Region-Object Relation-Aware Dense Captioning via Transformer

Shao, Zhuang; Han, Jungong; Marnerides, Demetris; Debattista, Kurt

Published in:
IEEE Transactions on Neural Networks and Learning Systems

DOI:
10.1109/TNNLS.2022.3152990

Publication date:
2025

Citation for published version (APA):
Shao, Z., Han, J., Marnerides, D., & Debattista, K. (2025). Region-Object Relation-Aware Dense Captioning via Transformer. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(3), 4184-4195. https://doi.org/10.1109/TNNLS.2022.3152990

Region-object Relation-aware Dense Captioning via Transformer

Zhuang Shao, Jungong Han, Demetris Marnerides, Kurt Debattista

***Abstract*—Dense captioning provides detailed captions of complex visual scenes. While a number of successes have been achieved in recent years, there are still two broad limitations: 1) Most existing methods adopt an encoder-decoder framework, where the contextual information is sequentially encoded using Long Short-Term Memory (LSTM). However, the forget gate mechanism of LSTM makes it vulnerable when dealing with a long sequence; 2)The vast majority of prior arts consider Regions of Interests (RoIs) equally important, thus failing to focus on more informative regions. The consequence is that the generated captions cannot highlight important contents of the image, which does not seem natural. To overcome these limitations, in this paper, we propose a novel end-to-end transformer-based dense image captioning architecture, termed Transformer-based Dense Captioner (TDC). TDC learns the mapping between images and their dense captions via a Transformer, prioritising more informative regions. To this end, we present a novel unit, named Region-Object Correlation Score Unit (ROCSU), to measure the importance of each region, where the relationships between detected objects and the region, alongside the confidence scores of detected objects within the region, are taken into account. Extensive experimental results and ablation studies on the standard dense-captioning datasets demonstrate the superiority of the proposed method to the state-of-the-art methods.**

***Index Terms*—Dense Image Captioning, Transformer-based Dense Image Captioner, Region-Object correlation score unit**

I. Introduction

Dense captioning has gained significant attention from both the engineering and research communities recently. On the one hand, it facilitates important practical applications [1], such as human-robot interaction [2], navigation for the blind, object detection [3] [4] or segmentation [5] and image-text retrieval [6] [7]. On the other hand, it poses substantial challenges to both computer vision and natural language processing research communities. Its complexity in generating richer and more detailed descriptions for local regions, compared to image captioning, hastens the emergence of more advanced captioning techniques.

Dense captioning stems from image captioning, and recent years have witnessed a rapid development of image captioning techniques. Many of these methods are based on encoder-decoder frameworks and inspired by the successful transfer of sequence to sequence training used for machine translation [8]. Broadly, image features are first extracted by a Convolutional Neural Network (CNN) as an encoder, and then fed into an RNN-based decoder that outputs the corresponding captions. However, such a captioning mechanism based on encoder-decoder frameworks fails to focus on areas that may be worthy of more attention at the training stage. To address this issue, many updated methods have been proposed. For example, [9] proposed aligned high-level information while [10]–[13] resorted to different forms of attention to aid guidance during training.

(a) (b)

Fig. 1. (a) An example of the RoI description created by the LSTM method COCG [14]. (b) The corresponding object detection results as context to guide the dense captioning.

Dense captioning is beyond image captioning due to the need to provide richer and more detailed descriptions for a given image. [15] took the initiative to develop a Fully Convolutional Localization Network (FCLN) for the dense captioning task, in which Regions of Interests (RoIs) are localized before being described. Afterwards, many follow-ups appeared, which can be generally categorized into two classes depending on whether the contextual information encoded in the model is used. At the early stage, the architecture was composed of a Faster R-CNN [16] module to detect RoIs and describe them with a Long Short-Term Memory (LSTM) [17], which was an advanced variant of Recurrent Neural Network (RNN). Unfortunately, this kind of framework only considered the RoIs but ignored possible contextual information that can be leveraged to improve training. To address this problem, [18] proposed to integrate the RoI features with image features as a global context to build up a joint and contextual fusion before captioning via an LSTM. However, the proposed global context seems too coarse, and there have been several methods that explored fine-grained contexts. For instance, [19] proposed a non-local similarity graph for the feature interaction between the target RoI and its neighboring RoIs. Also, supported by

Manuscript received xxx, xxx; revised xxx, xxx and xxx, xxx; accepted xxx, xxx. (Corresponding author: Jungong Han). This research was supported by the funds of China Scholarship Council under Grant No. 201909120012.

Zhuang Shao is with Warwick Manufacturing Group, University of Warwick, CV4 7AL, UK (e-mail: ZhuangShao@warwick.ac.uk).

Jungong Han is with the Department of Computer Science, Aberystwyth University, SY23 3DB, UK (e-mail: jungonghan77@gmail.com).

Demetris Marnerides is with Warwick Manufacturing Group, University of Warwick, CV4 7AL, UK (e-mail: dmarnerides@gmail.com).

Kurt Debattista is with Warwick Manufacturing Group, University of Warwick, CV4 7AL, UK (e-mail: K.Debattista@warwick.ac.uk).

data statistics, [14] revealed the close relationship between RoIs and detected objects via object detection, thus resulting in an architecture with contextual information considered.

Despite the preliminary success of the aforementioned methods, dense image captioning can, arguably, be considered still in its infancy. We believe a number of limitations still exist, two of which are critical. Firstly, LSTMs, as the dominant structures for the methods mentioned above, suffer from the nature of the forget gate mechanism: forgotten information after a sequence cannot be avoided, especially when the inputted sequence is long. In state-of-the-art methods, if the contextual information is encoded by an LSTM, and with time rolling, the initial object would be "forgotten" and thus it weakens the guidance function of context especially when there are interactions of multiple people and multiple objects. Hence, the training model may fail to "oversee" the objects so that it cannot guide the captioning process properly. As a result, this kind of gap often gives rise to the missing of descriptive objects, as illustrated in Fig. 1. Obviously, the object detection results as guided context for dense captioning are in good conditions, with accurate localizations and high confidence on the right. However, on the left, due to the aforementioned deficit of LSTM, the output caption does not include all three birds and a chair in its answer. Instead, it generates only a bird on a building, but forgets the other two birds and the chair.

Secondly, in the previous methods, e.g. [14], all the RoIs are treated with equal weights during training. However, in the real world, the useful information carried by each RoI can be hugely different. Also, the detection confidence scores of objects within and around the region may vary considerably from region to region. These all imply that the regions should be treated differently during model training. As shown in Fig. 2, it may make more sense if assigning larger weights to the RoIs with more information at the training stage. Concretely, in this example, on the left are two RoIs detected, but apparently, they have different IoUs with the overall object bounding boxes illustrated on the right. According to the descriptive languages of these RoIs, it is obvious that the caption of the one in red with a higher IoU with the objects on the right. Also, it contains much more information in its ground truth since its description reveals the theme of the image. In contrast, the RoI in yellow contains too detailed information and this kind of information is even far difficult for the human being to observe, not to mention attain it by machine learning. Inspired by the common exam strategy that a student should focus more on the basic questions accounting for a large proportion of marks, rather than concentrating on difficult ones, we hold a view that the informative regions deserve more priorities.

To alleviate the first issue, we propose a novel end-to-end dense captioning framework based on Transformer [20], which is currently popular in a great variety of computer vision tasks, termed **T**ransformer-based **D**ense **C**aptioner (**TDC**), to overcome the limitations of the forget mechanism of LSTM when encoding and decoding visual and language information. Fig. 3 gives an overview of TDC. Particularly, inspired by [14], we compose both object detection information and holistic image features as context. Along with the detected RoIs from Faster R-CNN Region Proposal Network (RPN) and contextual information, the visual information is projected into a visual representation by applying a dot product between them. The same operation is implemented on language information as well. At the decoding phase, a probability distribution for captions of detected RoIs is learnt by cross-modality attention of both visual and language encoding results. During encoding and decoding, all of the input vectors are aligned and computed together, hence it can overcome the forget problem.

In order to address the second limitation, we propose a module, which allocates weights for the language loss of each region at each step of training. The underlying assumption is that the regions comprising more objects with high detection confidence scores are more important, and thus, deserve priority. To this end, we propose a novel unit, which makes use of both the object detection score and the intersection of unit [21] (IoU), named **R**egion-**O**bject **C**orrelation **S**core **U**nit (ROCSU).

The major contributions of this work are summarized as:

- A novel end-to-end dense captioning framework based on the Transformer, dubbed TDC, is proposed. A distinct property of TDC is the advocate of a Transformer to capture the long-range contextual information among objects. It is clearly advantageous over LSTM that is impotent in capturing *long-range dependencies* among objects. To the best of our knowledge, this is the first work that builds up a Transformer-based architecture rather than an LSTM for the dense captioning topic.
- An RoI importance unit, named Region-Object Correlation Score Unit (ROCSU), drives the loss function to focus more on RoIs with more information. In doing so, our work, for the first time, weighs RoIs by jointly considering object-region relationships and object detection confidence scores during model training. It differs from treating each RoI equally at the training stage.
- Extensive experimental results on different challenging datasets show the superiority of the proposed method against the state-of-the-art methods.

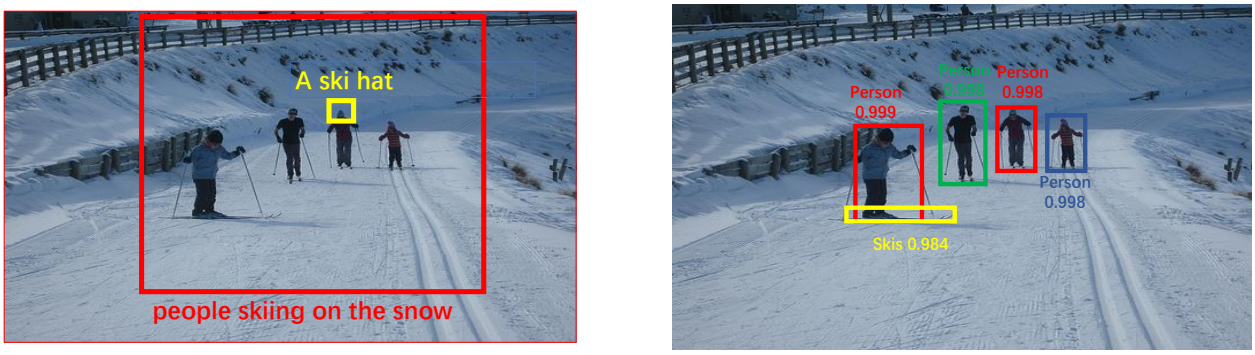
The rest of this paper is organized as follows: We discuss related work in Section II. In Section III, the proposed method is introduced in detail with a comprehensive analysis. Extensive experimental results are demonstrated in Section IV with both qualitative and quantitative analysis. Finally, we summarize this paper with a conclusion in Section V.

II. Related Work

In this section, we will review the related works from two aspects: image captioning and dense captioning.

A. Image Captioning

Earlier neural network models for image captioning [11], [22]–[24] encoded visual information using a single feature representation of the image [25] with very limited additional information. However, with the development of deep learning, more auxiliary information can be added up into a model



(a) (b)

Fig. 2. An example shows the RoIs with different IoUs should be weighted differently. (a) Two RoIs and their descriptions; (b) Object detection results.

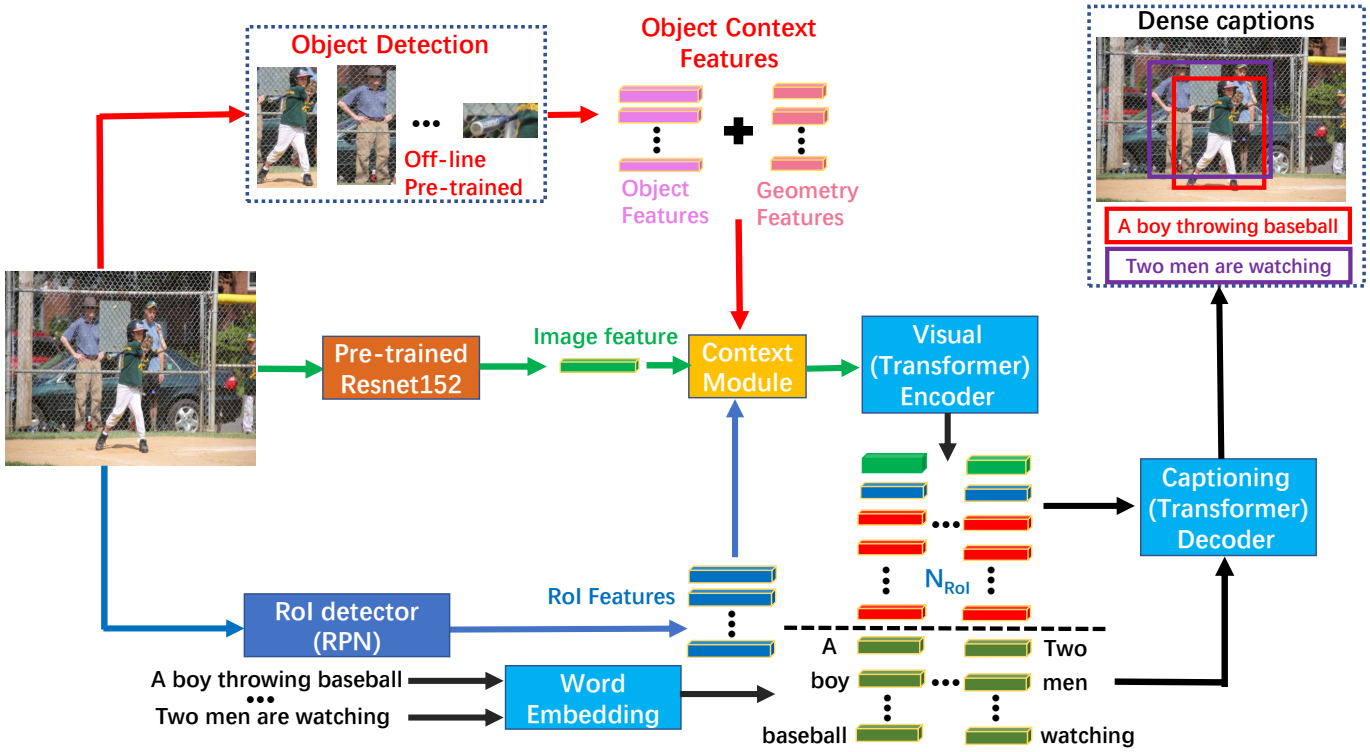


Fig. 3. The proposed TDC framework is made up of an RoI detector, context module, visual encoder and captioning decoder. Given an image, the RoI detector detects RoIs and the context module prepares contextual information generated via the pre-trained object detector for further use. After this, the visual encoder encodes visual information by attention, which gains a visual representation. Finally, after the word embeddings are conducted, visual representation and sentence information are decoded by the captioning decoder to generate dense captions for each RoI.

structure. [9] extracted region features from images with an R-CNN object detector [26] and generated separate captions for the regions as the captions of the given image. [10] proposed a method to generate image descriptions by first detecting words associated with different regions within the image [25]. In addition, [27] proposed an efficient concept learning module to get pseudo pairs.

To better focus on important parts of images and model their correspondent relations with words in captions, a series of variants of attention models have been incorporated. [12] proposed a semantic attention module, which combines the top-down and bottom-up attention together. Also, [28] involved geometric attention, which inspired [29] to develop a framework with two Graph Convolutional Networks to explore visual relationships. In recent years, with the advance of Natural Language Processing (NLP), the Transformer architecture [20] has led to significant performance improvements for various tasks. [30] proposed a Transformer-based model by extracting a single global image feature from the image as well as uniformly sampling features by dividing the image into 8x8 partitions. In the latter case, the feature vectors were fed in a sequence to the Transformer encoder [25].

B. Dense Captioning

Later on, dense captioning [15] emerged as a new task that requires an intelligent vision system to both localize and describe salient regions within an image in natural language. Existing dense captioning algorithms can be roughly categorized into two types: captioning with the guidance of contextual information and captioning without using contextual information.

1) Dense Captioning Without Context: In [15], Johnson *et al.* proposed a bilinear interpolation with a prototype of an RPN in Faster R-CNN. All the RoIs are represented by the same-size features, denoted as region features. Subsequently, they are passed through a fully-connected layer to determine if they are foreground (the descriptive region) or background. The locations of these regions are also amended at this stage via regression. At a later stage, region features are described by an LSTM language model, which is trained in an end-to-end manner.

2) Dense Captioning With Context: The work in [18] is conceptually similar to [15]. But the difference lies in that the image feature acted as the contextual information, which was fed into the captioning module together with RoIs. Despite an improved performance, the contextual information is just a kind of global and coarse information, thus leading to the failure to encode more detailed context information.

Subsequent works attempted to incorporate fine-grained context into the framework. For instance, [19] established a non-local similarity graph for the feature interaction between the target RoIs and its neighboring RoIs. Furthermore, it is noted that in [14], the authors argued that objects provide valuable cues to help locate captioning regions and generate descriptions for them via the use of data statistics. Inspired by this, the authors proposed to bring in local contextual information to guide the training of the model. To capture useful object information in an image, a novel framework for learning a complementary object context for each RoI was proposed using an LSTM. This context is derived from a concatenation of extracted object features and geometry information. The LSTM cell progressively accepts each object as input and decides whether to keep it or discard it. In the end, the context is also used as guidance information to help generate the descriptions and predict the bounding box offsets.

A close look at the method in [14] reveals that the entire algorithm carries out an encoding-decoding procedure. In the encoding procedure, the representations of each contextual object fused with its CNN feature and geometry features (relative coordinates) are encoded step by step with a guidance LSTM, where the guidance information is composed of region features. The output of this procedure is the contextual information denoted as c_i. For the decoding procedure, the authors tried two kinds of caption decoder frameworks, namely COCG and COCD, respectively. Although they both have a caption LSTM for captioning as well as a location LSTM for localization, the main difference between these two decoders is their context decoding architectures. Concretely, COCD adds another LSTM to decode context c_i while COCG removes this LSTM and turns the caption LSTM into a guidance LSTM to decode c_i. In conclusion, as shown in the section of experiments in [14], the COCG framework outperforms the COCD framework and other methods, thus obtaining the state-of-the-art results due to the alleviation of the vanishing gradient problem by the guidance LSTM unit inside.

III. Methodology

In this section, we first briefly describe the popular Transformer architecture, which is a fundamental component of our method. Then, we present the framework of our proposed TDC. Finally, we elaborate on the proposed ROCSU loss adaptation.

A. Preliminary Review of Transformer

1) Scaled dot-product attention: The scaled dot-product attention is a basic component of the Transformer [20] architecture. Given a query $q_i \in R^d$ in all T queries, a group of keys $k_t \in R^d$ and values $v_t \in R^d$, where $t = 1, 2, ..., T$, the output of dot-product attention is the weighted sum of the v_t values. The weights are determined by the dot-products of query q_i and keys k_t. Specifically, k_t and v_t are placed into respective matrices $K = (k_1, ..., k_T)$ and $V = (v_1, ..., v_T)$ [31]. The output from a query q_i is as follows:

$$A(q_i, K, V) = V \frac{\exp\left(K^T q_i / \sqrt{d}\right)}{\sum_{t=1}^{T} \exp\left(k_t^T q_i / \sqrt{d}\right)}, \tag{1}$$

where d is the dimension of q_i and $\sqrt{d}$ is to normalize the dot-product value. To capture detailed features of the input, an additional component called multi-head attention is introduced. The multi-head attention is composed of H parallel partial dot-product attention components, $\{h_j | j \in [1, H]\}$ refer to heads, with each head being independent. The realization of the attention resulting from the multi-head attention (*MA*) is given by:

$$\begin{aligned} MA(q_i, K, V) &= \text{concat}(h_1, h_2, ..., h_H) W^O, \\ h_j &= A(W_j^q q_i, W_j^K K, W_j^V V), \end{aligned} \tag{2}$$

where W_j^q, W_j^K, W_j^V denote the transfer weight matrices q, K, V for h_j. W^O is the weight matrix for each head. All of these weights are learned during training. This formula of attention is generic so that it can represent two kinds of attention according to where its input comes from. Specifically, when the query is from the decoder layer, and meanwhile, both the keys and values come from the encoder layer, it represents the mutual attention due to its cross-module attribute. The second multi-head attention is called self-attention, where the queries, keys, and values keep unchanged in both encoder and decoder.

2) Transformer: We now present the use of the Transformer on top of scaled dot-product attention. The basic unit of the Transformer is multi-head attention with feed-forward layers followed by layer normalization [32]. The feed-forward layers map the output of the multi-head attention layer by two linear projections and a ReLU activate function. The encoder and decoder of the Transformer are composed of multiple basic structures, and usually, their layer numbers are the same. The decoder of each layer takes the output of the corresponding

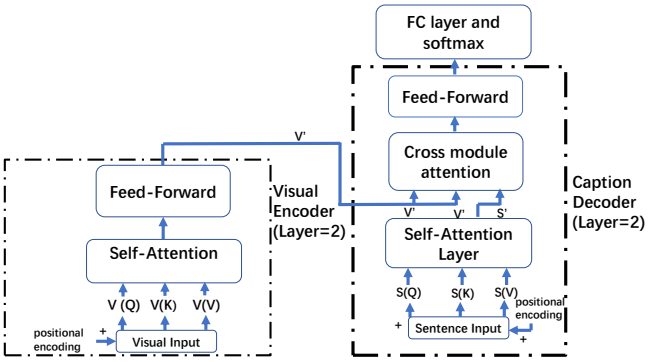


Fig. 4. Transformer structure in our dense captioning scenario, where the layer normalization is omitted.

encoder along with the output of the lower layer decoder output. Self-attention exists in both encoder and decoder. Cross-module attention between encoder and decoder is also applied in the decoder. Residual connection [33] and layer normalization [32] are implemented to all layers. Furthermore, because there is no recurrence module in a Transformer, to indicate positions for each vector, positional encoding (PE) of the input is used. PE occurs at the bottom of the multi-layer Transformer-based encoder and decoder stacks. The dimension of PE is the same as the input, so PE embedding can be added directly to the input. The realization of PE is as follows:

$$\begin{aligned} PE(pos, 2i) &= sin(pos/10000^{2i/d}), \\ PE(pos, 2i+1) &= cos(pos/10000^{2i/d}), \end{aligned} \tag{3}$$

where pos is the position of the embedded vector inside the input matrix, and i is the dimension of the encoded element in the input matrix, d is the total dimension of the input matrix.

B. Transformer in Dense Captioning Scenario

Fig. 4 shows the structure of the Transformer in this dense captioning scenario. To be specific, in the visual encoder, the input is encoded into visual features plus positional encodings, denoted as V. The self-attention layer takes three Vs at the positions of Q, K, V. After the output of the feed-forward layer denoted as V', on the other side, the embedded words plus positional encodings defined as S undergo the same self-attention. At the cross-module attention unit, these two modalities of data interact with each other to gain the output of cross-module attention, which proceeds to feed-forward to learn a captioning probability distribution by fully connected layers and a softmax.

C. Transformer-based Dense Captioner

In this section, we introduce our novel Transformer-based Dense Captioner. Given an image from an image set $\mathbf{I} = \{I_1, I_2, ...I_N\}$, our target is to detect an RoI set, denoted as $\mathbf{R} = \{r_1, r_2, ...r_M\}$ and then describe each of them with corresponding sentence set defined as $\mathbf{S} = \{s_1, s_2, ...s_M\}$. To achieve this goal, our proposed TDC consists of four parts with different functions, namely RoI detector, context module, visual encoder, and captioning decoder, each being elaborated in the following subsections. For ease of explanation, we omit the positional encodings in the following sections.

1) RoI Detector: Inspired by the success of the Faster R-CNN framework in the area of object detection [34], we adopt its Region Proposal Network (RPN) as our RoI detector. This RPN-based RoI detector is trained in an end-to-end manner together with the captioning downstream task to identify whether a region proposal is an RoI to be described. However, our framework not only uses RoI features from RPN; we integrate RoI features with contextual information as introduced in the next sections. Specifically, we use almost the same configuration as [14], however, we replace its backbone structure VGG16 [35] with a ResNet-101 due to its superiority of shortcut structure [33]. In addition, we leverage RoI Align [36] rather than RoI Pooling due to its better performance for small object detection. Via the RoI detector, given an image in $\mathbf{I}$, we get the RoI set $\mathbf{R} = \{r_1, r_2, ...r_M\}$ and its corresponding RoI feature set, denoted as $\mathbf{RF} = \{rf_1, rf_2, ...rf_M\}$.

2) Context Module: According to the data statistics in [14], the description of RoIs has a very close relationship with the objects detected in the image, and therefore, the prior knowledge of object detection can provide useful aids as contextual information for dense captioning. Inspired by this and to obtain such prior knowledge, we pre-trained a Faster-RCNN object detection network on the MS COCO dataset [37] with the same operation as [14]. This is used to create contextual information. In this way, we can gain a set of bounding box coordinates of detected objects $\mathbf{B_{obj}} = \{b_1, b_2, ...b_{obj_N}\}$ with their confidence scores $\mathbf{conf_{obj}} = \{conf_1, conf_2, ...conf_{obj_N}\}$. Additionally, to get features of each bounding box, we extract bounding box and image features with a pre-trained ResNet-152 network because the deeper neural network can capture more local features and it is more suitable for local bounding boxes. We denote corresponding bounding box features as $\mathbf{B} = \{bf_1, bf_2, ...bf_{obj_N}\}$. The image features are defined as $\mathbf{Imgf} = \{Imgf_1, Imgf_2, ...Imgf_N\}$. We also get the geometry information of each object bounding box, namely $\mathbf{G} = \{g_1, g_2, ...g_{obj_N}\}$. Same as [14], g_i, $i \in [1, obj_N]$ is the corresponding coordinate and size ratios of b_i. We only add up class information ahead. Finally, the information is merged together with image features extracted by a pre-trained ResNet-152 network as contextual information for each RoI detected.

3) Visual Encoder: Given the aforementioned visual features consisting of prepared context and RoI information, there is a visual encoder to learn a combined feature representation. We use both visual features (object features) and geometry information (relative bounding box coordinates in an image, and object class label) as context. These two kinds of features are firstly concatenated together as context encoding. Then, for the feature of each RoI detected, the object context encoding from the object detection is concatenated with image features as the final context information. For the context dimension, we first concatenate visual features and geometry features, then we use a linear layer to align the context with the size of RoI features and image features. For a fair comparison with the

state-of-the-art methods, we follow the configuration of [14]. We detect 10 objects for each image. For each RoI detected, we assign the features of these 10 objects as the encoding features of this RoI. First of all, we concatenate $\mathbf{B}$ with $\mathbf{G}$ to get the potential context for each RoI as $\mathbf{BG}$. Then it is allocated to each RoI and thus we get a context matrix denoted as $\mathbf{C} \in R^{M\times obj_N\times(d_F+d_G)}$, where d_F is the dimension of features and d_G is the dimension of geometry information. Because of the different dimensions of object features and RoI features, to align with the image and RoI features and eventually fuse the context information, a linear mapping from $R^{d_F+d_G}$ to R^d is formulated into:

$$\mathbf{C_{align}} = \mathbf{W_c}\mathbf{C} + \mathbf{b}, \tag{4}$$

where $\mathbf{W_c}$ and $\mathbf{b}$ are weight and bias, which can be learned in the linear layer for alignment. After we attain $\mathbf{C_{align}}$, we incorporate it with expanded image feature of given image I_i, whose image feature is $\mathbf{Imgf_i}$ and RoI feature is $\mathbf{Rf_i}$. Finally, we get the visual features $F^0 = (f_1^0, ...f_T^0) \in R^{M\times T\times d}$, $T = 2 + obj_N$ as the input of our visual encoder.

The encoding process is as follows:

$$\begin{aligned}
&V\left(F^l\right) = \varphi(PF(\omega(F^l)), \omega(F^l));\\
&\omega(F^l) = \begin{pmatrix} \varphi(MA(f_1^l, F^l, F^l), f_1^l \\ ... \\ \varphi(MA(f_T^l, F^l, F^l), f_T^l \end{pmatrix};\\
&\varphi(\alpha, \beta) = LayerNorm(\alpha + \beta);\\
&PF(\gamma) = M_2^l \max(0, M_1^l\gamma + b_1^l) + b_2^l,
\end{aligned} \tag{5}$$

where φ is layer normalization on residual output, PF represents the feed-forward unit, which is composed of two linear layers with a nonlinear transformation by an activation function. MA is the multi-head attention that is composed of *H* parallel partial dot-product attention components. ω is the output of assembled multi-head attention with a layer normalization by φ. M_1^l and M_2^l are the weights trained for the feed-forward layers, and b_1^l and b_2^l are corresponding biases. For the $t^(th)$ feature vector encoded inside the representation of an RoI, f_t^l is given as the query to the attention layer and the result is the weighted sum of each f_t^l, $t \in [1, T]$, which processed all the encoded features for an RoI, from global image feature to local RoI feature. Therefore, the output vector can gather the encoded the information from all kinds of features by rating their relationships one by one. In other words, it makes the encoder with a broad horizon so that it can avoid forgetting information with the bigger picture observed.

4) Captioning Decoder: With visual features encoded, the captioning process is as follows:

$$\begin{aligned}
&Y_{\leq t}^{l+1} = \varphi(PF(\omega(Y_{\leq t}^l)), \omega(Y_{\leq t}^l));\\
&\omega(Y_{\leq t}^l) = \begin{pmatrix} \varphi(MA((\delta(Y_{\leq t}^l)_1), F^l, F^l), \delta(Y_{\leq t}^l)_1 \\ ... \\ \varphi(MA((\delta(Y_{\leq t}^l)_t), F^l, F^l), \delta(Y_{\leq t}^l)_t \end{pmatrix};\\
&\delta(Y_{\leq t}^l) = \begin{pmatrix} \varphi(MA(y_1^l, Y^l, Y^l), y_1^l \\ ... \\ \varphi(MA(y_t^l, Y^l, Y^l), y_t^l \end{pmatrix};\\
&p(w_{t+1}|F^0, Y_{\leq t}^L) = soft\max(W_V Y_{t+1}^L),
\end{aligned} \tag{6}$$

where y_i^0 denotes a word token with an embedding dimension W_V, and $Y_{\leq t}^l = (y_1^l, ..., y_t^l)$,w_{t+1} is the probability of vocabulary bank at time step $t+1$. δ is the cross-module attention that uses the current representation of word embedding to attend to the visual representation from the corresponding layer of the encoder. φ represents the self-attention part in the decoder. However, different from the encoder, its inputs are words. It is noted that the restriction of time step means that the attention is only on the already generated words.

D. Training and Optimization

In this section, we introduce the training and optimization details. First, we show the loss function during training. Then in the second subsection, we explain our novel ROCSU.

1) Loss Function: In order to enforce both of the localization of detected RoIs and descriptive captions to be as close as training examples in an end-to-end manner, multiple loss function terms are leveraged during the Stochastic Gradient Descent [38] (SGD) at each training step in a training batch as follows:

$$L = L_{cls} + L_{reg} + \mathbf{rg_{score}} \times \mathbf{L_{caption}}^T, \tag{7}$$

where L_{cls} is the classification binary cross entropy loss function of Faster-RCNN RPN [16] for RoI detection, L_{reg} is the smooth l_1 loss [39] for coordinate regression of the location of detected RoIs. It is notable that $\mathbf{L_{caption}}$ is the cross entropy loss of $P = \{p(w_i|F^0; \theta), i \in [1, max]$, which is the probability distribution of descriptive sentence for RoIs in the RoI batch, and their ground truth sentences word by word. To allocate different weights for each detected RoI according to its importance, we design a module ROCSU, its output is denoted as $\mathbf{rg_{score}}$. We will introduce ROSCU in detail in the next subsection.

2) ROCSU: In this section, we introduce our novel unit ROSCU to measure the region score for each RoI according to its overlap with detected object bounding boxes as follow: Given an RoI r_i in $\mathbf{R} = \{r_1, r_2, ...r_M\}$ and detected object at a training step, the corresponding rg_{score_i} is computed as follows:

$$rg_{score_i} = \mathbf{BW} + IoU(r_i, \mathbf{B_{obj}})\mathbf{conf_{obj}}^T, \tag{8}$$

where $\mathbf{BW}$ is the basic weight preset for each RoI, IoU is the Intersection of Union between r_i and $\mathbf{B_{obj}}$. To assemble all the rg_{score_i} into vector, $\mathbf{rg_{score}}$ can be achieved.

IV. Experiment

In this section, we report and discuss the experiments conducted on three public datasets in order to evaluate the dense captioning performance of our proposed method.

We use the Visual Genome dataset (VG) [40] and the VG-COCO dataset, which is the intersection of VG V1.2 and MS COCO [37], as the evaluation benchmarks. The choice of datasets is the same as the state-of-the-art methods [14], [19] for a fair comparison. The details of each dataset as well as the adopted evaluate metrics are elaborated below:

1) VG: Visual Genome currently has three versions: VG V1.0, VG V1.2, VG V1.4. As the state-of-the-art methods have always used VG V1.0 and VG V1.2, we also conduct our experiments on VG V1.0 and VG V1.2. The training, validation and test splits are chosen similarly as [14], [15], [19]. There are 77,398 images for training and 5,000 images for validation and testing [14].

2) VG-COCO: As demonstrated in [14], the target bounding boxes of VG V1.0 and VG V1.2 are much denser than the bounding boxes in other object detection benchmark datasets such as MS COCO and ImageNet [41]. For example, each image in the training set of VG V1.2 contains an average of 35.4 objects, whilst the average value for MS COCO is only 7.1. To get proper object bounding boxes and caption region bounding boxes for each image, following the configuration in [14], the intersection of VG V1.2 and MS COCO is used in our paper, which is denoted as VG-COCO in which there are 38,080 images for training, 2,489 images for validation and 2,476 for testing.

3) Evaluation Metrics: For evaluation, to comply with evaluation metrics of the state-of-the-art methods, we use the same metric as in [14], [15], [19] called mean Average Precision (mAP). It measures the precision for both localization and description of RoIs. Following the threshold setting in [15], average precision is computed with combinations of different IoU thresholds (0.3, 0.4, 0.5, 0.6, 0.7) for the evaluation of RoI locations and different Meteor [42] thresholds (0, 0.05, 0.10, 0.15, 0.20, 0.25) for the evaluation of language similarity with the ground truth. In the end, the mean value of these APs is the mAP score. For each test image, top boxes with high confidence after non-maximum suppression [43] (NMS) with an IoU threshold of 0.7 are generated. The final results are generated by the second round of NMS under the IoU threshold of 0.5.

A. Implementation Details

The experiments are carried out on Linux Ubuntu Server with an Intel i7-5960X CPU@3.0GHz, 64GB RAM and NVIDIA GTX 2080 Ti GPU. Specifically, in the proposed method, all the image features, RoI features, and object bounding box features consist of 2048 dimensions. The image batch size is set to 1, the detected RoI batch size in a training step is 32, and the maximum iteration is $1,000,000$ for VG-COCO, and $2,000,000$ for VG V1.0 and VG V1.2. The learning rate decrease factor is 0.1 at step $480,000, 640,000, 800,000$ for VG-COCO, and $1,200,000, 1,500,000, 1,800,000$ for VG

TABLE I
The mAP (%) performance of dense captioning algorithms on VG-COCO dataset

Method	mAP(%)
FCLN [15]	4.23
JIVC [18]	7.85
Max Pooling [14]	7.86
COCD [14]	7.92
COCG [14]	8.90
ImgG [14]	7.81
COCG-LocSiz [14]	8.76
COCG&GT [14]	9.79
TDC+ROCSU	**11.58**

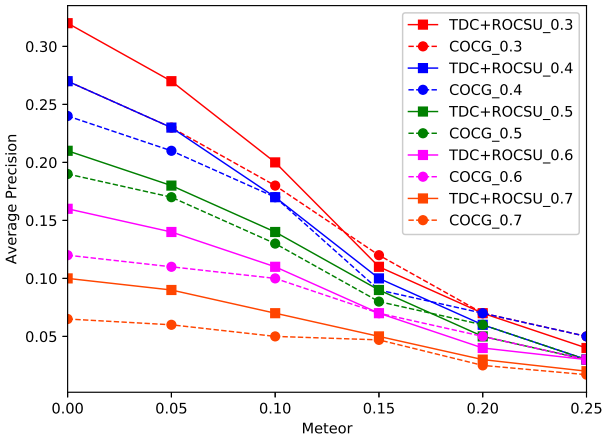


Fig. 5. Average precision with different Meteor scores and different IoU thresholds on the VG-COCO dataset.

TABLE II
The mAP (%) performance of dense captioning algorithms on VG V1.0 dataset

Method	mAP(%)
FCLN [15]	5.39
JIVC [18]	9.31
ImgG [14]	9.25
COCD [14]	9.36
COCG [14]	9.82
CAG-Net [19]	10.51
TDC	10.64
TDC+ROCSU	**11.49**

TABLE III
The mAP (%) performance of dense captioning algorithms on VG V1.2 dataset

Method	mAP(%)
FCLN [15]	5.16
JIVC [18]	9.96
ImgG [14]	9.68
COCD [14]	9.75
COCG [14]	10.39
TDC	10.33
TDC+ROCSU	**11.90**

V1.0 and VG V1.2. The basic learning rate is set to 0.001, momentum is 0.9, and weight decay is 0.0005. The $\mathbf{BW}$ is set as a matrix with all values 0.75.

It is noted that the RoI detector and object detector are trained separately. The RPN based RoI detector is trained online as a part of the entire architecture, while the object detection framework is pre-trained offline. They cannot be trained together because they are designed for different tasks. RPN is trained for selecting potential RoIs. It is a binary classification and regression problem while the object detector is used to create more comprehensive object information. In addition, this kind of training settings keeps the same with [14] for a fair comparison.

TABLE IV
The mAP (%) performance of ablation studies on VG-COCO Dataset

Method	mAP(%)
TDC	11.47
TDC+img+RoI	9.50
TDC+RoI	10.24

TABLE V
The mAP (%) performance of different ROSCU weighting schemes on VG V1.0 dataset

Method	mAP(%)
$ROSCU_Norm$	9.25
$ROSCU_Ones$	9.82

B. Quantitative Results and Analysis

1) Results on VG-COCO Dataset: On the VG-COCO dataset, we conduct extensive experiments to compare our approach and other baseline methods. These baselines are categorized into two groups: state-of-the-art methods including Max Pooling, COCD, COCG, ImgG, COCG-LocSiz and COCG&GT in [14] and earlier methods including FCLN [15] and JIVC [18]. mAP values are provided in Table I. In the following section, we denote our proposed method as TDC+ROSCU, and the method treats each RoI equally without ROSCU as TDC. Table I shows significant improvement in mAP. First of all, compared with the state-of-the-art LSTM method, i.e. COCG, the mAP increases dramatically by about 30%. The gap between TDC+ROSCU is even larger, reaching almost three times the mAP of the FCLN method. The results demonstrate the superiority of TDC+ROSCU, which comes from the broad horizon gained of TDC in encoding and decoding and the focus on informative RoIs from ROCSU. It should be noted that even against ground truth localization of each RoI plus the state-of-the-art method COCG denoted as COCG&GT, TDC+ROSCU still outperforms it by an 18.28% mAP increase.

2) Results on VG V1.0 Dataset: TDC+ROSCU is also evaluated on the VG V1.0 dataset. In order to have a fair comparison with state-of-the-art methods, we adopted the same setting as used in [14], [19]. The mAP results are shown in Table II. It can be seen that TDC+ROSCU outperforms the state-of-the-art methods by a significant margin on this dataset also. Overall, our method achieves a 17% mAP increase against the COCG method [14]. Furthermore, the comparison with CAG-Net in [12] also shows the superiority of TDC+ROSCU, with 9.32% mAP improvements. The improvement is, to a large extent, due to the Transformer in TDC+ROSCU that can provide a broad vision for RoI captioning. In addition, ROSCU can capture more important information. It is also noted that the TDC method by itself also achieves 10.64, which surpasses the state-of-the-art methods. This clearly demonstrates the suitability of the Transformer-based model. On top of that, TDC+ROSCU outperforms TDC by a 0.85 mAP increase, which shows the importance of ROSCU.

3) Results on VG V1.2 Dataset: We also evaluate our proposed TDC+ROSCU method on the VG V1.2 dataset. As with the VG V1.0 experiments, we adopted the same settings as [14], [19]. The mAP results are shown in Table III. It can be observed that the TDC+ROSCU method obtains a relative gain of 14.5% on VG V1.2 with an mAP of 11.90, compared with the state-of-the-art COCG (10.39). It is worth noting that the mAP achieved by our method is more than twice the mAP of the FCLN method. Furthermore, the TDC method without our contributed ROSCU achieves 10.33, which is very close to COCG. However, it is still far (around 15%) from the TDC+ROSCU method, which again shows the effectiveness of ROSCU.

4) AP Values Comparison with Different Threshold Combinations: Fig. 5 shows quantitative comparisons between the baseline (COCG) and TDC+ROCSU. With the Meteor threshold of 0, our TDC+ROCSU method achieves a significant improvement. This is mainly because ROCSU can make the model focus on RoIs with more information. Furthermore, TDC+ROCSU performs better than COCG at nearly all parameters. This shows both the encoding and decoding powers of our TDC and the capability of ROSCU to help the model to grab the important regions.

5) Ablation Studies: To validate the effectiveness of our ROCSU component, we remove it and only leave TDC with the same feature encoding method as TDC+ROCSU, which is denoted as TDC. We can see the value drops by 0.11 due to the equal weights of each RoI allocated during the training stage as the regions that deserve higher priorities are not used.

To validate the function of comprehensive feature encoding, we also propose a wide range of experiment settings. We maintain TDC and adopt different ways of feature encoding. For example, the configuration of image and RoI features with TDC is defined as TDC+image+RoI. It is obvious that with object guidance, the performance improves sharply by 1.93 whilst TDC+Img+RoI achieves even worse results than TDC+RoI possibly because the image features may be too compact to understand, and thus, weaken its own function to guide dense captioning. To better clarify why TDC+ROSCU can achieve better dense captioning ability, we also illustrate an example and analyze the reason in depth in the next section.

Furthermore, to validate the effectiveness of ROSCU setting in Eq. 8. We have also adopted two more kinds of ROSCU weighting schemes. The first one is as follows:

$$ROSCU_{Norm} = \frac{\mathbf{BW} + IoU(r_i, \mathbf{B_{obj}})\mathbf{conf_{obj}}^T}{\sum_{i=1}^{N_{rg}} IoU(r_i, \mathbf{B_{obj}})}, \quad (9)$$

where N_{rg} is the total number of RoIs in the RoI batch, other vectors and actors are the same with Eq. 8. This weighting scheme is denoted as $ROSCU_{Norm}$. It is observed that it only achieves an mAP of 9.25, only 80% of ROSCU when using Eq. 8. This is mainly due to the weakened value by the normalization term, which undermines the function of ROSCU.

Another weighting scheme we adopted is denoted as $ROSCU_{Ones}$. It differs from Eq. 8 in the value of **BW**. For $ROSCU_{Ones}$, we adopted a matrix of all ones as the basic weight of each RoI. The performance of $ROSCU_{Ones}$ is better than $ROSCU_{Norm}$ with an mAP of 9.82. It is still lower than ROSCU setting using Eq. 8, which demonstrates the superiority of the chosen ROSCU score function.

C. Qualitative Results and Analysis

In this section, we show qualitative results and analysis to help evaluate the experimental results in a more subjective way. In the first subsection, we present four examples from VG-COCO, VG V1.0 and VG V1.2 dataset respectively with the visualisation of all RoIs and the descriptions of them. In the second subsection, we will display results, in comparison with the COCG method and also the provided ground truth.

1) Examples of RoIs and Captions by TDC+ROCSU: Four complete examples of dense captioning results by TDC+ROCSU targeted on an image are shown in Fig. 6. From this visualization, we can clearly see the decent quality of both localizations and captions of RoIs achieved by TDC+ROCSU. To begin with, the model is able to capture the grammar of natural languages fairly well. A majority of the generated sentences comply with plain English grammar recognised by humans and are completely readable and understandable. We should owe this to the powerful encoding and decoding capability to learn representative features in order to correspond with visual and language clues as well as be aware of intra-modality connections with each other. Furthermore, it is easy to see the proposed model has a very good command of commonly used ways of description (e.g., in the first example, 'with structure' is used three times correctly. This attributes to the function of ROCSU. Its aim is to attend more on RoIs overlapping more with objects. As we all know, 'with structure' can easily bridge multiple entities together so it is more likely to occur in the RoIs with more attention. Hence, a good command of 'with structure' complies with the doctrine of ROCSU.

2) Ablation Studies: To have a discussion about the experimental results of TDC+ROSCU and TDC in depth, in this section, we will analyze the importance of each part of our contributions, TDC and ROSCU separately. To be specific, we provide the top-5 visualization results according to region confidence of both TDC+ROSCU and TDC methods with the object detection results in the same image from VG-COCO as shown in Fig. 7 although we have given quantitative analysis in the last section.

In Fig. 7, it is clear that due to the power of TDC to process sequential data, both methods can generate decent captions for common regions in the dataset that only describe the action of a person ('a man skiing on the mountain in the middle' and 'trees covered in snow' at the left top of the image). The only difference is TDC+ROSCU provides the 'pine trees', which is more detailed. The region almost has no overlap with the objects detected in (c) and according to Eq. 8, there are no extra weights on this region while training by TDC+ROSCU. However, these good results, to a large extent, come from plenty of training samples from images with similar scenes in the dataset.

Furthermore, there are two examples showing that ROSCU works better if the given region has more overlaps with objects, thus enabling the ROSCU to give more priority to this region even though it is focusing on more detailed information. Specifically, ROSCU helps the machine to recognise the red hat for the orange region instead of the helmet in the results of the TDC method due to more weights allocated to (a) during the training stage than (b), which derives from more overlaps with objects (specifically the IoU with the person with a score of 1.000) in (c) than in (b). Also, based on the same explanation, with the aid of ROSCU, it can benefit from the bigger weight so that it is relatively easier to recognise the colour of the jacket (yellow not brown and yellow in (b)) on the man.

Finally, from the red box in (a), it is easy to observe that for a given region that corresponds with different semantics in the image, ROSCU can show its superiority due to a high weight in the training from the summation of overlap with different detected objects. Because of this, ROSCU can encourage the generation of captions that link different semantics in the image in order to create more comprehensive descriptions that are likely to reveal the theme of the whole image rather than detailed descriptions.

3) Results with COCG and the Ground Truth: Fig. 8 shows the comparison results of our TDC+ROCSU method and the ground truth as a reference to measure their performance in randomly sampled RoIs. From these results, it is also visible that TDC+ROCSU performs better in both localisation and description of RoIs. This can be reflected by higher IoUs and Meteor displayed in the graph. It is noted that TDC+ROCSU is likely to accurately find the salient semantic in ground truth. It might be due to the joint ability of captioning modelling by both TDCs that learns better feature representation and their relationships and ROCSU, which focuses on RoIs that have more overlaps with semantic objects. We argue that it is not proper to owe this superiority to a unique module. For instance, in the first subfigure, without TDC, the close relationship between object surfboard and woman cannot be perfectly built up. Instead, it may suffer from the forget shortcoming like LSTM methods, losing the guidance from the word surfboard. Without ROCSU, this kind of informative RoI may not gain a priority, therefore causing a decrease in performance.

V. Conclusion

In this paper, a novel end-to-end trainable Transformer-based Dense Captioning Captioner (TDC) was proposed to facilitate the encoding and decoding of both visual and language features. This TDC can encode and decode both visual

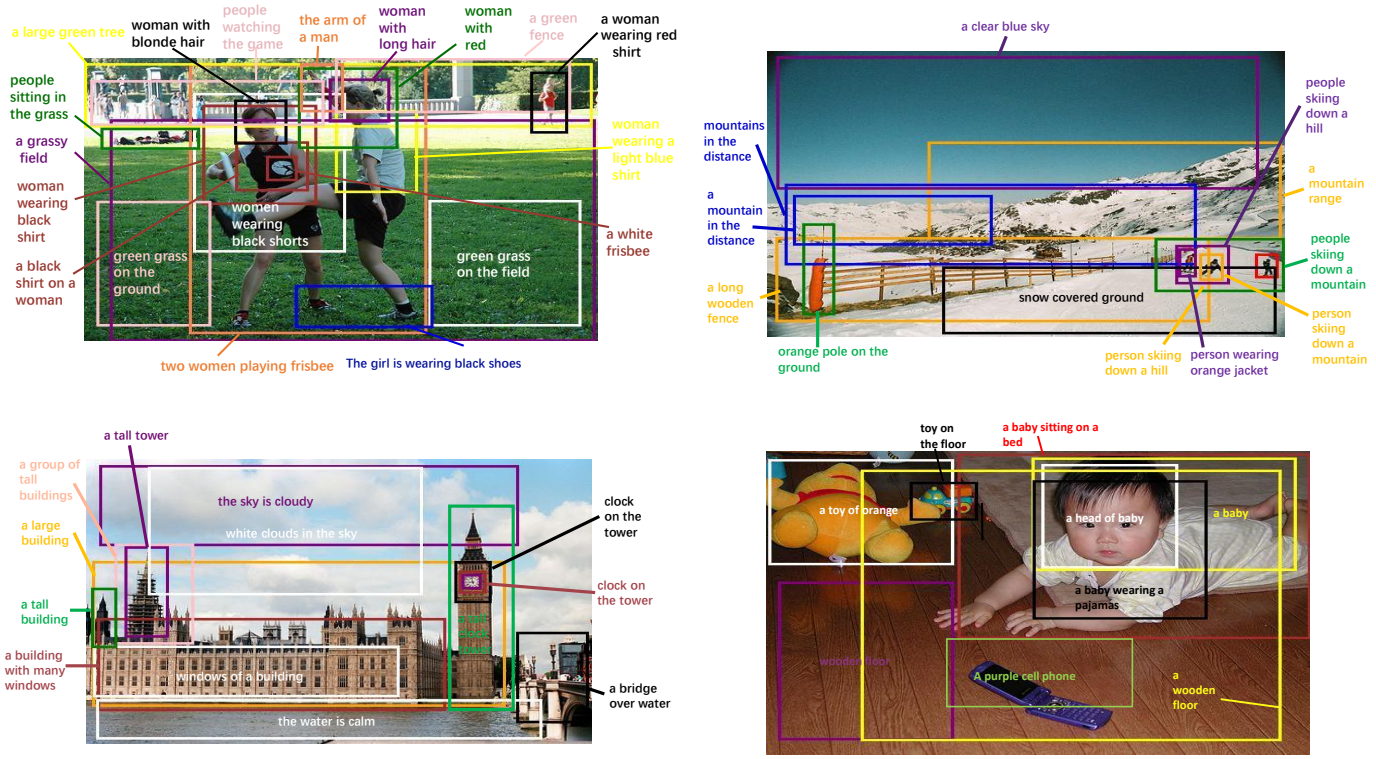


Fig. 6. Detected RoIs with their corresponding captions by TDC+ROCSU of three different datasets: VG-COCO, VG V1.0 and VG V1.2. Specifically, two examples at the top are from VG-COCO, whilst the left bottom one from is VG V1.0 and the right bottom is from VG V1.2.

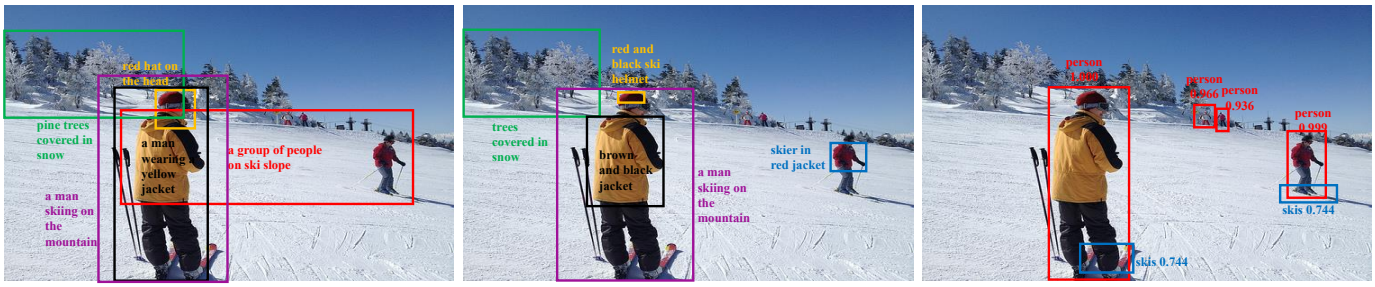


Fig. 7. Dense captioning results of TDC+ROSCU and TDC method on VG-COCO dataset along with their object detection results. (a). Dense captioning results of TDC+ROSCU (Top-5 results according to confidence). (b). Dense captioning results of TDC (Top-5 results according to confidence). (c). Object detection results of the same image.

features and language features effectively with the guidance of object detection information. To make the model pay more attention to the detected RoIs with more information, particularly, we proposed another innovative unit, named ROCSU, to measure the importance of an RoI. Doing so allows the model to give higher priority to them, thus learning more useful knowledge. Experiments on several public datasets show that the TDC+ROSCU method outperforms the state-of-the-art significantly. This framework is easily to be transplanted to similar applications due to its flexibility. In our future work, we will apply the proposed TDC+ROCSU to the application of image captioning, dense video captioning [44] etc. though there might be some changes for ROSCU module according to the specific task.

REFERENCES

[1] Y. Miao, Z. Lin, X. Ma, G. Ding, and J. Han, "Learning transformation-invariant local descriptors with low-coupling binary codes," *IEEE Transactions on Image Processing*, vol. 30, pp. 7554–7566, 2021.

[2] Z. R. Khavas, S. R. Ahmadzadeh, and P. Robinette, "Modeling trust in human-robot interaction: A survey," in *International Conference on Social Robotics (ICSR)*. Springer, 2020, pp. 529–541.

[3] J. Cao, Y. Pang, J. Han, and X. Li, "Hierarchical regression and classification for accurate object detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[4] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

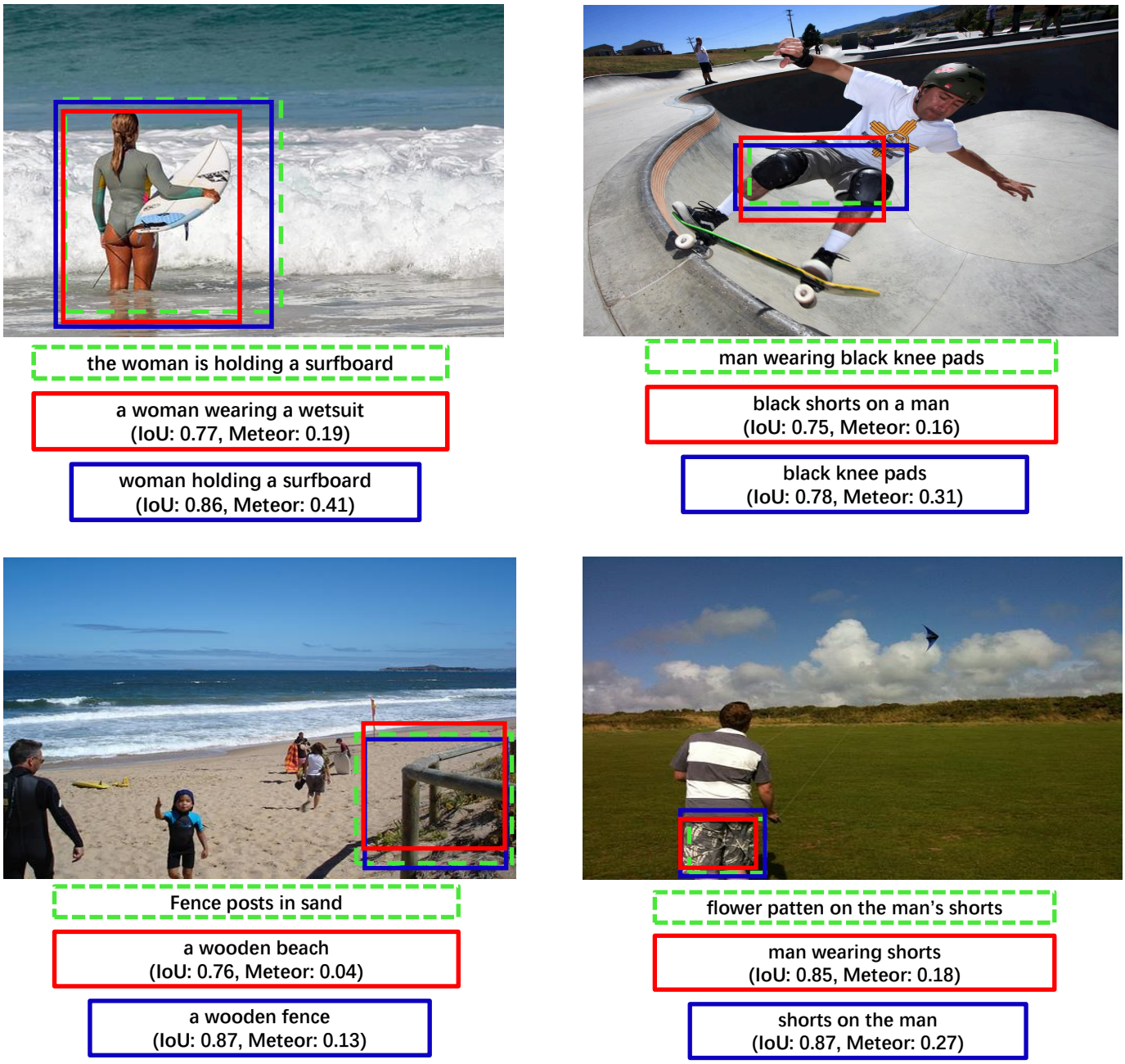


Fig. 8. Qualitative comparisons between baseline (COCG) and our method (TDC+ROCSU). The green box refers to the ground truth, the red box and the blue box are the prediction results of COCG and TDC+ROCSU respectively (Best viewed in color).

[5] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 655–12 663.

[7] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image–text matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5412–5425, 2020.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[9] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.

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

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[12] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4651–

4659.
[13] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 375–383.
[14] X. Li, S. Jiang, and J. Han, "Learning object context for dense captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 8650–8657.
[15] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 4565–4574.
[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 91–99.
[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[18] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2193–2202.
[19] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6241–6250.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
[21] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658–666.
[22] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *International Conference on Learning Representations (ICLR)*, 2015.
[23] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning (ICML)*, 2014, pp. 595–603.
[24] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.
[25] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 11 137–11 147.
[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587.
[27] K. Fu, J. Li, J. Jin, and C. Zhang, "Image-text surgery: Efficient concept learning in image captioning by generating pseudopairs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 5910–5921, 2018.
[28] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3588–3597.
[29] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.
[30] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, 2018, pp. 2556–2565.
[31] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8739–8748.
[32] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[34] J. Cao, Y. Pang, S. Zhao, and X. Li, "High-level semantic networks for multi-scale object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
[35] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1943–1955, 2015.
[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 2961–2969.
[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European conference on computer vision (ECCV)*, 2014, pp. 740–755.
[38] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.
[39] K. Miyaguchi and K. Yamanishi, "Adaptive minimax regret against smooth logarithmic losses over high-dimensional l1-balls via envelope complexity," in *International Conference on Artificial Intelligence and Statistics AISTATS*, 2019, pp. 3440–3448.
[40] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.
[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
[42] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the second workshop on statistical machine translation*, 2007, pp. 228–231.
[43] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *International Conference on Pattern Recognition (ICPR)*, vol. 3, 2006, pp. 850–855.
[44] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017, pp. 706–715.

Zhuang Shao is currently a Ph.D candidate with Warwick Manufacturing Group at University of Warwick, Coventry, UK. He holds a BEng in Electronic & Information Engineering (Northwestern Poly-technical University, 2015), an MSc in Information & Communication Engineering (Tianjin University, 2018). His research interests include image captioning, video captioning and machine learning.

Jungong Han is currently a Chair Professor and the Director of the Research of Computer Science, Aberystwyth University, U.K. He also holds an Honorary Professorship with the University of Warwick, U.K. His research interests include computer vision, artificial intelligence, and machine learning.

Demetris Marnerides has previously worked as a Research Fellow at the Warwick Manufacturing Group (WMG), University of Warwick. He holds a BA in Physics (University of Cambridge, 2013), an MSc in Scientific Computing (University of Warwick, 2015), and a PhD in Engineering (University of Warwick, 2019). His research topics include Machine Learning, Computer Vision, Image Processing and HDR Imaging.

Kurt Debattista is Professor at WMG, University of Warwick. He holds a PhD from the University of Bristol. His research has focused on high-fidelity rendering, high-dynamic range imaging, applications of vision, and applied perception.