Comprehension evaluation on RefGTA. Our speaker model exhibits high comprehension performance, and its ensembling exceeds that of re-SLR.

5. Experiments

First, we explain the datasets used in our study. Next, we describe the results of comprehension, ranking, and generation evaluation in this order. Finally, we evaluate the generated sentences by humans.

We refer to the state-of-the-art method for generation [23] as "SLR." The SLR originally used VGGNet [20] as its image feature encoder. We also used ResNet152 [19] which achieved better performance on image classification. We compared re-implemented SLR and our model that we refer as "re-SLR", and "our SR" respectively. We set re-SLR with ResNet as a baseline. Our SLR implies our SR with the re-implemented listener.

5.1. Datasets

We conducted experiments on both existing datasets (RefCOCO, RefCOCO+ [24] and RefCOCOg [17]) and our dataset (RefGTA). Our primarily purpose is the evaluation on RefGTA, whilst we used existing datasets to evaluate versatility of our method on various objects with real images. Yu *et al*. [23] collected more sentences for the test sets of RefCOCO and RefCOCO+; therefore, we used these for generation evaluation.

5.2. Comprehension Evaluation

We compared comprehension performance of the speaker, listener, and reinforcer. Given a sentence r, each comprehension by reinforcer and speaker is calculated by $o^* = \arg\max_i F(r, o_i)$ and $o^* = \arg\max_i P(r|o_i)$, respectively. We used ground truth bounding boxes for all the objects. We only compared our method with the state-of-the-art model for generation [23] because our purpose is generation and we cannot compare methods for comprehension (e.g. [25]) fairly.

	Test
baseline: re-SLR (Reinforcer)	55.89%
baseline: re-SLR (Speaker)	55.99%
Our SR (Reinforcer)	55.46%
Our SR (Speaker)	56.38%
Our SR (Reinforcer) + rank loss	**57.55 %**
Our SR (Speaker) + rank loss	56.64%

Table 4. Accuracy of classifying ranked pairs. Ranking loss improved its performance.

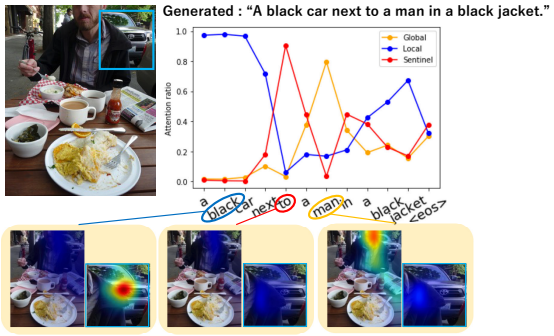


Figure 6. Generation example on RefCOCOg and each attention transition. Each of attention values corresponds to the sum of the softmax probability divided by local, global and sentinel in Eqn. 7 respectively and their sum equals to one for each word.

Results on existing datasets: First, we demonstrate the comprehension performance on RefCOCO, RefCOCO+ and RefCOCOg in Table 2. Although our speaker demonstrates comparable or better performance in most cases, we focus on the sentence generation, and the model with higher comprehension performance does not always generate better sentences. Because both the listener and reinforcer used in [23] have a similar role as described in Sec. 3, we obtained similar results from our SR and our SLR.

Results on RefGTA: Next, we demonstrate the comprehension performance of the system on RefGTA in Table 3. Although the listener's comprehension accuracy is better for re-SLR, our speaker's comprehension accuracy is higher than that of the re-SLR, and our model is best when ensembling a speaker and listener models. The accuracy on Table 3 is higher than the accuracies on Table 2 because we constructed large-scale dataset limiting targets to humans.

5.3. Ranking Evaluation on RefGTA

We evaluated the ranking accuracy by classifying a given pair into two classes; whether the given two expressions are correctly ranked or not. First, we extracted the set of ranking pair as described in Sec. 4.2. The number of all pairs is 13,023. The results are shown in Table 4. "Rank loss" implies that we adopt the ranking loss for both speaker and reinforcer as we explain in Sec. 3.3. Both of them improved the ranking performance by rank loss. This implies that rank loss helps our model learning expressions comprehended by humans correctly and quickly.

5.4. Generation Evaluation

Qualitative results on existing datasets: Generated sentence example on RefCOCOg is shown in Fig. 6. While the

	features	RefCOCO				RefCOCO+				RefCOCOg	
		Test A		Test B		Test A		Test B		val	
		Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr
SLR [23]	VGGNet	0.268	0.697	0.329	1.323	0.204	0.494	0.202	0.709	0.154	0.592
SLR+rerank [23]	VGGNet	0.296	0.775	0.340	1.320	0.213	0.520	0.215	0.735	0.159	0.662
re-SLR	VGGNet	0.279	0.729	0.334	1.315	0.201	0.491	0.211	0.757	0.146	0.679
re-SLR+rerank	VGGNet	0.278	0.717	0.332	1.262	0.198	0.476	0.206	0.721	0.150	0.676
baseline: re-SLR	ResNet	0.296	0.804	0.341	1.358	0.220	0.579	0.221	0.798	0.153	0.742
Our Speaker only	ResNet	0.301	**0.866**	0.341	**1.389**	**0.243**	**0.672**	0.222	**0.831**	0.163	0.746
Our SR (w/o local attention)	ResNet	0.289	0.760	0.328	1.278	0.214	0.542	0.210	0.753	0.156	0.666
Our SR (w/o global attention)	ResNet	0.307	0.845	0.335	1.331	0.237	0.654	0.220	0.822	0.163	0.714
Our SR (w/o sentinel attention)	ResNet	0.303	0.851	0.340	1.358	0.238	0.663	0.219	0.819	0.164	0.746
Our SR	ResNet	0.307	0.865	0.343	1.381	0.242	0.671	0.220	0.812	0.164	0.738
Our SR+rerank	ResNet	0.310	0.842	**0.348**	1.356	0.241	0.656	0.219	0.782	0.167	0.773
Our SLR	ResNet	0.310	0.859	0.342	1.375	0.241	0.663	0.225	0.812	0.164	0.763
Our SLR+rerank	ResNet	**0.313**	0.837	0.341	1.329	0.242	0.664	**0.228**	0.787	**0.170**	**0.777**

Table 5. Generation results using automatic evaluation. We used the test set of RefCOCO and RefCOCO+ extended by Yu *et al*. [23]. While the generation qualities are high by the speaker only in RefCOCO and RefCOCO+, rerank improves the performance in RefCOCOg.

value of local attention is high when explaining the target car, the value of global attention becomes high when mentioning objects outside of the target. When switching from local attention to global attention, the value of sentinel attention that holds the sentence context becomes higher.

Quantitative results on existing datasets: Next, we discuss the quantitative evaluation based on the automatic evaluation metric, CIDEr [35] and Meteor [4]. Because ground-truth sentences are referring expressions, we can evaluate them to some extent. Our re-implemented rerank did not improve the generation performance although Yu *et al*. [23] reported that reranking improves performance. In RefCOCO and RefCOCO+, the generation qualities are high by the speaker only. Meanwhile, in RefCOCOg, rerank helped to improve the performance. This is because while the model should generate one phrase in RefCOCO and RefCOCO+, the model should generate a full sentence in RefCOCOg and has to solve more complex problems including satisfying language structures.

Qualitative results on RefGTA: Next, we demonstrate the generated sentence examples on RefGTA in Fig. 7. While the baseline method (re-SLR) demonstrates lower capability in capturing the outside of the target than our method, our method can generate sentences that can identify the target easier especially in the right-side examples. As shown in the left-bottom example, while the baseline method generates a brief and sufficient description, our SR+rank loss also generates the same one. Attention visualization is shown in Fig. 8. While the local attention value is high when describing the target, the global attention value is high when mentioning "building," which is outside of the target.

Quantitative results on RefGTA: Finally, we demonstrate the quantitative evaluation on RefGTA. In our study, the ideal metric should assign a high score to a sentence that can be easily comprehended by humans correctly and quickly. While CIDEr calculates the average similarity between a generated sentence from an object o_i and ground-truth sentences $\{r_{i1}, \cdots, r_{im}\}$; we define the ranking-weighted CIDEr (R-CIDEr) which utilizes weighted similarity scores between them by the inverse of their rank.

	Test			
	Meteor	CIDEr	R1-CIDEr	R2-CIDEr
baseline: re-SLR	0.263	0.966	0.994	0.976
Our Speaker only	0.278	1.014	1.038	1.025
Our SR (w/o local attention)	0.208	0.557	0.570	0.561
Our SR (w/o global attention)	0.276	1.036	1.065	1.047
Our SR (w/o sentinel attention)	0.278	1.022	1.049	1.033
Our SR	**0.279**	1.036	1.065	1.048
Our SR+rank loss	0.277	**1.047**	**1.078**	**1.059**
Our SLR	0.278	1.030	1.054	1.041

Table 6. Generation evaluation on RefGTA. Without rank loss, Our SR with all modules is the best. Furthermore, ranking improved performance.

	Unspecified color	Black wearing	White wearing	All	All (selected)
baseline	69.52%	60.96%	71.09%	67.48%	73.17%
Our SR	**75.67%**	64.61%	**72.93%**	71.52%	75.83%
Our SR+rank loss	74.89%	**68.76%**	72.30%	**72.25%**	**76.94%**

Table 7. Left three columns: the evaluation of humans' comprehension accuracy when divided by clothing types as seen in Fig. 3, All: The rate for which annotators were able to select the correct target, All (selected): The accuracy ignoring "impossible to identify" choices.

The weight of the sentence r_{ij} is calculated as $w(r_{ij}) = \left(rank(r_{ij}) \sum_j rank(r_{ij})^{-1}\right)^{-1}$. This metric assigns a high score to sentences where a human identified the referred objects correctly and quickly. In Table 6, R1-CIDEr implies using ranking by humans' comprehension accuracy and time required, and R2-CIDEr implies using ranking by only humans' comprehension accuracy. In particular, R1-CIDEr that we optimized is improved by the ranking loss. Rerank was not applicable in RefGTA.

5.5. Human Evaluation on RefGTA

Human comprehension evaluation: First, we evaluated human comprehension for the generated sentences by each method. We used 600 targets extracted randomly from the test data, and requested 10 AMT workers to identify the referred persons while measuring the time. If no referred target exists, we allow them to check a box, "impossible to identify." The results including clothing type evaluations are shown in Table 7. Our model outperformed the baseline method, and the rank loss improved the performance in black wearing case. Our SR+rank loss was the best for the overall performance.

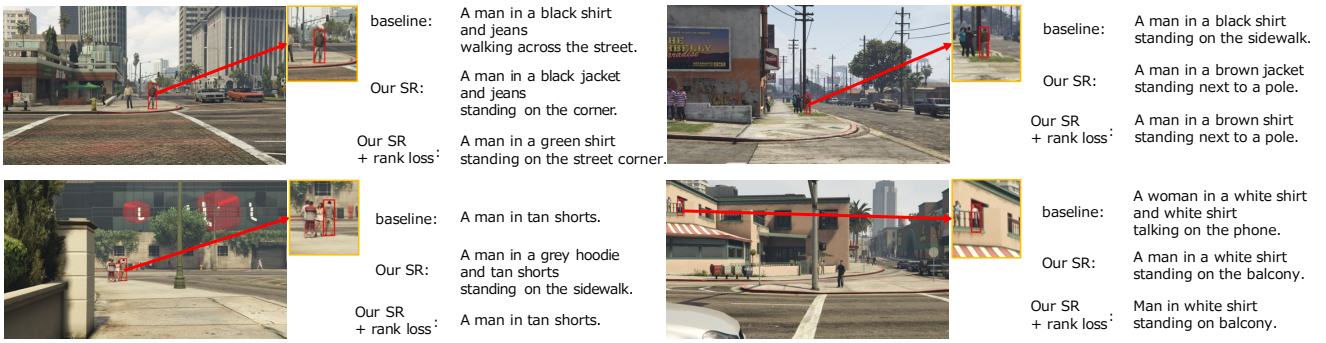


Figure 7. Comparison of generated sentences by each method on RefGTA. Rank loss implies to be trained with ranking.

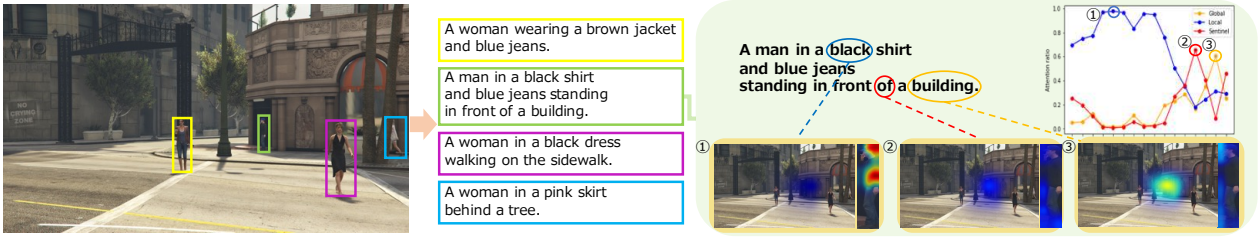


Figure 8. Generation example on RefGTA and each attention transition. Each sentence is generated from an object of the same color.

	Accuracy only	Accuracy and time
baseline	30.08%	30.86%
Our SR	34.08%	33.28%
Our SR+rank loss	**35.83%**	**35.86%**
difference between proposed methods	1.75%	2.58%

Table 8. Comparison of our generated sentences in terms of humans' comprehension accuracy and the time required to locate the referred objects. For all methods, the sum of accuracies is 100%. When including time in the comparison, the difference between the proposed methods increases. This shows the efficacy of using rank loss.

Time evaluation: Next, we evaluated whether our method improved performance based on the time required by humans to locate referred objects. We evaluated as follows: first, all sentences are ranked by humans' comprehension accuracy; subsequently, sentences that are comprehended correctly by all workers (i.e., comprehension accuracy is 100%) are ranked by the average time; finally, for the remaining sentences, we calculated the ratio of the number of instances that are ranked first in each method (if there are 2 or 3 instances ranked first, the number is counted as 1/2, and 1/3 respectively.) The obtained results (see Table 8) show that rank loss improved not only human comprehension accuracy but also the time.

Human comprehension evaluation considering saliency: Finally, we evaluated our method when the saliency of the target changes. We evaluated the relationship between humans' comprehension accuracy and targets' saliency. We calculated saliency as described in Sec. 4.3. We present the results in Fig. 9. As shown, our model performs better on the lower saliency area because mentioning salient contexts around the targets helped humans to comprehend them. The difference between the methods becomes smaller as the saliency becomes higher.

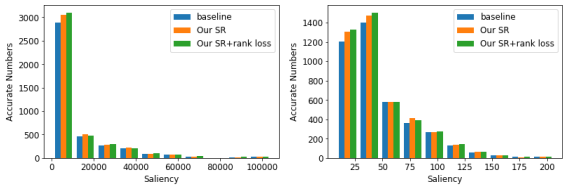


Figure 9. Relationship between the number of people who answered correctly and saliency calculated as Fig. 5.

6. Conclusions

We herein focused on generating referring expressions that allowed for humans to identify referred objects correctly and quickly. We proposed a model that could utilize relationships between targets and contexts around them to generate better sentences even when the compositions of the images were complex, and the targets were not sufficiently salient. We also proposed a method to optimize referring expressions that are easy for target identifications with additional annotations. To evaluate our system, we constructed a new dataset, RefGTA. We demonstrated that our method improved referring expression generation not only on the existing automatic evaluation metric, but also on the newly proposed automatic evaluation metric and human evaluation.

Acknowledgment

This work was supported by JST ACT-I Grant Number JPMJPR17U5, partially supported by JSPS KAKENHI Grant Number JP17H06100 and partially supported by JST CREST Grant Number JPMJCR1403, Japan. We would like to thank Atsushi Kanehira, Hiroaki Yamane and James Borg for helpful discussions.

References

[1] Rockstar Games. Grand Theft Auto V. http://www.rockstargames.com/V.

[2] Rockstar Games. PC single-player mods. https://support.rockstargames.com/articles/200153756/Policy-on-posting-copyrighted-Rockstar-Games-material.

[3] Rockstar Games. Policy on posting copyrighted Rockstar Games material. https://support.rockstargames.com/articles/115009494848/PC-Single-Player-Mods.

[4] Alon Lavie and Abhaya Agarwal. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Workshop on Statistical Machine Translation*, pages 228–231, 2007.

[5] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object Referring in Videos with Language and Human Gaze. In *CVPR*, 2018.

[6] Constantine Papageorgiou and Tomaso Poggio. A Trainable System for Object Detection. In *IJCV*, 2000.

[7] Emiel Krahmer and Kees van Deemter. Computational Generation of Referring Expressions: A survey. In *Computational Linguistics*, volume 38, pages 173–218, 2012.

[8] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating Pedestrians via Simultaneous Detection Segmentation. In *ICCV*, 2017.

[9] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from Synthetic Humans. In *CVPR*, 2017.

[10] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding Referring Expressions in Images by Variational Context. In *CVPR*, 2018.

[11] Hugo Jair Escalante, Carlos A. Hernández, Jesus A. Gonzalez, A. López-López, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated IAPR TC12 benchmark. In *CVIU*, 2010.

[12] Jette Viethen and Robert Dale. The Use of Spatial Relations in Referring Expression Generation. In *INLG*, 2008.

[13] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In *CVPR*, 2017.

[14] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural Baby Talk. In *CVPR*, 2018.

[15] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring Expression Generation and Comprehension via Attributes. In *ICCV*, 2017.

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. CyCADA: Cycle Consistent Adversarial Domain Adaptation. In *ICML*, 2018.

[17] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*, 2016.

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Fei-Fei Li, Larry Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[20] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[21] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[22] Laurent Itti, Christof Koch, and Ernst Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. In *Trans. on PAMI*, volume 20, pages 1254–1259, 1998.

[23] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A Joint Speaker-Listener-Reinforcer Model for Referring Expressions. In *CVPR*, 2017.

[24] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *ECCV*, 2016.

[25] 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 *CVPR*, 2018.

[26] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is Faster R-CNN Doing Well for Pedestrian Detection? In *ECCV*, 2016.

[27] Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. Image Caption with Global-Local Attention. In *AAAI*, 2017.

[28] Margaret Mitchell, Kees van Deemter, and Ehud Reiter. Natural Reference to Objects in a Visual Domain. In *INLG*, 2010.

[29] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016.

[30] Michael Oren, Constantine Papageorgiou, Pawan Sinha, Edgar Osuna, and Tomaso Poggio. Pedestrian Detection Using Wavelet Templates. In *CVPR*, 1997.

[31] Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. Learning Distributions over Logical Forms for Referring Expression Generation. In *EMNLP*, 2013.

[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[33] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 2018.

[34] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian Detection: A Benchmark. In *CVPR*, 2009.

[35] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015.

[36] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling Relationships in Ref-

erential Expressions with Compositional Modular Networks. In *CVPR*, 2017.

[37] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *CVPR*, 2017.

[38] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. ReferIt Game: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014.

[39] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded Pedestrian Detection Through Guided Attention in CNNs. In *CVPR*, 2018.

[40] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical Sequence Training for Image Captioning. In *CVPR*, 2017.

[41] Tao Zhou and Jie Yu. Natural Language Person Retrieval. In *AAAI*, 2017.

[42] Terry Winograd. Understanding natural language. In *Cognitive psychology*, volume 3, pages 1–191, 1972.

[43] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring Visual Relationship for Image Captioning. In *ECCV*, 2018.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Larry Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[45] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling Context Between Objects for Referring Expression Understanding. In *ECCV*, 2016.

[46] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answerings. In *ICCV*, 2017.