

# IMPROVING IMAGE PARAGRAPH CAPTIONING WITH DUAL RELATIONS

Yun Liu<sup>†</sup>, Yihui Shi<sup>†</sup>, Fangxiang Feng<sup>†</sup>, Ruifan Li<sup>†\*</sup>, Zhanyu Ma<sup>†‡</sup>, Xiaojie Wang<sup>†</sup>

<sup>†</sup> School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China

<sup>‡</sup> Beijing Academy of Artificial Intelligence, Beijing, China

{yunliu, yhshi, fxfeng, rfli, mazhanyu, xjwang}@bupt.edu.cn

## ABSTRACT

Image paragraph captioning aims to generate multiple descriptive sentences for an image. However, most previous methods ignore the explicit relations among objects resulting in unsatisfactory performance. In this paper, we propose a novel model (i.e., DualRel) to capture spatial and semantic relations among objects. Specifically, the spatial relation embedding is obtained solely from images using a predefined geometry pattern. With the help of captions, the semantic relation embedding is learned in a weakly supervised manner. These two relation embeddings are then interacted with regional features of objects through a relation-aware attention interaction. It first obtains a visual context vector using regional features. Then with the visual context vector, we obtain the corresponding spatial and semantic relation-aware vectors using attentions. These three vectors are fused with two gates for language decoding to further generate a paragraph. Experimental results on Stanford benchmark dataset show that DualRel achieves remarkable improvements<sup>1</sup>.

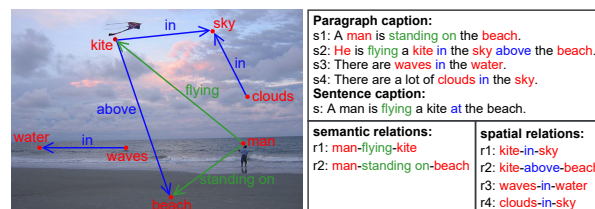
**Index Terms**— Image paragraph captioning, relation embeddings, relation-aware attention.

## 1. INTRODUCTION

The task of image paragraph captioning [1] aims to automatically generate multiple sentences for an image. Compared with the conventional sentence captioning task, such as [2, 3, 4], paragraph captioning is more challenging which requires more demanding visual and linguistic details. Recently, following the pioneering work of Krause et al. [1], various variants have been proposed [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. In these methods, objects in images are often extracted by an object detector and then represented as regional features. A language decoder is then adopted to implicitly learn the relations among objects for generating captions. However, these methods do not explicitly model specific types of relations among objects and the subsequent handling is not targeted. Thus, relations among objects are not effectively captured and utilized resulting in unsatisfactory performance.

\* Corresponding author.

<sup>1</sup>Code released at <https://github.com/fuyun1107/DualRel>



**Fig. 1.** Semantic and spatial relations for an image and its paragraph and sentence descriptions are shown. The corresponding semantic and spatial relation are grouped into two boxes. Relations in the sentence caption are relatively fewer.

Intuitively, relations among objects, e.g., the semantic and spatial relations could significantly enrich the details for image paragraph captioning. As shown in Fig. 1, multiple objects including ‘beach’, ‘kite’, ‘water’, ‘man’ and ‘clouds’ appears in an image. And relations of these objects including spatial relations (‘above’ and ‘in’) and semantic relations (‘flying’ and ‘standing on’) are described in paragraph caption. Compared with the sentence captioning, paragraph captioning includes more relations, as shown in Fig. 1. Note that the spatial relation can be obtained solely from images, but the semantic relation should be learned from both images and language. Furthermore, when describing an image, human first observes a salient object and then he/she pays attention to the object’s relations with other objects. In Fig. 1, take the first sentence as an example; we first observe a man in the image, and then further notice that he is ‘standing on’ (i.e., a relation) the beach. This cognition process motivates a plausible use of these relations.

In this paper, we propose DualRel model to explicitly capture semantic and spatial relations among objects for image paragraph captioning. 1) We design a relation embedding module which contains spatial and semantic relation encoder. The spatial relation encoder emphasizes the spatial position between overlapping objects. The semantic relation encoder captures the language related semantic relations between objects, which is learned in a weakly supervised manner with the help of captions. 2) We design a relation-aware interaction module which is responsible for incorporating semantic and spatial relations embeddings with regional features of objects. Specifically, we first obtain a visual context vectors us-

ing regional features. Then with the visual context vectors, we obtain the corresponding semantic and spatial relation-aware vectors. Finally, these three vectors are fused with gates for language decoding to generate a paragraph.

We highlight our major contributions as follows.

1) We propose DualRel model for image paragraph captioning. DualRel explicitly models semantic and spatial relations among objects to enrich details of generated paragraphs.

2) We design a relation embedding module and a relation-aware interaction module. In the former, semantic relation embedding which is learned in a weakly-supervised manner, and spatial embedding is obtained solely from images. In the latter, relation embeddings and object-level features are fused.

3) We conduct extensive experiments on Stanford benchmark dataset. Experimental results showcase that our model achieves remarkable improvements compared with strong baselines on several popular metrics.

## 2. RELATED WORK

Previous works on image paragraph captioning can be divided into two categories: implicit relations modeling methods [1, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16] and explicit relation modeling methods [17, 18]. The former methods generate paragraphs by first extracting individual objects in images and then representing them as regional features. A subsequent language decoder is adopted to implicitly learn the relations among objects. Major improvements have been made by enhancing coherence among sentences [5, 6], by using reinforcement learning [8, 12], by distilling fine-grained information with attention mechanism [7, 10, 14], by exploring semantic richer visual encoding [11], and by using hierarchical supervision over words and sentences [9]. All these methods are not explicitly capture the relations among objects, which are beneficial for generating rich and coherent captions.

Explicit relations modeling approaches are relatively few. Che et al. [17] explicitly infer relations between objects and then take valid relations into an LSTM for generating captions. However, this method requires a large scale training data for relations classification. Most recently, Yang et al. [18] proposed OR-ATT model to encode relations among objects. OR-ATT and our DualRel have two major differences. First, OR-ATT only considers coarse-grained relation, while our DualRel exploits specific semantic and spatial relation. Second, OR-ATT uses a naïve addition of relation encoding and object regional features, while our DualRel uses relation-aware interaction to fuse relations with object features.

## 3. PROPOSED MODEL

Image paragraph captioning is to generate a paragraph  $Y=\{y_1, \dots, y_T\}$  with  $T$  words for an image. The architecture of our DualRel model is illustrated in Fig. 2.

### 3.1. Relation Embedding Module

**Spatial Relation Encoder.** We detect  $N$  objects as  $C=\{c_1, \dots, c_N\}$  in an image with Faster R-CNN [19]. Let  $V=\{v_1, \dots, v_N\}$ ,  $v_k \in \mathbb{R}^{2048}$  be the visual features. We have bounding boxes  $B=\{b_1, \dots, b_N\}$ , and each box  $b$  is denoted as  $(x, y, w, h)$  with its center coordinates, width and height.

We note that spatial relations (e.g., ‘in front of’) can often be observed from images. Our spatial relation embeddings  $P = \{p_{ij} : p_{ij} \in \mathbb{R}^D\}$  is given as a predefined geometry pattern. Like [20], the geometric relation of two bounding boxes is defined as a vector  $\lambda(i, j) \in \mathbb{R}^4$ , i.e.,

$$\lambda(i, j) = \left( \log\left(\frac{|x_i - x_j|}{w_i}\right), \log\left(\frac{|y_i - y_j|}{h_i}\right), \log\left(\frac{w_i}{w_j}\right), \log\left(\frac{h_i}{h_j}\right) \right) \quad (1)$$

We project  $\lambda(i, j)$  into a high dimensional space as  $E_b(i, j)$ . Then our spatial relation embeddings  $p_{ij}$  is given as,

$$\begin{aligned} v_k'' &= \text{ReLU}(W_p v_k + b_p) \\ p_{ij} &= f_p(\text{Concat}(E_b(i, j), v_i'', v_j'')) \end{aligned} \quad (2)$$

where  $W_p \in \mathbb{R}^{D \times 4D}$  and  $b_p \in \mathbb{R}^D$ , and  $v_k''$  is the low dimensional projection of the object’s feature vector. The learnable nonlinear function  $f_p(\cdot)$  is a two-layer MLP.

**Semantic Relation Encoder.** The semantic relation encoder captures relations (e.g., ‘flying’) between two objects. To this aim, the encoder requires both visual perception and inference combined with language knowledge. 1) The object category embedding  $E_c(i, j)$  of two objects is defined as,

$$E_c(i, j) = \text{ReLU}(W_c \text{Concat}(W_g c_i, W_g c_j) + b_c) \quad (3)$$

where  $W_c \in \mathbb{R}^{D \times 2D}$  and  $b_c \in \mathbb{R}^D$ .  $W_g \in \mathbb{R}^{D \times D_c}$  is a fixed object category embedding matrix initialized by GloVe.  $D_c$  is the total number of object categories. 2) The semantic relation embeddings  $E = \{e_{ij} : e_{ij} \in \mathbb{R}^D\}$  is given as,

$$\begin{aligned} v_k' &= \text{ReLU}(W_s v_k + b_s) \\ e_{ij} &= f_e(\text{Concat}(E_b(i, j), E_c(i, j), v_i', v_j')) \end{aligned} \quad (4)$$

where  $W_s \in \mathbb{R}^{D \times 4D}$  and  $b_s \in \mathbb{R}^D$ , and  $v_k'$  is the low dimensional projection of the object’s feature vector  $v_k$ . The learnable nonlinear function  $f_e(\cdot)$  is a two-layer MLP.

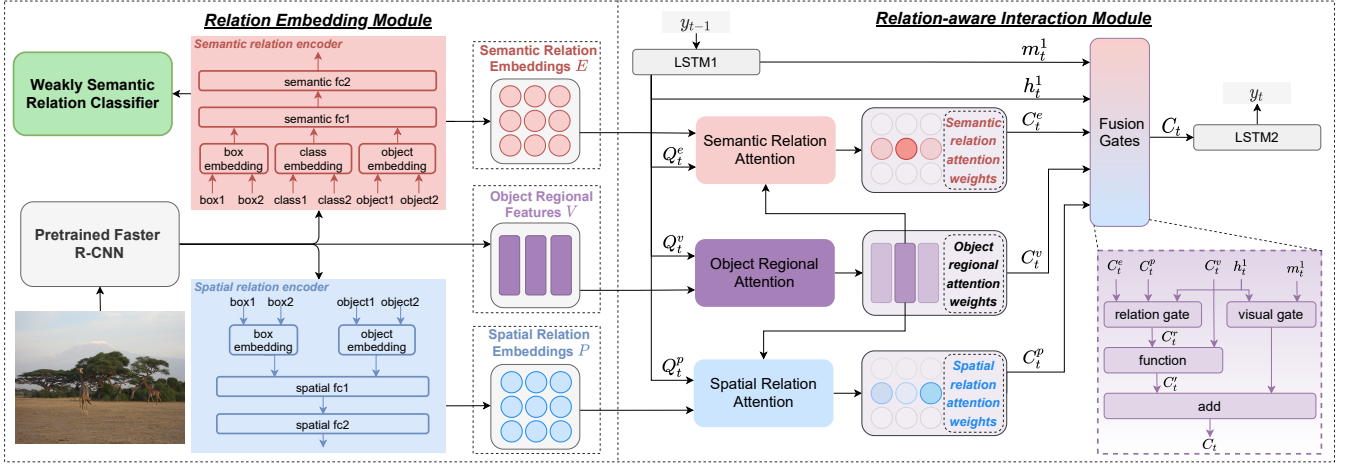
3) To boost semantic relations, we design a weakly supervised multi-label classifier. We feed the semantic relation embedding  $e_{ij}$  into a linear layer to obtain the category scores,

$$\hat{r}_{ij} = W_r e_{ij} + b_r \quad (5)$$

where  $W_r \in \mathbb{R}^{D_r \times D}$  and  $b_r \in \mathbb{R}^{D_r}$ ,  $D_r$  is the number of semantics relations. We collect semantic relation triplets from Visual Genome. Note that the collected triplets only provide common prior knowledge but not the ground truth.

### 3.2. Relation-aware Interaction Module

**Relation-aware Attentions.** With the object regional features  $V$ , semantic and spatial relation embeddings  $\{E, P\}$ ,



**Fig. 2.** The framework overview of our DualRel. It comprises relation embedding and relation-aware interaction modules.

we construct relation-aware attentions. **1)** We first obtain the visual context vector  $C_t^v$  at time step  $t$  as follows,

$$\alpha_{it}^v = \text{Softmax}(W_{v1} \tanh(W_v v_i + W_{vh} h_t^1))$$

$$C_t^v = \sum_{i=1}^N \alpha_{it}^v v_i \quad (6)$$

where  $W_v \in \mathbb{R}^{D \times 4D}$ ,  $\{W_{vh}, W_{v1}\} \in \mathbb{R}^{D \times D}$ .  $\alpha_{it}^v$  denotes an attention weight for object feature  $v_i$  at time step  $t$ .

**2)** With visual context vector, we employ semantic and spatial relation attention to generate relation-aware context vectors. Semantic relation context vector  $C_t^e$  is generated as,

$$\alpha_{it}^e = \text{Softmax}(W_{e1} \tanh(W_e e_{mi} + W_{eh} h_t^1))$$

$$C_t^e = \sum_{i=1}^N \alpha_{it}^e e_{mi} \quad (7)$$

where  $W_e$ ,  $W_{eh}$  and  $W_{e1}$  are all in  $\mathbb{R}^{D \times D}$ .  $\alpha_{it}^e$  denotes an attention weight for semantic relation embeddings  $e_{mi}$ . Here,  $e_{mi} \in \mathbb{R}^D$  is the  $i$ -th spatial relation embedding for the  $m$ -th object region. We obtain the  $m$ -th object region by taking the maximum of the object regional attention, i.e., considering the salient object when generating captions. Similarly, we obtain the spatial relation context vector  $C_t^p$  at time step  $t$ .

**Fusion Gates.** We fuse the three context vectors  $C_t^v$ ,  $C_t^e$  and  $C_t^p$  for language decoding. **1)** We observe that there often exists only one dominant relation in captions. Thus we design a relation gate  $g_r$  to control the degree of two relations,

$$g_r = \sigma(W_{re} C_t^e + W_{rp} C_t^p + W_{r1} h_t^1)$$

$$C_t^r = C_t^p \odot g_r + C_t^e \odot (1 - g_r) \quad (8)$$

$$C_t' = \text{LayerNorm}(C_t^v + W_r(C_t^r))$$

where  $W_{re}$ ,  $W_{rp}$ ,  $W_{r1}$  and  $W_r$  are in  $\mathbb{R}^{D \times D}$ ,  $\sigma$  denotes the logistic function,  $C_t^r$  is relation context vector, and  $\odot$  denotes element-wise product. Then, we obtain the relation-aware visual context vector  $C_t'$ .

**2)** We define a visual gate  $g_v$  to decide when to use relation-aware visual context information or language context information to obtain final context vector  $C_t$  as,

$$g_v = \sigma(W_{vx} x_t + W_{v1} h_t^1)$$

$$C_t = C_t' \odot g_v + \tanh(m_t^1) \odot (1 - g_v) \quad (9)$$

where  $W_{vx} \in \mathbb{R}^{D \times 3D}$  and  $W_{v1} \in \mathbb{R}^{D \times D}$ .  $m_t^1$  denotes memory cell of LSTM1.  $x_t = \text{Concat}(h_{t-1}^2, \bar{v}, W_e y_{t-1})$  at time  $t$ , in which  $h_{t-1}^2 \in \mathbb{R}^D$  is hidden state of LSTM2,  $W_e \in \mathbb{R}^{D \times D_w}$  is word embedding matrix,  $D_w$  is vocabulary size.  $\bar{v} = \frac{1}{N} \sum_{i=1}^N v_i$  is global representation of the image.

### 3.3. Objective and Learning

Our loss function  $\ell$  is defined as a combination of two terms, a word-level loss  $\ell_{XE}$  and a semantic relation classification loss  $\ell_R$ , i.e.,  $\ell = \ell_{XE} + \alpha \ell_R$ , where  $\alpha$  is a parameter. **1)** The word-level loss is a cross-entropy (XE) loss, i.e.,

$$\ell_{XE}(\theta) = - \sum_{t=1}^T \log(p_\theta(y_t | y_{1:t-1})). \quad (10)$$

**2)** The semantic relation classification loss is to encourage the model to learn embeddings with weakly prior information. Like [21], we use a multi-label classification loss as follows,

$$\ell_R = \log(1 + \sum_{p \in \Omega_{neg}} e^{\hat{r}_{ij}^p}) + \log(1 + \sum_{q \in \Omega_{pos}} e^{-\hat{r}_{ij}^q}) \quad (11)$$

where  $\hat{r}_{ij}^p$  is score for one class output. The set  $\Omega_{neg}$  represents two objects  $o_i$  and  $o_j$  without a certain relation  $t$  (i.e.,  $r_{ij}^t$ ), and the set  $\Omega_{pos}$  represents two objects having a certain relation. **3)** We further optimize our model with self-critical sequence training (SCST) [22] as follows,

$$\nabla_\theta \ell(\theta) = -E_{w^s \sim p_\theta} [(r(w^s) - r(w^g)) \nabla_\theta \log p_\theta(w^s)] \quad (12)$$

where  $w^s$  and  $w^g$  denote the sampled and greedily decoded paragraph.  $r(\cdot)$  denotes a reward from metrics and  $p_\theta$  represents DualRel model.

**Table 1.** Results comparison with state-of-the-art models.

Model	B@1	B@2	B@3	B@4	M	C
Regions-Hierarchical [1]	41.90	24.11	14.23	8.69	15.95	13.52
RTT-GAN [5]	42.06	25.35	14.92	9.21	18.39	20.36
DAM [7]	35.00	20.20	11.70	6.60	13.90	17.30
SCST [8]	43.54	27.44	17.33	10.58	17.86	30.63
DCPG-VAE [6]	42.38	25.52	15.15	9.43	18.62	20.93
TOMS [27]	43.10	25.80	14.30	8.40	18.60	20.80
CAE-LSTM [11]	-	-	-	9.67	<b>18.82</b>	25.15
DHPV [9]	43.35	26.73	16.92	10.99	17.02	22.47
CVAP [10]	42.01	25.86	15.33	9.26	16.83	21.12
CRL [12]	43.12	27.03	16.72	9.95	17.42	31.47
Dual-CNN [15]	41.60	24.40	14.30	8.60	15.60	17.40
VREN [17]	41.94	24.99	15.01	9.38	17.40	14.71
IMAP [14]	44.45	27.93	17.14	10.29	17.36	24.07
S2TD [16]	44.47	27.38	16.87	10.17	17.64	24.33
OR-ATT [18]	44.55	28.54	18.19	11.18	17.97	33.12
Our DualRel	<b>45.30</b>	<b>28.91</b>	<b>18.46</b>	<b>11.30</b>	17.86	<b>34.02</b>

## 4. EXPERIMENT

### 4.1. Dataset, Metrics and Settings

**Dataset and Metrics.** Following previous works, we conduct experiments on Stanford benchmark dataset [1]. The dataset contains 14575/2487/2489 pairs for training/validation/test. There are on average 67.5 words in each paragraph with 5.7 sentences. Popular evaluation metrics, including BLEU@{1, 2, 3, 4} [23], METEOR [24] and CIDEr [25] are used. Recently proposed F metrics of BERTScore [26] are also used.

**Settings.** We use pretrained Faster R-CNN [19] to extract  $N$  of 30 object regions. The vocabulary size  $D_w$  equals 4963. We obtain  $D_r$  of 201 class of semantic relations and keep  $D_c$  of 500 objects to classify semantic relations. Besides, we set dimension  $D$  be 512. For training, first, we use the total loss in the first 25 epochs and use only word-level loss for next 5 epochs. The learning rate is initialized as  $2 \times 10^{-4}$  and decayed by 0.8 for every three epochs. Based on the validation set, we set hyper-parameter  $\alpha$  to be 0.3. Second, we adopt SCST configuration to train 40 epochs. The learning rate is initialized as  $2 \times 10^{-5}$  and decayed by 0.8 for every three epochs. We use the batch size of 10 and Adam optimizer.

### 4.2. Baselines and Main Results

**Baselines.** We compare our DualRel model with the following baselines, including Regions-Hierarchical [1], RTT-GAN [5], DAM [7], SCST [8], DCPG-VAE [6], TMOS [27], CAE-LSTM [11], DHPV [9], CVAP [10], CRL [12], Dual-CNN [15], VREN [17], IMAP [14], S2TD [16] and OR-ATT [18].

**Results.** Table 1 reports the performance of all methods. We observe that our DualRel method achieves the best scores in terms of B@{1-4} and CIDEr. First, our DualRel model outperforms SCST significantly on all metrics (tie in METEOR). Second, our method outperforms the recent IMAP method. Although IMAP uses a complex attention mechanism, DualRel still achieves better performance. In addition,

**Table 2.** F metrics on BERTScore of DualRel and SCST.

Model	F (idf)	F (rescale)	F (rescale & idf)
SCST	83.85	35.38	16.35
DualRel	<b>84.37</b> (+0.52)	<b>36.54</b> (+1.16)	<b>16.96</b> (+0.61)

**Table 3.** Ablation study on our DualRel model.

Model	B@1	B@2	B@3	B@4	M	C
DualRel	<b>45.30</b>	<b>28.91</b>	<b>18.46</b>	<b>11.30</b>	<b>17.86</b>	<b>34.02</b>
w/o Spatial Rel Encoder	44.56	28.35	18.28	11.11	17.66	33.13
w/o Semantic Rel Encoder	44.33	18.21	18.13	10.88	17.49	32.29
w/o Rel-aware Interaction	43.26	27.29	17.58	10.55	17.62	32.63

our DualRel achieves 45.30, 18.46 and 34.02 on B@1, B@2, and CIDEr scores, which outperforms OR-ATT by 1.7%, 1.5% and 2.7%. OR-ATT only exploits coarse-grained relations among objects and simply utilizes a weighted fusion of object regional features and relation embeddings. In contrast, our DualRel explicitly captures spatial and semantic relations and effectively fuses these relations embeddings through relation-aware attention interaction. Thus, our DualRel generally can generate better paragraphs. In addition, the comparison performance evaluated on recently proposed F metrics of BERTScore are reported in Table 2. We observe that DualRel outperforms SCST model in all BERTScore metrics.


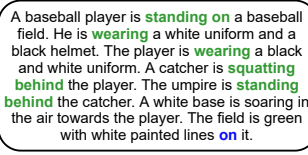
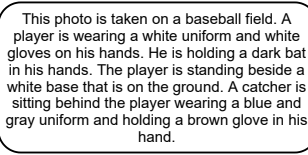

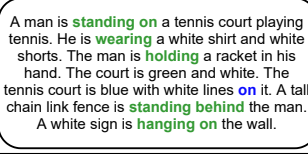
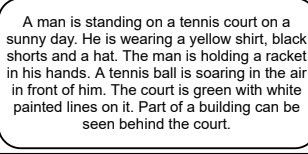
### 4.3. Ablation Studies

**On Components.** We provide ablation studies by removing key components. The experimental results are reported in Table 3. The full DualRel model achieves the best performance compared with the other variants. Without using spatial relation, the variant of Dual model cannot understand spatial relations among objects, resulting in lower metric scorers. The semantic relation is more outstanding compared with spatial relation, so the corresponding scores drops more dramatically if removed. The relation-aware interaction plays the most important role in our full Dual-Rel model. The scores drops the most if the relation-aware interaction is not used.

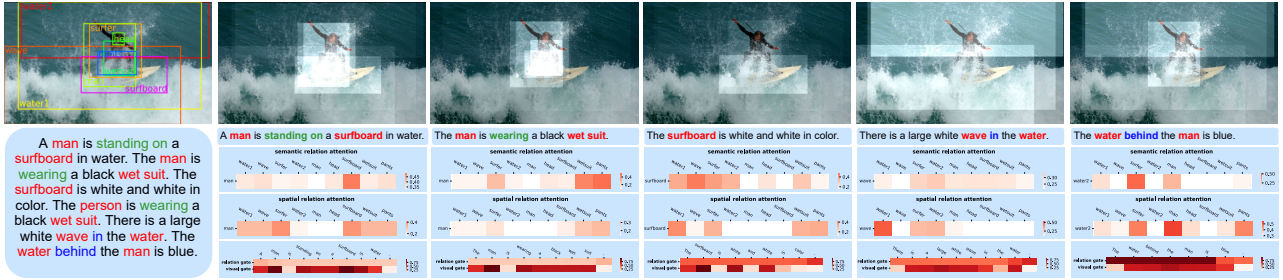
**On Parameter  $\alpha$ .** We train the DualRel without SCST on the validation set. Grid search is performed on the parameter  $\alpha$  in our loss function. The model performance of  $\alpha$  from 0 to 1 with an increment 0.1 on CIDEr score is {18.72, 20.01, 19.63, **20.6**, 19.32, 19.72, 19.84, 19.47, 19.71, 19.37, 19.29}. Without the semantic relation loss, i.e.,  $\alpha = 0$ , the model performs the worst. The highest score is achieved at 0.3 of  $\alpha$ . With the increase of semantic relation loss, the model tends to learn the bias which reduces the model performance.

### 4.4. Qualitative Analysis

Fig. 3 shows two paragraph examples generated by our DualRel and SCST. Compared with SCST, our DualRel produces paragraphs that contain more accurate and detailed relations among objects. For the first image, SCST generates wrong and even unreasonable relations like ‘baseball-

SCST	DualRel (Ours)	Ground Truth
 <p>A baseball player is <b>standing on</b> a baseball field. He is <b>wearing</b> a white uniform. The player is <b>holding</b> a bat in his hands. The catcher is <b>wearing</b> a red helmet. The <b>baseball is wearing black pants</b>. The umpire is <b>standing behind</b> the player. The field is green and white. The <b>dirt is a brown dirt</b>. The field has green grass <b>on</b> it.</p>	 <p>A baseball player is <b>standing on</b> a baseball field. He is <b>wearing</b> a white uniform and a black helmet. The player is <b>wearing</b> a black and white uniform. A catcher is <b>squatting behind</b> the player. The umpire is <b>standing behind</b> the catcher. A white base is soaring in the air towards the player. The field is green with white painted lines <b>on</b> it.</p>	 <p>This photo is taken on a baseball field. A player is wearing a white uniform and white gloves on his hands. He is holding a dark bat in his hands. The player is standing beside a white base that is on the ground. A catcher is sitting behind the player wearing a blue and gray uniform and holding a brown glove in his hand.</p>
 <p>A man is playing tennis <b>on</b> a tennis court. He is <b>wearing</b> a white shirt and white shorts. The man is <b>holding</b> a racket in his hands. <b>The man is wearing a black shirt. The court is blue and white. The court is green and white.</b> There is a large white wall <b>behind</b> the man. The tennis court is white. The wall is white and white.</p>	 <p>A man is <b>standing on</b> a tennis court playing tennis. He is <b>wearing</b> a white shirt and white shorts. The man is <b>holding</b> a racket in his hand. The court is green and white. The tennis court is blue with white lines <b>on</b> it. A tall chain link fence is <b>standing behind</b> the man. A white sign is <b>hanging on</b> the wall.</p>	 <p>A man is standing on a tennis court on a sunny day. He is wearing a yellow shirt, black shorts and a hat. The man is holding a racket in his hands. A tennis ball is soaring in the air in front of him. The court is green with white painted lines on it. Part of a building can be seen behind the court.</p>

**Fig. 3.** Paragraph examples generated by SCST and our DualRel model. Spatial relations (in blue) and semantic relations (in green) are highlighted. The relations (in red) stand for incorrect sentences are also shown.



**Fig. 4.** Visualization of relation-aware attentions and two gates during paragraph generation. For clarity, we only show the relation attention of main object in a sentence, such as ‘man’ in the first sentence and ‘surfboard’ in the third sentence.

wearing-pants’, but fails to capture the relation ‘catcher-squatting behind-player’. In contrast, our model can capture relations, the ‘lines-on-court’ and ‘sign-hanging on-wall’ for the second image. In addition, SCST may produce failure sentences, like ‘The dirt is brown dirt’ in the first example. But our DualRel model does not make similar mistakes.

#### 4.5. Visualization

**Attention Visualization.** We visualize the attention distribution, including both regional attention and relation attention, when generating multiple sentences of a paragraph. The attention visualization is shown in Fig. 4. From the object regional attention, we observe that DualRel is able to focus on the correct image regions when generating the corresponding sentences. In addition, from the results of spatial and semantic relation attention, our model can also capture the pairs of objects that probably have relations with the corresponding sentences. For example, when generating the sentence ‘A man is standing on a surfboard in water.’, DualRel captures the relation between ‘man’ and ‘surfboard’. Moreover, DualRel captures another relation between ‘man’ and ‘wet suit’.

**Gate Visualization.** We visualize outputs of relation gate and visual gate at different time steps. The gate visualization is shown in Fig. 4. The results show that our gate mechanisms play an important role during paragraph generation. For example, when generating the sentence ‘The man is wearing a black wet suit.’, the output score of the relation gate is always

**Table 4.** Human evaluation on generated paragraphs.

Model	Rich	Coh	Fin	Div	Relv
SCST	26.7	20.0	36.7	32.2	33.3
DualRel w/o $\ell_R$	13.3	23.3	20.0	30.8	23.3
DualRel	60.0	56.7	43.3	37.0	43.3

less than 0.5, which means the semantic relation should be given. When generating the sentence ‘The water behind the man is blue.’, the output score of the relation gate is always close to 1. In addition, when predicting objects and relations words such as ‘man’, ‘surfboard’ and ‘wearing’, the visual gate outputs a score close to 1. In contrast, when generating other words like ‘is’ and ‘the’, the visual gate close to 0.

#### 4.6. Human Evaluation and Relation Statistics

**Human Evaluation.** For human evaluation on generated paragraphs, we design five factors. 1) **Rich** (i.e., Relation-richness) means that a paragraph includes as many relations as possible. 2) **Coh** (i.e., Coherence) means that a paragraph is logical and well organized. 3) **Fin** (i.e., Fine-grain) means that a paragraph is detailed and includes as much content as possible in the image. 4) **Div** (i.e., Diversity) means that sentences in a paragraph have different descriptions, i.e., few repetition on sentences. 5) **Relv** (i.e., Relevance) means that a paragraph only contains content within an image.

We then randomly sample 50 images and the correspond-



**Table 5.** Statistics of relations in generated paragraphs.

Model	Spatial	Semantic	Total	Average
SCST	8239	18416	26655	10.71
Our DualRel	14585	26023	40608	16.31
Ground Truth	12987	22157	35144	14.12

ing paragraphs generated by SCST, DualRel w/o  $l_R$  and full DualRel from the testing set. These pairs of images and paragraphs are randomly assigned to 30 volunteers. They are asked to choose the paragraphs most fit for five given factors. The statistics of their proportions are reported in Table 4. Compared with those generated by SCST and DualRel w/o  $l_R$ , paragraphs generated by DualRel are generally with highest proportions. Among the five factors, relation-richness has the highest proportion of 60.0% by DualRel.

**Relation Statistics.** To quantitatively show the relation-richness, we count the number of semantic and spatial relations in paragraphs generated by our DualRel and SCST. The testing set contains 2489 images and the corresponding paragraphs. The statistics of relations with their total and average are also reported in Table 5. For each paragraph, on average, our DualRel model involves 16.31 relations, while SCST contains only 10.71 relations. In addition, our DualRel even produce more relations than those in ground truth paragraphs.

## 5. CONCLUSION

In this paper, we propose a novel model (i.e., DualRel) for image paragraph captioning. Our DualRel captures the semantic and spatial relations for effectively generating detailed paragraphs for given images. A relation embedding module and a relation-aware interaction module are designed. Extensive experiments on Stanford benchmark dataset show that our DualRel model outperforms strong baselines. In the future, we will extend our model with transformer-based methods.

## 6. ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303300 and Subject II under Grant 2019YFF0303302, in part by the National Natural Science Foundation of China under Grants 61922015, 61906018, 62076032, 62192784 and U19B2036, in part by Beijing Natural Science Foundation Project No. Z200002, and in part by the Fundamental Research Funds for the Central Universities under Grant 2021RC36.

## 7. REFERENCES

- [1] Jonathan Krause et al., “A hierarchical approach for generating descriptive image paragraphs,” in *CVPR*, 2017, pp. 317–325.
- [2] Oriol Vinyals et al., “Show and tell: A neural image caption generator,” in *CVPR*, 2015, pp. 3156–3164.
- [3] Peter Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in *CVPR*, 2018, pp. 6077–6086.
- [4] Marcella Cornia et al., “Meshed-memory transformer for image captioning,” in *CVPR*, 2020, pp. 10578–10587.
- [5] Xiaodan Liang et al., “Recurrent topic-transition gan for visual paragraph generation,” in *ICCV*, 2017, pp. 3362–3371.
- [6] Moitrey Chatterjee and Alexander G Schwing, “Diverse and coherent paragraph generation from images,” in *ECCV*, 2018.
- [7] Ziwei Wang et al., “Look deeper see richer: Depth-aware image paragraph captioning,” in *ACM MM*, 2018, pp. 672–680.
- [8] Luke Melas-Kyriazi, Alexander M Rush, and George Han, “Training for diversity in image paragraph captioning,” in *EMNLP*, 2018, pp. 757–761.
- [9] Siying Wu et al., “Densely supervised hierarchical policy-value network for image paragraph generation,” in *IJCAI*, 2019, pp. 975–981.
- [10] Zheng-Jun Zha et al., “Context-aware visual policy network for fine-grained image captioning,” *TPAMI*, vol. 44, no. 2, pp. 710–722, 2022.
- [11] Jing Wang et al., “Convolutional auto-encoding of sentence topics for image paragraph generation,” in *IJCAI*, 2019.
- [12] Yadan Luo et al., “Curiosity-driven reinforcement learning for diverse visual paragraph generation,” in *ACM MM*, 2019.
- [13] Xu Yang et al., “Hierarchical scene graph encoder-decoder for image paragraph captioning,” in *ACM MM*, 2020.
- [14] Chunpu Xu et al., “Interactive key-value memory-augmented attention for image paragraph captioning,” in *COLING*, 2020.
- [15] Ruifan Li et al., “Dual-CNN: A convolutional language decoder for paragraph image captioning,” *Neurocomputing*, vol. 396, pp. 92–101, 2020.
- [16] Yihui Shi et al., “S2TD: A tree-structured decoder for image paragraph captioning,” in *ACM MMAsia*, 2021.
- [17] Wenbin Che et al., “Visual relationship embedding network for image paragraph generation,” *TMM*, vol. 22, no. 9, pp. 2307–2320, 2019.
- [18] Li-Chuan Yang, Chih-Yuan Yang, and Jane Yung-jen Hsu, “Object relation attention for image paragraph captioning,” *AAAI*, vol. 35, no. 4, pp. 3136–3144, 2021.
- [19] Shaoqing Ren et al., “Faster r-cnn: Towards real-time object detection with region proposal networks,” *NIPS*, 2015.
- [20] Simao Herdade et al., “Image captioning: Transforming objects into words,” in *NeurIPS*, 2019.
- [21] Ningyu Zhang et al., “Document-level relation extraction as semantic segmentation,” *IJCAI*, pp. 3999–4006, 2021.
- [22] Steven J Rennie et al., “Self-critical sequence training for image captioning,” in *CVPR*, 2017, pp. 7008–7024.
- [23] Kishore Papineni et al., “BLEU: a method for automatic evaluation of machine translation,” in *ACL*, 2002, pp. 311–318.
- [24] Satandeep Banerjee and Alon Lavie, “METEOR: An automatic metric for mt evaluation with improved correlation with human judgments,” in *ACL*, 2005, pp. 65–72.
- [25] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, “Cider: Consensus-based image description evaluation,” in *CVPR*, 2015, pp. 4566–4575.
- [26] Tianyi Zhang et al., “BERTScore: Evaluating text generation with bert,” in *ICLR*, 2020.
- [27] Yuzhao Mao et al., “Show and tell more: Topic-oriented multi-sentence image captioning,” in *IJCAI*, 2018, pp. 4258–4264.