Look Deeper See Richer: Depth-aware Image Paragraph Captioning

Ziwei Wang∗
School of Information Technology and Electrical Engineering
The University of Queensland
ziwei.wang@uq.edu.au

Yadan Luo∗
School of Information Technology and Electrical Engineering
The University of Queensland
lyadanluol@gmail.com

Yang Li
School of Information Technology and Electrical Engineering
The University of Queensland
y.li9@uq.net.au

Zi Huang
School of Information Technology and Electrical Engineering
The University of Queensland
huang@itee.uq.edu.au

Hongzhi Yin
School of Information Technology and Electrical Engineering
The University of Queensland
h.yin1@uq.edu.au

ABSTRACT
With the widespread availability of image captioning at a sentence level, how to automatically generate image paragraphs is yet well explored. Describing an image by a full paragraph involves organizing sentences orderly, coherently and diversely, inevitably leading higher complexity than by a single sentence. Existing image paragraph captioning methods give a series of sentences to represent the objects and regions of interests, where the descriptions are essentially generated by feeding the image fragments containing objects and regions into conventional image single-sentence captioning models. This strategy is difficult to generate the descriptions that guarantee the stereoscopic hierarchy and non-overlapping objects. In this paper, we propose a Depth-aware Attention Model (DAM) to generate paragraph captions for images. The depths of image areas are firstly estimated in order to discriminate objects in a range of spatial locations, which can further guide the linguistic decoder to reveal spatial relationships among objects. This model completes the paragraph in a logical and coherent manner. By incorporating the attention mechanism, the learned model swiftly shifts the sentence focus during paragraph generation, whilst avoiding verbose descriptions on a same object. Extensive quantitative experiments and the user study have been conducted on the Visual Genome dataset, which demonstrate the effectiveness and the interpretability of the proposed model.

CCS CONCEPTS
• Computing methodologies → Natural language generation; Scene understanding;

∗Co-first authors

KEYWORDS
Paragraph Captioning; Depth Estimation; Attention Mechanism.

ACM Reference Format:

1 INTRODUCTION

Figure 1: The paragraph captions for an example image generated by the state-of-the-art model[20] and the proposed DAM, and written by the human annotators, respectively.

The volume of user-generated multimedia content such as images is exploding on social websites due to rapid advancement in both software technologies and hardware devices. Understanding the multimedia content and describing it in natural language is an essential task to benefit multimedia data management, retrieval, and sharing. Image captioning, a process of generating textual descriptions of images, is now an emerging research direction in
multimedia and computer vision fields. Effective auto-captioning methods are in high demand, whose outputs are expected as the medium to represent multimedia content in a human understandable manner.

Image captioning has been widely studied in recent years. Its description form can be summarised into two categories: sentence-level captioning and paragraph captioning, where the former focuses on generating one sentence to describe an image and the latter generates a series of sentences as a paragraph to provide a more comprehensive textual representation of the image.

Most of the image captioning methods follow an encoder-decoder process. In [35], images are encoded by the Convolutional Neural Network (CNN), where the features extracted from the last fully-connected layer are fed to an Long-Short-Term Memory (LSTM) decoder to generate a sentence. Aiming at describing images in a focused and detailed manner, previous work incorporate object detection [16, 17], object attributes [36, 41, 44], object relationships [5, 21, 33, 42, 43] and attention mechanism [26, 28, 38] to enhance the recognition on the major targets and enrich the expression with critical details.

Despite the fact that image captioning has made great progress, describing an image by a single sentence could barely convey sufficient semantic content and is limited to a fixed perspective. Therefore, recent work shifts their attention to image paragraph captioning. [20] utilises Faster-RCNN [30] to detect regions of interest, whose features are passed to a word and sentence recurrent neural network to model the hierarchy of paragraphs. [22] adds additional visual attention and topic-transition GAN to learn topic transition between sentences. Nevertheless, the existing image paragraph captioning approaches limit themselves in 2D recognition, and therefore lack cognition on locative relationship between objects in realistic 3D space, which is likely to lead to a poor performance on preposition (e.g., “behind” and “in front”) and the order of sentences. Besides, object-detection based models require extra tedious and expensive human labelling. Meanwhile, significant overlaps among bounding boxes with detected objects and regions may result in redundant representations of a same object or scene appearing in multiple boxes.

In order to alleviate above issues, we propose a Depth-aware Attention model (DAM) for image paragraph captioning. To the best of our knowledge, it is the first paper to leverage depth estimation in captioning task. It reconstructs a 2D image to a 3D scene by assigning estimated depths to different areas of the image, which understands the scene embedded in the image in a more diverse way. Additionally, our framework enables image paragraph captioning to be further applied in autonomous driving or assisting for visual disability. For example, users can ask the system to narrate a scene in a specific or relative spatial location, such as “what is going on in front of the blue car?”, or “what is happening 30m away?”. To demonstrate the superior performance of DAM, a real example is provided in Figure 1.

Figure 2 gives an overview of the proposed framework. Technically, by applying a dual-stream neural network at the encoder stage, DAM learns both visual and depth representations simultaneously, which are discriminative on capturing precise interactions between objects and differentiating regions at various depths. The attention mechanism is further incorporated with double LSTM layers to capture the shift of description subject on a realistic 3D space at the decoder stage. Different from all existing image paragraph captioning methods, the proposed DAM understands the scene with hierarchy of spaces rather than simply selecting object proposals, and therefore significantly reduces the sentence redundancy. The
subtle prepositions of location and the prepositional phrases can be then successfully achieved.

The contributions in this paper are three-fold:

1) To the best of our knowledge, it is a new way of thinking to leverage scene depth in captioning task, with which the generated paragraphs showcase rich diversity, coherence and sensitivity to spatial object-object relationships.

2) Different from existing image paragraph methods, a dual-stream structure equipped with the attention mechanism is used to enrich a depth-aware image representation and enable an intelligent shift of the sentence focus, rather than simply relying on extracting overlapped object proposals.

3) Quantitative experiments and user study conducted on the Visual Genome Paragraph dataset illustrate the effectiveness and the comprehensiveness of the proposed depth-aware model.

The rest of this paper is organised as follows: Sec.2 reviews depth estimation and image caption method. Sec.3 introduces a basic depth-aware paragraph generation model and full model with spatial-depth attention. Sec.4 contains experiments, and Sec.5 discusses our proposed model.

2 RELATED WORK

2.1 Visual Captioning

2.1.1 Single Sentence Captioning. Image captioning has attracted wide attention in computer vision community in recent years. Generally, image captioning methods could be roughly divided into two trends: template based, and language based. Early template based work [9, 15, 18] treat the problem as a ranking and template retrieval task while they could hardly achieve a satisfactory performance as its compositional nature of language highly limits in terms of diversity and flexibility. The language-based models [3, 7, 13, 32, 39, 40] intend to learn the mapping from visual features to linguistic embeddings to form a with full sentence. With the rapid development of deep learning techniques, the accuracy and effectiveness of the image captioning model learned with neural networks has been powered up. In Vinyals et al. [35], the authors propose an end-to-end neural networks architecture to generate the sentence to convey visual content. The framework consists of an image encoder and a language decoder, where the encoder is a convolutional neural network (CNN) for generating the visual representation of images, and the decoder is a recurrent neural network (RNN) for caption generation. Emerging work follow the encoder-decoder structure, and add attention mechanism [2, 10, 26, 28, 38], object attribute [36, 41, 44] to enrich representations for a better semantic expression. Reinforcement learning[25, 29, 31] has also been utilised in image captioning as the optimising strategy.

2.1.2 Paragraph Captioning. Even though describing image content by a sentence has made great progress, massive underlying information, such as the background and geometric structure conveyed in the image would be missed if only summarising them with a single sentence. As an extension of image captioning, image paragraph captioning has been proposed in [20] to tackle the above issue of sentence-level captioning. Specifically, [20] detects object regions to preserve major objectness and uses hierarchical recurrent neural network to capture hierarchy of sentences and words. Liang et al.[22] add an attention mechanism and recurrent topic-transition generative adversarial network to generate personalised and reasonable paragraph. However, existing image paragraph captioning methods highly rely on object detection model trained with RGB data, which inevitably makes sentences repeatedly describe a same object-object relationship and lack of discrimination on relative spatial relationship between objects.

2.2 Depth Estimation

Our work is inspired by a recent work in [4] that exploits depth estimation for object detection and semantic segmentation. Among related literature on depth estimation, the deep neural networks learned in both supervised and unsupervised settings have been fully investigated. Supervised approaches [8, 24] require the ground-truth calculated with a fixed camera baseline distance and focal length during the training procedure, and predict the depth map given a monocular colour image during the testing. Most of unsupervised methods [11, 12] formulate disparity map prediction as an image reconstruction problem at training time, and then use monocular image to predict disparity map. Since our strategy is agnostic to depth estimation methods, we adopt the recent state-of-the-art unsupervised monocular depth estimation with left-right consistency method [12] to predict depth map for image captioning.

3 THE PROPOSED APPROACH

In this section, we introduce the proposed Depth-aware Attention Model (DAM) for image paragraph captioning. As shown in Figure 2, given an image, the associated depth map is firstly generated from its single monocular image. Secondly, a dual-stream convolutional neural network extracts the RGB-based image visual feature and the depth map deep representation simultaneously, based on which the final representation of the original image is projected by a fully-connected layer. The generated final representation is a depth-aware deep visual feature, which is forwarded as the input of a double-layer LSTM. The attention model cooperating with the LSTM in our work is depth-aware, which considers the depth of each area when selecting the attended area and working out its corresponding description at each time point. The input of this novel attention model is the visual feature map and the depth map of the original image. The core mechanism of the model is illustrated in Figure 4.

3.1 Problem Formulation

We denote image and its corresponding depth map as I and d respectively. The objective of the proposed model is to generate a paragraph of description $S = \{S_1, ..., S_N\}$ given $I$ and $d$, where $N$ is the length of the paragraph.

3.2 Preliminary Work on Depth Estimation

The depth map $d$ can be measured or estimated given a RGB image, and the illustration could be seen in Figure 3. Currently, it is not practical to obtain measured ground truth depth map for a large variety of scenes. Thus we adopt the depth estimation method [12] to augment depth map given a single monocular image.
The task of monocular image depth estimation is formulated as follows. At testing time, given a single monocular image \( I \), the corresponding pixel-wise depth map \( d \) is generated,

\[
\hat{d} = g(I),
\]

where \( g(\cdot) \) denotes the learned projection.

The objective of this unsupervised depth estimation with left-right consistency method (DELRC) [12] is to learn a function \( f(\cdot) \) given a calibrated pair of binocular images at training time. Specifically, during training, we have access to stereo pair of left and right RGB image \( I_l \) and \( I_r \). DELRC attempts to find dense correspondence field \( d_l \), such that, when \( d_l \) is applied to \( I_l \), it reconstructs the right RGB image \( \hat{I}_r \). Similarly, by applying the learned right-to-left dense correspondence field \( d_r \) to the right RGB image \( I_r \), DELRC reconstructs left RGB image \( \hat{I}_l \).

Assuming that the images are rectified, \( d_l \) corresponds to image left-to-right disparity map, where a scalar value per pixel will be learned. Given the baseline distance \( b \) between two cameras and the camera focal length \( f \), we can then recover the depth \( \hat{d} \) from predicted disparity \( d_l \),

\[
\hat{d} = b f / d_l .
\]

Three losses are defined to train the model: the appearance matching loss, the disparity smoothness loss and the left-right disparity consistency loss.

At testing time, DELRC predicts disparity \( d_l \) given input image \( I \), while right-to-left disparity \( d_r \) is only used in training. To convert predict disparity \( d_l \) to extract depth \( \hat{d} \), we still need to know the camera baseline length and focal length. However, for the purpose of separating objects in different spatial layout in our cascade model, it is not necessary to predict the absolute depth \( \hat{d} \). Therefore, in this scenario, we rescale the disparity \( d_l \) to the range of \([0, 255]\). Then we define the rescaled disparity \( d' \) as relative depth \( \hat{d} \).

### 3.3 The Depth-aware Captioning Model

In this section, we discuss how to exploit estimated depth map \( d \) for image paragraph captioning. A basic depth-aware encoder-decoder framework is first proposed in Sec. 3.3.1, followed by the advanced Depth-aware Attention Model (DAM) in Sec. 3.3.2.

#### 3.3.1 The Depth-aware Basic Model

Most existing image captioning work follow the encoder-decoder structure [35]. Given a pair of image \( I \) and paragraph caption \( S \), the encoder-decoder maximises the following objective function directly:

\[
\theta^* = \arg \max_\theta \sum_{(S, I)} \log p(S|I; \theta),
\]

where \( \theta \) are the model parameters. We propose a basic depth-aware encoder-decoder model, which takes into account the additionally measured or augmented depth-map. This model aims to maximise the probability of the target paragraph caption by using the following formulation, given an image \( I \) and its associated depth-map \( d \):

\[
\theta^* = \arg \max_\theta \sum_{(S, I, d)} \log p(S|I, d; \theta).
\]

Applying the chain rule, the joint probability over \( S \) is calculated as follows:

\[
\log p(S|I, d) = \sum_{t=0}^{N} \log p(S_t|I, d, S_0, ..., S_{t-1})
\]

where \( N \) is the number of words in this particular paragraph example, and we also drop the model parameters for clearer illustration. In an encoder-decoder model, it is natural to model conditional probability with RNNs, where the conditional probability at each time step \( t \) is modelled as:

\[
\log p(S_t|I, d, S_0, ..., S_{t-1}) = f(h_t)
\]

where \( h_t \) is hidden state at time \( t \), which is a fixed vector that represents variable length of word sequence we condition upon up to \( t - 1 \). In this paper, we adopt the double-layer LSTM units. \( h_t \) is modelled as:

\[
h_t = \text{LSTM}(x_t, h_{t-1}, m_{t-1})
\]

\[
x_t = v_d
\]

where \( x_t \) is the input at time \( t \), \( h_{t-1} \) is the hidden state \( t - 1 \), \( m_{t-1} \) is the cell state at time \( t - 1 \), \( v_d \) is the depth-aware visual representation of the image.

For the for representation of images, we use last fully connected layer outputs of CNN as the image feature. A dual-stream CNN is designed to extract both the RGB-based global image feature \( v_g \) and a deep representation of the depth feature \( u_g \) simultaneously. \( v_d \), a fused depth-aware visual feature, is generated by the linear projection from the concatenation of \( v_g \) and \( u_g \):

\[
v_d = W_c[v_g; u_g],
\]

where \( W_c \in \mathbb{R}^{D \times 2D} \) is the projection parameter to be learned, \( D \) denotes the embedding size and \( v_g, u_g \in \mathbb{R}^{D \times 1} \)

#### 3.3.2 The Depth-aware Attention Model

Traditionally, visual attention is a mechanism to help RNN reason about where is the most salient visual region that the language model should attend to at the time step \( t \). Its input contains the image convolutional feature \( V \in \mathbb{R}^{D \times k} \) from the last convolutional layer before average pooling layer, \( k \) denotes the number of features regions. Inspired by the spatial attention model [26], similarly we feed the attention module with the visual and depth feature extracted from the last convolutional layers (i.e., \( V \) and \( U \in \mathbb{R}^{D \times k} \) at the same time. The
proposed attention mechanism is capable of inferring where and how deep the language model should be currently focusing on.

In the depth-aware attention model (Figure 4), we compute the context vector \( c_t \) in each time step \( t \) when predicting the next word probability. The context vector \( c_t \) could be defined as:

\[
c_t = g(V, U, h_t),
\]

where \( g() \) is the function of depth-aware attention, \( V = [v_1, ..., v_k] \), \( v_i \in \mathbb{R}^{D \times 1} \) is the visual feature corresponding to \( i \)-th area of the image. Similarly, \( U = [u_1, ..., u_k], u_j \in \mathbb{R}^{D \times 1} \) is the depth feature corresponding to \( i \)-th region in the image. \( h_t \) is the hidden state at time step \( t \) from the LSTM. Therefore, Equation 6 could be rewritten as:

\[
\log p(S_t|I, d, S_0, ..., S_{t-1}) = f(h_t, c_t).
\]

Given \( V \in \mathbb{R}^{D \times k}, U \in \mathbb{R}^{D \times k} \) and \( h_t \in \mathbb{R}^{D \times 1} \), we feed them through a fully-connected (FC) layer. The outputs from FC layer are forwarded through a softmax function to obtain the attention scores over \( k \) regions:

\[
z_t = w_k^T \tanh(W_{vu}[V;U] + (W_g h_t)1^T),
\]

\[
\alpha_t = \text{softmax}(z_t),
\]

where \( 1 \in \mathbb{R}^{1 \times 2k} \) is a vector filled with 1. \( W_{vu}, W_g \in \mathbb{R}^{c \times D} \), \( e \) denotes a latent embedding space for dimension reduction and \( w_k \in \mathbb{R}^{c \times 1} \) are model parameters. \( \alpha_t \in \mathbb{R}^{1 \times 2k} \) is the attention weights over the concatenation of \( V \) and \( U \). The context vector \( c_t \) can be calculated as a weighted sum:

\[
c_t = \sum_{i=1}^{k} \alpha_{ti} v_i + \sum_{i=1}^{k} \alpha_{tk+i} u_i,
\]

where the combination of \( c_t \) and \( h_t \) are forwarded through a FC layer to predict most possible next word \( S_t \) as described in equation 11.

4 EXPERIMENTS

In this section, we conduct evaluation experiments on the benchmark image paragraph annotation dataset by comparing our method with the baseline and the state-of-art models. A comprehensive user study is also conducted to discuss the performance of captioning achieved by different models from the user perspective in terms of five criteria.

4.1 Experimental Settings

4.1.1 Dataset. We conduct the experiments on the benchmark dataset - Visual Genome Paragraph Dataset [20], which is created for image paragraph generation task. The training set contains 14,575 image-paragraph pairs, the validation set has 2,487 pairs, and the testing set has 2,489 pairs. In average, each paragraph caption contains 67.5 words and 5.7 sentences, each of which consists of an average of 11.9 words.

4.1.2 Implementation Details. Given the estimated depth map, we encode the depth map to a three-channel image. We simultaneously extract RGB and depth features of each image through dual-stream convolutional neural networks (CNN), which are not sharing weights. Here we adopt the architecture of ResNet-152 [14] as encoder, on which the last two layers are removed for fine-tuning. The last convolutional map is obtained for attention model. The language model is a two-layer LSTM with the hidden size of 512. The dimensionality of the word embedding is empirically set to 1,024. We use stochastic optimisation method Adam [19] for parameter updating with the learning rate of 1e-4. The batch size is fixed at 32. All the experiments are tested on a server with 40 Intel(R) Xeon(R) E5-2690 CPUs and 2 Titan X GPUs.

4.1.3 Evaluation Metrics. We report the performance of image paragraph captioning using the COCO captioning evaluation tool [6] with following metrics: BLEU [27], METEOR [1], CIDEr [34]. BLEU is originally designed for evaluating the equality of machine translation, which is essentially the geometric mean of \( n \)-gram precision scores with brevity penalty. For example, generally speaking, BLEU-3 describes how precisely 3-gram phrases are generated by the proposed algorithm according to the groundtruth. METEOR is another machine translation oriented metric, which is the harmonic mean of precision and recall. More recently, CIDEr is proposed specifically for image descriptions evaluation, which is essentially a \( tf-idf \) weighted \( n \)-gram similarity.

4.2 Compared Methods

Sentence-Concat (Neuraltalk) [17]: This baseline samples and concatenates five sentences caption from a Neuraltalk captioning model [17], which is trained on MS-COCO captions dataset [23]. The first sentence uses beam search (beam size = 2) and the rest are sampled (greedy search). Specifically, every image is represented as 19 detected region features and 1 global feature.

Sentence-Concat (NIC) [35]: Similar to Neuraltalk Sentence-Concat, NIC model is also trained on MS-COCO sentence caption dataset. It predicts five sentences and concatenates them into a paragraph. Different from Sentence-concat (Neuraltalk), NIC model does not detect regions. Therefore the encoded image feature is simply the output from the last fully-connect layer of CNN.

Image-Flat (NIC) [35]: This model uses a flat representation for both images and language. Essentially it has a same architecture as
When the model is trained on sentence captioning data, which has limited details, these fine-grained visual cues are effective. We present our main results in Table 1. In this section, we compare DAM with baseline methods sentence-concat, template basic version, Depth-aware Basic model, to demonstrate the effectiveness of leveraging depth information through the comparisons to the above three baseline methods. Comparing to Sentence-Concat models, Depth-aware Basic outperforms Sentence-Concat (Neuraltalk) by 133% and 68% on CIDEr and BLEU-4 respectively. Depth-aware Basic has similar architecture to Image-Flat model but with a depth CNN. It improves Image-Flat especially on BLEU-4 and CIDEr, which indicates that the proposed depth-aware model is able to learn better long phrases than vanilla NIC model.

DAM achieves comparable results in terms of METEOR and BLEU-4 and performs the best in CIDEr and BLEU-1,2,3 among all compared methods. Comparing to the Region-Hierarchical model, DAM makes a 53.1% relative improvements on CIDEr from 11.3 to 17.3. DAM also outperforms Depth-aware Basic in all metrics except BLEU-4.

Finally, we note that, during the training, the model learns the length of paragraph and the number of sentences in very early stage. However, it takes a much longer time to learn how to generate a paragraph with diversity.

### 4.3 Quantitative Analysis of DAM with the state-of-the-arts

We present our main results in Table 1. In this section, we compare the proposed depth-aware attention model (DAM) with its basic version, Depth-aware Basic model, to demonstrate the effectiveness of the depth-aware attention mechanism. We further compare DAM with baseline methods sentence-concat, template based, image-flat and state-of-the-art hierarchical model [20] to have a comprehensive performance study.

Two Sentence-Concat models are tested in our experiments. Both methods perform poorly among all baselines, indicating that the pre-trained sentence captioning models are not suitable for paragraph captioning task. Comparing Neuraltalk and NIC Sentence-Concat methods, we can see that Neuraltalk method performs slightly better than NIC does. The reason is that Neuraltalk detects regions of interest and use them as visual cues for captioning generating. When the model is trained on sentence captioning data, which has limited details, these fine-grained visual cues are effective.

Image-Flat adopts the same model as Sentence-Concat(NIC). However it outperforms Sentence-Concat model in a great margin, especially on BLEU-3,4 and CIDEr. The main difference is that Image-Flat is trained on the task of paragraph generation, so that it shows the importance of using appropriate paragraph training data.

Depth-aware Basic model demonstrates the effectiveness of leveraging depth information through the comparisons to the above three baseline methods. Comparing to Sentence-Concat models, Depth-aware Basic outperforms Sentence-Concat (Neuraltalk) by

### Table 1: Performance comparison using BLEU, METEOR and CIDEr on Visual Genome Paragraph dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence-Concat (Neuraltalk[17])</td>
<td>12.1</td>
<td>6.8</td>
<td>31.1</td>
<td>15.1</td>
<td>7.6</td>
<td>4.0</td>
</tr>
<tr>
<td>Sentence-Concat (NIC[35])</td>
<td>9.3</td>
<td>7.1</td>
<td>22.3</td>
<td>10.7</td>
<td>4.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Image-Flat (NIC[35])</td>
<td>13.4</td>
<td>14.7</td>
<td>34.8</td>
<td>19.4</td>
<td>10.9</td>
<td>6.0</td>
</tr>
<tr>
<td>Region-Hierarchical ([20])</td>
<td><strong>14.8</strong></td>
<td>11.3</td>
<td>34.2</td>
<td>17.4</td>
<td>8.5</td>
<td>4.0</td>
</tr>
<tr>
<td>Depth-aware Basic</td>
<td>13.4</td>
<td>15.9</td>
<td>34.3</td>
<td>19.9</td>
<td>11.6</td>
<td>6.7</td>
</tr>
<tr>
<td>Depth-aware Attention Model (DAM)</td>
<td>13.9</td>
<td><strong>17.3</strong></td>
<td><strong>35.0</strong></td>
<td><strong>20.2</strong></td>
<td><strong>11.7</strong></td>
<td><strong>6.6</strong></td>
</tr>
<tr>
<td>Human</td>
<td>19.2</td>
<td>28.6</td>
<td>42.9</td>
<td>25.7</td>
<td>15.6</td>
<td>9.7</td>
</tr>
</tbody>
</table>
The depth-aware attention mechanism learns where to focus on \( \alpha \) to the image size \((224 \times 224)\) from Visual Genome Paragraph testing split. We simply interpolate the image without paying much attention to depth. However, while predicting the last word "to", DAM decides to attend to surrounding area in the depth map to find the next object to describe. It does not rely much on RGB image. We can observe that DAM model learns the alignment between the attended area and the word that reflects the human intuition especially during object describing and focus transition.

**4.6 User Study**

To better understand how satisfactory are the sentences generated from different methods, we conduct an user study to compare our DAM against the state-of-the-art approach, and even ground-truth. A total number of 30 evaluators (15 females and 15 males) from different education backgrounds, including computer science (10), mechanical engineering (6), arts (6), business (8), are invited and a subset of 500 images is randomly selected from testing set of the Visual Genome Paragraph subset and the corresponding paragraph in a more coherent order.
generated paragraphs from DAM, Region-Hierarchical [20] and human-written respectively, for the subjective evaluation.

The user study is defined as following inspired by [37, 41]. We separate evaluators into two groups, and give them a detailed instruction on criteria and a simple example before they starting evaluation. We show the first group full machine generated paragraphs by different models and human-annotated paragraph without giving any hints of the method names. They are then required to evaluate the quality of each paragraph by giving a score between 1 and 3 (larger is better) according to the four pre-defined criteria. For each criterion, we use the average score to reflect the quality of a method. For example, in terms of "Coherence", the average score of all paragraphs generated by DAM is 1.9. In contrast, we show the second group one pair of paragraphs at a time (an automatically generated paragraph and the ground truth) and require the evaluators to determine which one is generated by a system and which one is generated by the human beings. Based on the statistics of the evaluators’ responses, we calculate a metric of Turing, which is defined as the percentage of captions that pass the Turing Test. For better visualisation, we rescale all the metrics to 0-1 in Figure 7.

Quality control is a good mechanism for enhancing the quality of evaluation and for filtering inconsistent and unreliable marking. We control the evaluation by inserting an assessed image-paragraph pair inside a 30-image batch, and monitor the evaluator’s consistency on it. Evaluators that fail to give a same mark on this test image-paragraph pair are not able to submit the evaluation and have to remark the image batch.

(1) **Spaciousness**: whether the paragraph uses correct locative prepositional phrases to describe objects;
(2) **Coherence**: judge the logic and readability of the sentences;
(3) **Relevance**: whether the paragraph contains the major and critical objects/actions/relationships in the image;
(4) **Diversity**: judge the richness and diversity of the representation of relationships and objects;
(5) **Turing**: whether the generated paragraph passes the Turing Test.

The experimental results are shown in Figure 7, from which it is observed that the proposed method DAM generates better quality paragraphs with respect to spaciousness, coherence, relevance and diversity compared with the state-of-the-art method [20]. Besides, the paragraph captioning of the proposed approach passes more cases in Turing test, which proves its expression is more natural and flexible. Admittedly, there still exists gap between human-written paragraphs and the machine-generated paragraphs, since minor small grammar errors laying in the generated paragraph are easy to tell while the human-written descriptions preserve good language conventions, potentially causing most of cases fail in Turing Test.

![Figure 6: Visualisation of depth-aware attention of DAM.](image)

![Figure 7: The evaluation result of user study on the Visual Genome Paragraph dataset with respect to five major metrics. All the metrics are rescaled to 0-1 for better visualisation. (Better viewed in colour)](image)

### 5 CONCLUSION AND FUTURE WORK

In this paper, we propose a deep depth-aware framework to strengthen image paragraph captioning by enriching raw data with extra geometric information. The estimated depth map assists the model to recognise the subtle and locative relationships between objects and generates well-organised paragraph in a coherent way. An attention mechanism is further incorporated to predict coherent subject transition between consecutive sentences during the paragraph captioning. It improves the diversity of expressions in the meantime. Quantitative experiments have shown the superiority of the proposed method compared with the state-of-the-arts. Since the proposed algorithm is agnostic to any depth estimation method, we plan to incorporate a more appropriate depth estimation module to see a better performance on image paragraph captioning task in the near future.

### 6 ACKNOWLEDGEMENT

This work is partially supported by ARC FT130101530 and NSFC No. 61628206.