
Article:
Shao, Z., Han, J., Debattista, K. et al. (1 more author) (2023) Textual context-aware dense captioning with diverse words. IEEE Transactions on Multimedia, 25. pp. 8753-8766. ISSN 1520-9210

https://doi.org/10.1109/tmm.2023.3241517

Textual Context-Aware Dense Captioning with Diverse Words

Zhuang Shao, Jungong Han, Kurt Debattista, Yanwei Pang

***Abstract*—Dense captioning generates more detailed spoken descriptions for complex visual scenes. Despite several promising leads, existing methods still have two broad limitations: 1) The vast majority of prior arts only consider visual contextual clues during captioning but ignore potentially important textual context; 2) current imbalanced learning mechanisms limit the diversity of vocabulary learned from the dictionary, thus giving rise to low language-learning efficiency. To alleviate these gaps, in this paper, we propose an end-to-end enhanced dense captioning architecture, namely Enhanced Transformer Dense Captioner (ETDC), which obtains textual context from surrounding regions and dynamically diversifies the vocabulary bank during captioning. Concretely, we first propose the Textual Context Module (TCM), which is integrated into each self-attention layer of the Transformer decoder, to capture the surrounding textual context. Moreover, we take full advantage of the class information of object context and propose a Dynamic Vocabulary Frequency Histogram (DVFH) re-sampling strategy during training to balance words with different frequencies. The proposed method is tested on the standard dense captioning datasets and surpasses the state-of-the-art methods in terms of mean Average Precision (mAP).**

***Index Terms*—Dense Captioning, Enhanced Transformer Dense Captioner, Textual Context Module, Dynamic Vocabulary Frequency Histogram**

I. Introduction

Dense captioning originates from image captioning [1]. Rather than generating a single caption for the entire image, dense captioning aims to detect objects in images and describe them in natural language. Thanks to local region descriptors that provide rich and dense semantic labeling of the visual elements, dense captioning can benefit other tasks, including visual question answering [2], image segmentation [3].

Most existing image captioning methods adopt an encoder-decoder architecture, which was inspired by the successful transfer of sequence to sequence training used for machine translation [4] in earlier years. To be more specific, a Convolutional Neural Network (CNN) acts as an encoder, extracting features of a given image before the features are decoded by a trainable Recurrent Neural Network (RNN). However, the resulting captioning algorithms based on simple encoder-decoder frameworks do not prioritise the important parts of feature maps. Therefore, many follow-up methods aim to address this issue. Among them, [5] proposed aligned high-level information while [6], [7] resorted to different forms of attention to learn a group of weights to give more priorities according to the importance of feature maps. On top of attention mechanisms, further work has steered these advances along two orthogonal directions to improve the overall performance of image captioning. First, rapid progress of the Transformer [8] framework in many computer vision research fields, such as object detection [9], has helped to develop a Transformer-based image captioner. For instance, [10] proposed a Transformer-based structure with grid features to alleviate the semantic noise in attention. Subsequently, [11] proposed RSTNet to integrate spatial information to flatten grid features and adaptive attention to bridge the gap between non-visual signals and words. Second, recent advances in image captioning have increased the diversity and distinctiveness of the generated captions. [12] proposed a framework with context-object split latent spaces to generate more diverse captions for a given image while [13] proposed another metric named *CIDErBtw* to supervise the training in order to increase the distinctiveness between images with a similar theme.

In general, *dense captioning* is more challenging than image captioning due to the higher requirement of attaining more detailed and comprehensive descriptions of a given image. [16] pioneered the dense captioning task and designed a Fully Convolutional Localization Network (FCLN), constituting of a detector to detect all the RoIs and a decoder to generate the text descriptions of them one by one. Subsequently, many other solutions were presented, which can be broadly categorized into two classes: with or without the context encoded in their architectures. At the early stage, the architecture adopted a Faster R-CNN [17] to localize RoIs followed by captioning them independently using a Long Short-Term Memory (LSTM) module [18]. Such a pipeline processed the RoIs independently but under-explored possible contextual information that can be leveraged to improve training. To remedy this situation, [19] integrated the RoI features with image features. This can be regarded as a global context due to the fusion before captioning via an LSTM decoder. However, this global context seems too coarse, thus providing ambiguous clues for the training. To overcome this shortcoming, there have been several methods that explored fine-grained contexts. For example, [15] proposed a non-local similarity graph to mutually interact the target RoI with its neighboring RoIs. Alternatively, with the support of data statistics, [14] made

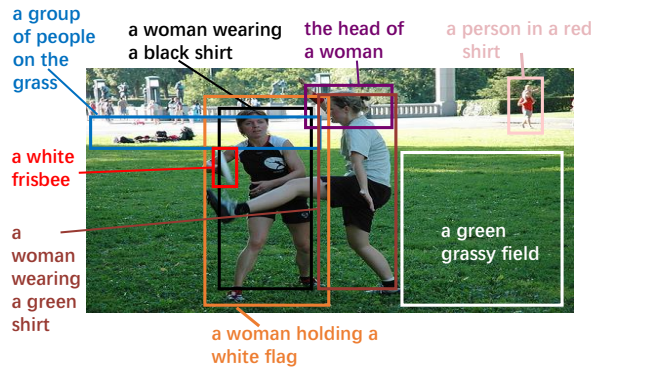
Manuscript received xxx, xxx; revised xxx, xxx and xxx, xxx; accepted xxx, xxx. (Corresponding author: Jungong Han). This research was supported by the funds of China Scholarship Council under Grant No. 201909120012.

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

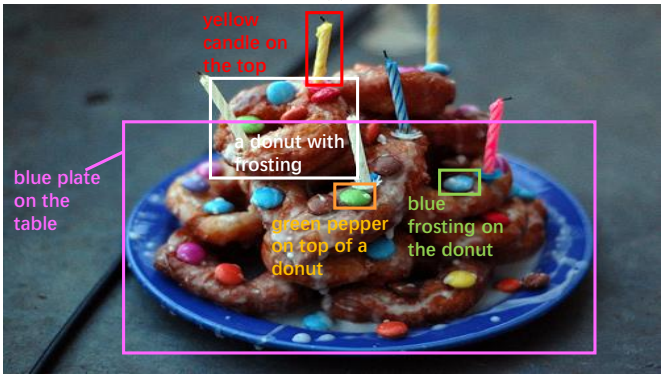
Jungong Han is with the Department of Computer Science, University of Sheffield, S1 4DP, UK (e-mail: jungonghan77@gmail.com).

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

Yanwei Pang is with the School of Electrical and Information Engineering, Tianjin University (e-mail: pyw@tju.edu.cn).



(a)



(b)

Fig. 1: Two examples of neglecting language context during dense captioning from [14] and [15], respectively.

use of the close relationship between RoIs and detected objects via object detection, which naturally considered the contextual information in their architecture. Moreover, inspired by the Transformer architecture, [20] proposed a Transformer-based Dense Captioner (TDC), which also considered different importance of each detected RoI by a Region-Object Correlation Score Unit (ROCSU).

In spite of the limited overall success of the aforementioned methods, developing dense captioning remains incomplete. It is believed that several limitations still exist, in which two of them are particularly critical. First of all, existing methods [14] [15] [20] only take advantage of visual context to guide the captioning process in their decoders but TOTALLY overlook the importance of language context from the surroundings of a detected RoI. In other words, the language information of detected RoIs are processed independently without any interactions, which is inefficient. It is likely to cause some mistakes during caption inference. We demonstrate this via two examples taken from the state-of-the-art methods [20] in Fig. 1. Here, when captioning the last word for the orange RoI in the middle, [14] merely leveraged the visual representation of itself plus the previous embeddings of those predicted words (in this case 'a', 'woman', 'holding','a','white') and the object context trained off line. Thus, they can only guarantee a relatively good grammatical structure but fail to refer to the correct object. This inefficient clue usage leads to the mistake to caption 'flag' instead of the correct answer 'frisbee'. In a similar situation, [15] only took advantage of the visual representation of itself plus the same previous embeddings and the visual features of surrounding RoIs. As shown in Fig. 1b, when inferring the word after 'green' for the orange box, only surrounding visual clues are added as a clue. As a result, it writes the incorrect word 'pepper' instead of 'frosting'. However, it is believed that the language clues from surrounding RoIs are always valuable, thus deserving more attention. Take the same examples, if these two models could look around to see the captioning words of the red RoI in Fig. 1a 'a white frisbee', or the word 'frosting' in the white and green boxes in Fig. 1b, it would significantly help the word inference process, thus improving the accuracy of generated captions. To alleviate this problem, we propose a novel structure, termed Textual Context Module (TCM), which can be integrated into each self-attention layer in the Transformer decoder to selectively capture useful surrounding textual context.

Secondly, previous methods unconsciously employed imbalanced learning due to imbalanced training data. Hence, the learned model tends to output descriptions with words that appeared frequently in training samples only. The consequence is that the diversity of vocabulary learned from the dictionary is rather limited, thus leading to lower language-learning efficiency. Without special treatment, the situation would not be remedied. To demonstrate this issue, we run the context as guidance (COCG) method [14] and noticed that only about 48.13% (940 out of 1953) of the words in the vocabulary bank were learned and appeared in the test period, which indicates both a low learning efficiency and low diversity of test captions. In order to overcome this limitation, we take full advantage of the class information of object context via the extra words in the dictionary and propose a novel Dynamic Vocabulary Frequency Histogram (DVFH) re-sampling strategy to re-balance words with different frequencies during training.

To sum up, the major contributions of this work are fourfold:

- We propose an enhanced end-to-end dense captioning framework based on the Transformer, dubbed Enhanced Transformer-based Dense Captioner (ETDC). A distinct property of ETDC is characterized by taking into account the surrounding textual context and providing more diverse textual words.
- A novel module, named Textual Context Module (TCM), which can be integrated into each self-attention layer in the Transformer, is proposed to select important and useful textual context during word inference.
- We make full use of the class information of object context as the extra words in the dictionary and propose a novel Dynamic Vocabulary Frequency Histogram (DVFH) re-sampling strategy during training to balance words with different frequencies such that the generated captions can be more diverse.
- Extensive experimental results on different datasets show the superiority of the proposed method against the state-of-the-art methods by a wide margin in terms of mean Average Precision (mAP).

The rest of this paper is organized as follows: To begin with, we review the prior works in Section II. Then, in Section III, we present the proposed methodology and expound on the details of ETDC. Extensive experimental results of our proposed method are demonstrated in Section IV with both qualitative and quantitative analysis. Finally, we draw a conclusion and discuss future work in Section V.

II. Related Work

A. Image Captioning

Most of the earlier work solved image captioning via retrieval based methods, designing sets of templates in the retrieval caption pools [21] with straightforward visual feature encoders [6]. However, descriptions of the image to sentences that already exist in the caption pools cannot associate with new objects in all images of a dataset [21]. To remedy this, with the successes of deep learning techniques and the improvement of computer hardware, numerous deep learning methods were proposed. Initially, [22] integrated the use of image-text embedding model and multi-modal sentence generation models with a CNN as the encoder and an LSTM as the decoder and [23] further proposed an reinforcement learning framework. Nevertheless, these works did not consider the spatial information which is crucial for comprehensive and complete description generation during captioning. Therefore, a series of follow-ups focused on the attainment of fine-grained visual and sentence features. In [5], a fine-grained region feature extractor from images was designed by an R-CNN object detector [24] and it produced region-level captions for the given image.

The initial encoder-decoder frameworks, described above, treated each region with equal importance. Doing so fails to focus on the more important part that can provide decisive visual clues for captioning. To tackle this issue, different forms of attention models have been proposed due to their plug-and-play nature. [7] proposed a model on top of semantic attention, which was composed of both top-down and bottom-up attention. Moreover, [25] developed a framework with two Graph Convolutional Networks (GCNs) to explore visual relationships. Recently, the Transformer architecture [8] has brought the advance of Natural Language Processing (NLP) and found application in many computer vision tasks as well. [26] firstly proposed a Transformer-based model for the image captioning task by extracting a single global image feature from the image as well as uniformly sampling features by dividing the image into patches before the feature vectors were input sequentially into the Transformer encoder [8]. Transformer-based solutions also concentrated on improving some hidden properties. [10] brought in grid features to coordinate with RoI features to reduce the semantic noise in the attention mechanism of the Transformer architecture, while [11] proposed RSTNet to integrate spatial information to flatten grid features and adaptive attention to bridge the gap between non-visual signals and words. On the other hand, several prior arts made efforts to improve the diversity and distinctiveness of the captions. [12] proposed context-object split latent spaces to produce captions with more diversities and [13] proposed a new evaluation metric called *CIDErBtw* to supervise the training process in order to improve the distinctiveness between different images with a similar theme.

B. Dense Captioning

To meet the requirement of achieving richer and more detailed descriptions, dense captioning [16] was proposed as a new task that requires an intelligent vision system to both localize and describe multiple salient regions within an image using natural language. Existing dense captioning algorithms can be approximately categorized into two categories: captioning with and without the guidance of contextual information.

1) Dense Captioning Without Context: [16] proposed a framework, which is composed of a prototype of a Region Proposal Network (RPN) in Faster R-CNN as an encoder and an LSTM as a decoder. All the anchors are firstly represented by features of the same size. They are then passed through the RPN and a fully-connected layer to determine if they are foreground (the descriptive region) or background. If an anchor is recognized as foreground, it is named RoI with its corresponding feature. At the same time, the bounding box coordinates of these RoIs are also slightly adjusted via regression. Finally, RoIs are described by an LSTM language model.

2) Dense Captioning With Context: [19] was the first work to add contextual information to guide the dense captioning task, in spite of its high conceptual similarity to [16]. The slight change of the framework lay in that the image feature acted as the global context, and it was input into the caption decoder with RoI features. Even though the added context led to a better performance eventually, this kind of contextual information is global and too coarse to encode fine-grained context information as guidance information.

To capture more fine-grained and detailed context, several subsequent attempts were presented. For example, [15] established a non-local similarity graph to interact a target RoI with its neighboring RoIs. In this scenario, the guidance is the weighted sum of neighboring RoIs. Furthermore, [14] argued that objects in images can provide valuable cues to help locate RoIs and generate descriptions for them via data statistics. Inspired by this, the authors brought in an off-line object detector as guidance information to guide the training of the model. To be specific, the entire algorithm is essentially an encoding-decoding procedure. In the encoding end, the representations of each contextual object fused with its CNN feature and geometry features (relative coordinates) are encoded one by one with a uniquely designed module, termed guidance LSTM. Here, the guidance information is composed of RoI features, which finally obtains the contextual information denoted as c_i. In addition, in the decoding end, the authors attempted to deploy two kinds of caption decoder, namely context as guidance (COCG) and context is decoded with an LSTM (COCD). Although they both have a caption LSTM for captioning as well as a location LSTM for the adjustment of RoI bounding boxes, the main difference is their respective frameworks to decode hidden states.

Moreover, inspired by Transformer architecture, [20] proposed a Transformer-based Dense Captioner (TDC), in which

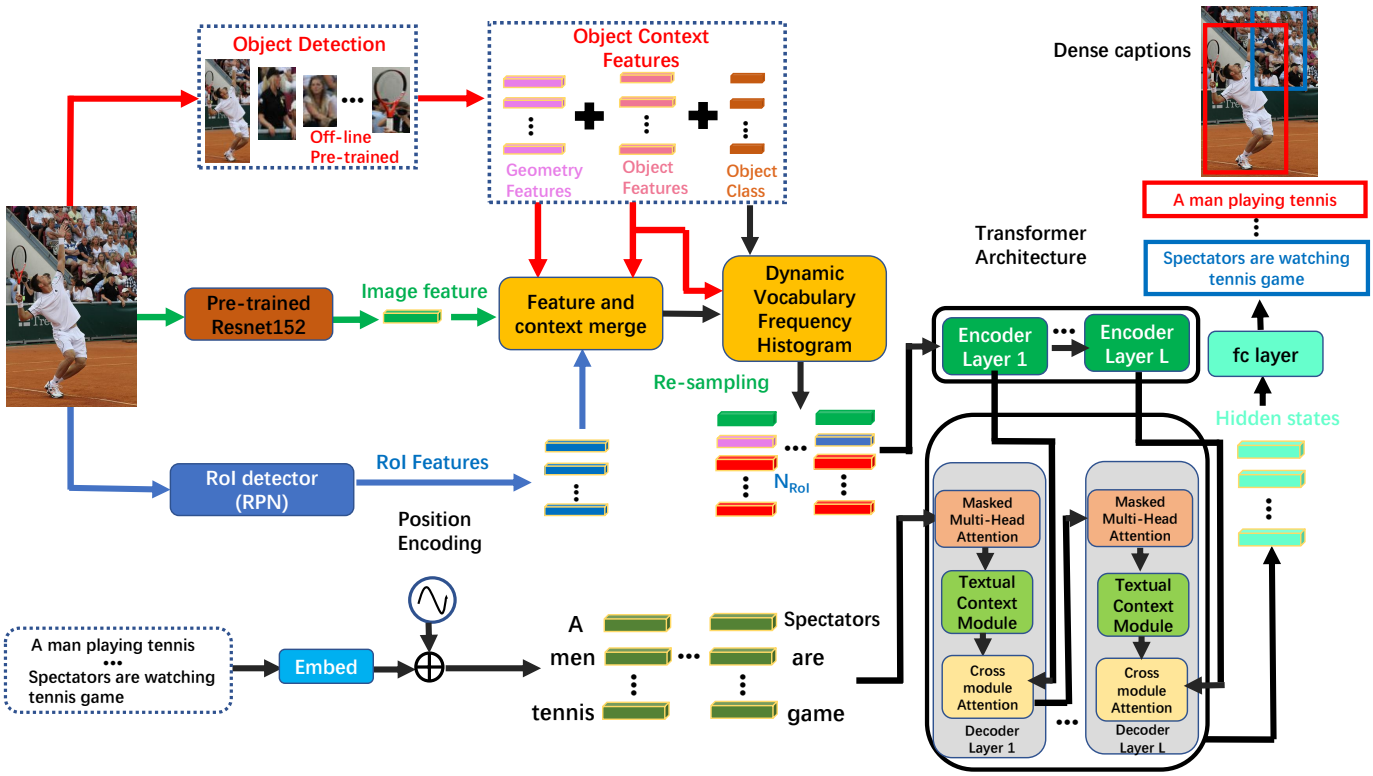


Fig. 2: The proposed ETDC framework is made up of an RoI detector, a feature and context merge module, followed by our Dynamic Vocabulary Frequency Histogram (DVFH), Transformer-based encoder and decoder with a novel Textual Context Module (TCM) inside each decoder layer. Given an image, the RoI detector finds RoIs and the feature and context merge module prepares contextual information generated via the pre-trained object detector for further use. Afterwards, the merged feature with object class information is input into the DVFH for resampling to balance word frequency. Then, the encoder encodes visual information using attention, which provides a visual representation. Finally, after the word embeddings are conducted, visual representation and sentence information are decoded by the caption decoder that integrates language context from surrounding RoIs with TCM to generate dense captions for each RoI.

different importance of each detected RoI was considered and by a Region-Object Correlation Score Unit (ROCSU).

III. Methodology

The overall framework of our proposed ETDC is shown in Fig. 2. Given an image, the Faster R-CNN based RoI detector firstly detects RoIs that are to be described. These are subsequently merged with visual object context from a pre-trained object detector in an offline manner with image features extracted from a pre-trained ResNet-152 network as [20]. This merged visual representation is then input into the proposed Dynamic Vocabulary Frequency Histogram (DVFH) module for re-sampling with object class information. After this, the re-sampled visual features are input into Transformer-based encoder to capture the internal relationships. Together with word embeddings and positional encodings they are then input to the decoder, in which the integrated Textual Context Module (TCM) identifies language context from the surrounding RoIs, eventually predicting sentences for each detected RoI.

In the following of this section, we will first generally review Transformer architecture in the scenario of dense captioning. After it, we explain our proposed Enhanced Transformer-based Dense Captioner. Next, we will introduce the deployment of our Textual Context Module. Then, we will give details of our novel Dynamic Vocabulary Frequency Histogram. Finally, we show our training and optimization details.

A. Preliminary Review of Transformer in Dense Captioning Scenario

Fig. 3 shows the structure of the Transformer [8] in this dense captioning scenario. Generally, following encoder-decoder framework, it consists of two parts, termly visual encoder and caption decoder. The numbers of visual encoder layers and caption decoder layers keep the same with each other, denoted as L. ($L = 2$ in Fig. 3). The Transformer layer setting is an empirical value. It is based on two aspects. On one hand, dense captioning is a compound task, consisting of two complex task: RoI localization and RoI captioning. We found that many works on similar compound task empirically adopted 2 as the Transformer layer number. For example, [27] and [28] for dense video captioning, and [29] for pedestrian search. Another reason for this number is that more Transformer layer stacks are likely to cause out of memory issue

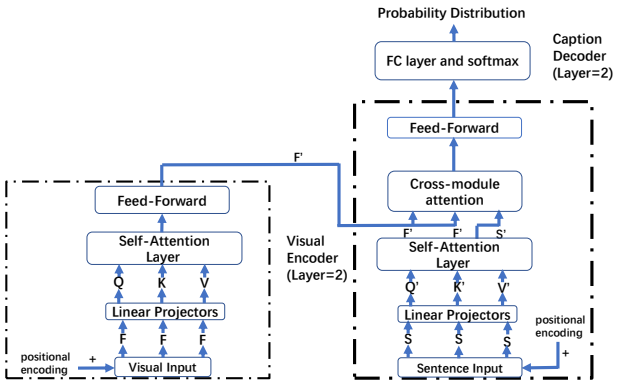


Fig. 3: Transformer structure in our dense captioning scenario, where the layer normalization is omitted.

because the compound tasks are often much heavier than single computer vision tasks.

1) Visual encoder: To be specific, in the visual encoder, visual features added with positional encoding (denoted as F) are first fed in as the input. We adopt positional encoding (PE) procedure in [8] with *sin* and *cos* functions.

It should be noted that PE operation only 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. After the visual features are added with PE, the output is denoted as F, which is input into three linear projectors to attain three different vectors Q, K, V. These three vectors are fed into the visual encoder, the visual encoding procedure is given by:

$$\begin{aligned} V\left(F^l\right) &= \varphi(PF(\omega(F^l)), \omega(F^l)); \\ \omega(F^l) &= \begin{pmatrix} \varphi(MA(f_1^l, F^l, F^l), f_1^l) \\ \cdots \\ \varphi(MA(f_T^l, F^l, F^l), f_T^l) \end{pmatrix}; \\ \varphi(\alpha, \beta) &= LayerNorm(\alpha + \beta); \\ PF(\gamma) &= M_2^l \max(0, M_1^l \gamma + b_1^l) + b_2^l, \end{aligned} \tag{1}$$

where φ is layer normalization [30] on residual output, PF represents the feed-forward layer which consists of two linear layers with a nonlinear activation function in between. 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 bias vectors. F^l is the input of the l^{th} encoding layer. f_t^l is given as the query to the encoding layer and l is the l^{th} encoding layer. Note that F^0 is the aforementioned visual feature F added by positional encodings. MA is a fine-grained component called multi-head attention, which is composed of H parallel partial dot-product attention components. Its realization is as follows:

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

where $\{h_j | j \in [1, H]\}$ refer to the index of each independent head. W_j^q, W_j^K, W_j^V denote the linear linear projectors to the input q, K, V for h_j. W^O is the weight matrix for each head. It is noted that when the query comes from the decoder layer, and both the keys and values are from the encoder layer, it represents cross-module attention. In contrast, if the queries, keys, and values are all from encoder or decoder, this kind of multi-head attention is named self-attention. A is the scaled dot-product attention operation realized by the equation below.

$$A(q_i, K, V) = V \frac{\exp\left(K^T q_i / \sqrt{d}\right)}{\sum_{t=1}^{T} \exp\left(k_t^T q_i / \sqrt{d}\right)}, \tag{3}$$

where $q_i \in R^d$ is a query in all T queries that composes q_i, 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]. d is the dimension of q_i and $\sqrt{d}$ is to normalize the dot-product value.

In the end, with the output of l encoding layers, the encoded visual features, F^l, as a part of the input, is fed into the caption decoder.

2) Caption Decoder: The caption decoder, which is made up of L decoding layers, is formulated as follows:

$$\begin{aligned} S_{\leq t}^{l+1} &= \varphi(PF(\omega(S_{\leq t}^l)), \omega(S_{\leq t}^l)); \\ \omega(S_{\leq t}^l) &= \begin{pmatrix} \varphi(MA((\delta(S_{\leq t}^l)_1), F^l, F^l), \delta(S_{\leq t}^l)_1) \\ \cdots \\ \varphi(MA((\delta(S_{\leq t}^l)_t), F^l, F^l), \delta(S_{\leq t}^l)_t) \end{pmatrix}; \\ \delta(S_{\leq t}^l) &= \begin{pmatrix} \varphi(MA(s_1^l, S^l, S^l), s_1^l) \\ \cdots \\ \varphi(MA(s_t^l, S^l, S^l), s_t^l) \end{pmatrix}; \\ p(w_{t+1} | F^0, S_{\leq t}^L) &= soft\max(W_V S_{t+1}^L), \end{aligned} \tag{4}$$

where $s_i^0, i = 1...t$ stands for a word token with an embedding dimension d_{emb}. $W_V \in R^{V_s \times d_{emb}}$ is the word embedding result for the whole vocabulary bank, where V_s is the vocabulary size. $S_{\leq t}^l = (s_1^l, ..., s_t^l)$ is the predicted words before time step $t+1$. w_{t+1} is the probability of each word in the vocabulary bank at time step $t+1$. δ is the cross-module attention that attends the current representation of word embedding to the visual representation F^l from the corresponding layer of the visual encoder. φ represents the self-attention part in the decoder. However, different from the encoder, its inputs are words. The multi-head attention mechanism of δ and φ *MA* is same with Eq. 2 and Eq. 3 whereas their inputs are different. It is noted that the restriction of time step means that the attention is only on the already generated words.

B. Enhanced Transformer-based Dense Captioner

In this section, we introduce our novel Enhanced Transformer-based Dense Captioner. Given an image from an image set $I = \{I_1, I_2, ...I_N\}$, our target is to detect an RoI set, denoted as $R = \{r_1, r_2, ...r_M\}$ and then describe each of them with a corresponding sentence set defined as $S = \{s_1, s_2, ...s_M\}$. To this end, our proposed ETDC is made

up of five components, namely RoI detector, feature-context merge module, dynamic vocabulary frequency histogram, visual encoder, and caption decoder. We will elaborate other four parts in this section except for dynamic vocabulary frequency histogram module in Section III-D.

1) RoI detector: We adopt the Region Proposal Network (RPN) [32] as our RoI detector. It is trained in an end-to-end manner, together with the captioning task, to identify whether a region proposal is an RoI to be described. It is noted that our framework does not only use RoI features from RPN, but integrates RoI features with contextual information as introduced in the next section. Specifically, we use a configuration similar to [14], however, we replace its backbone structure VGG16 [33] like [20] with a Resnet-101 due to its superior shortcut structure [34]. In this way, given an image in I, we obtain the RoI set $R = \{r_1, r_2, ...r_M\}$ and its corresponding RoI feature set, denoted as $RF = \{rf_1, rf_2, ...rf_M\}$.

2) Feature-context merge module: Discussed in [14], the descriptions of RoIs have a very close relationship with the objects detected in the image. Therefore, the prior knowledge, i.e. 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 R-CNN object detection network on the MS COCO dataset [35] with the same operation as [14] and [20]. This is used to create contextual information and get a fair comparison. This way enables us to obtain a set of bounding box coordinates of detected objects $\mathbf{B_{obj}} = \{b_1, b_2, ...b_{obj_N}\}$ with their confidence scores $\mathbf{conf_{obj}} = \{conf_1, conf_2, ...conf_{obj_N}\}$ and object features $\mathbf{of} = \{of_1, of_2, ...of_{obj_N}\}$. 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 $\mathbf{B} = \{bf_1, bf_2, ...bf_{obj_N}\}$. To simultaneously provide global features, the image features are extracted by the same pre-trained ResNet-152 network and 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 denoted as $\mathbf{cls} = \{cls_1, cls_2, ...cls_{obj_N}\}$.

With the aforementioned visual features constituting prepared context and RoI information, it is the role of the feature-context merge module to merge them. 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 $\mathbf{C} \in R^{M \times obj_N \times (d_F + d_G)}$, d_F and d_G are the dimensions of features and geometry information. Because of the different dimensions of object features and RoI features, to align with the image and RoI features and fuse the context information, a linear mapping from $R^{d_F+d_G}$ to R^d is generated as follows:

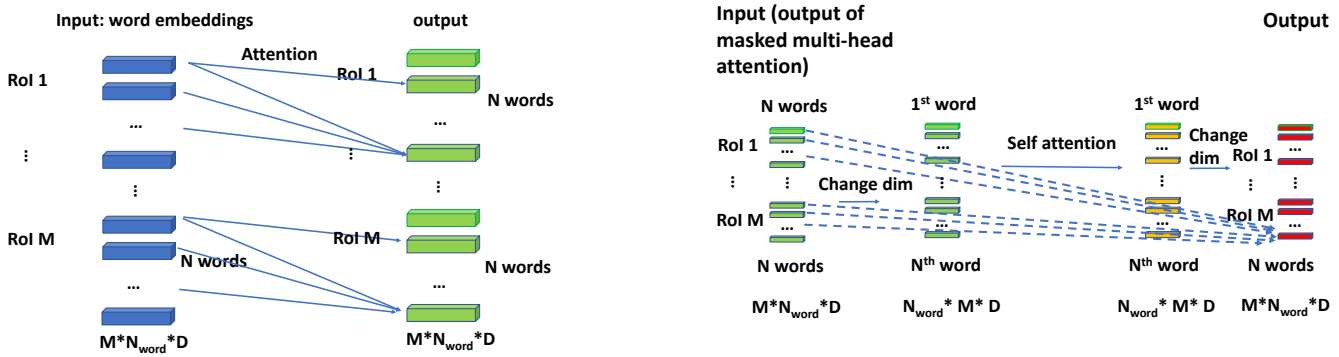
$$\mathbf{C_{align}} = \mathbf{W_c C} + \mathbf{b}, \tag{5}$$

where $\mathbf{W_c}$ and $\mathbf{b}$ are weight and bias, which can be learned in the linear layer for alignment. After we attain C_{align}, we incorporate it with expanded image feature of a given image I_i, $\mathbf{Imgf_i}$ and RoI feature $\mathbf{Rf_i}$. 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.

3) Visual Encoder and Caption Decoder: Our visual encoder and caption decoder are based on the Transformer architecture in [8]. We also applied the Multi-Head mechanism into our architecture as shown in the last section. Given the merged visual features $F'^0 = (f_1'^0, ...f_T'^0)$, which is the re-sampled visual feature from DVFH (will introduce in III-D), the visual encoder tries to learn a best mapping from F'^0 to $V = (v_1, v_2, ...v_L)$ via attention mechanism, where the subscript is the index of the encoding layer and L is the total layer of the visual encoder. Similar to the visual encoder, the caption decoder first takes the word embeddings denoted as $\mathbf{E} = \{e_1, e_2, ...e_M\} \in R^{M \times Len \times d_{emb}}$, where M is the RoI number, Len is the length of sentences, and d_{emb} is the embedding dimension, and feeds it to masked multi-head attention layer in which an upper triangular mask is used to avoid the exposure of the word information at the inference time stamp and onwards. After this masked multi-head layer, the output is sent to our proposed TCM (discussed in III-C) to gain language context from surrounding RoIs, the output size of the TCM is the same as $\mathbf{E}$. Finally, the output of TCM and the corresponding layer visual feature from V are fed into the cross-module attention module to gain multi-modality features as hidden states to generate sentences for RoIs.

C. Textual Context Module

In this section, we demonstrate our novel Textual Context Module. The main idea of this module is to capture the language context from surrounding RoIs detected. The inputs of the masked multi-head attention are the word embeddings in the shape of $M \times N_{words} \times d_{emb}$, where M is the RoI number, N_{words} is the fixed word number, and d_{emb} is the word embedding size. Traditional Transformer masked multi-head takes different word embeddings in sentences and deploys a masked self-attention operation to avoid cheating (the machine should NOT encode the word information at the inference time step and onward) according to Eq. 4. The machine can only refer to the word information before the inference time step and the unavailable information is blocked by an upper triangular matrix [8]. After this masked multi-head attention, the output feature is directly sent to cross-module attention to interact with visual features. However, due to the batch-processing mechanism of the Transformer, the attention operation is targeted on the last two dimensions. And a word can only attend to the words before the time step of its corresponding RoI. This limits the horizon of word inference. Therefore, after the output of multi-head masked attention, TCM is proposed to alleviate it. The whole process is given by following steps:



(a) The traditional masked multi-head attention (example of one head)

(b) Our proposed Textual Context Module.

Fig. 4: The comparison of traditional masked multi-head attention and our proposed Textual Context Module.

$$
\begin{aligned}
&S^{l+1}_{\leq t} = \varphi(PF(\omega(Y^l_{\leq t})), \omega(Y^l_{\leq t}));\\
&\omega(Y^l_{\leq t}) = \begin{pmatrix} \varphi(MA(((Y^l_{\leq t})_1), F^l, F^l)), (Y^l_{\leq t})_1) \\ ... \\ \varphi(MA(((Y^l_{\leq t})_t), F^l, F^l), ((Y^l_{\leq t})_t) \end{pmatrix};\\
&Y^l_{\leq t} = \begin{pmatrix} TCM((\delta(S^l_{\leq t})_1) \\ ... \\ TCM((\delta(S^l_{\leq t})_t)) \end{pmatrix};\\
&\delta(S^l_{\leq t}) = \begin{pmatrix} \varphi(MA(s^l_1, S^l, S^l), s^l_1) \\ ... \\ \varphi(MA(s^l_t, S^l, S^l), s^l_t) \end{pmatrix};\\
&p(w_{t+1}|F^0, S^L_{\leq t}) = soft\max(W_V S^L_{t+1}),
\end{aligned}
\tag{6}
$$

where $\delta(S^l_{\leq t})$ is the input and $Y^l_{\leq t}$ is the output of TCM layer, both $\in R^{M\times Len\times d_{emb}}$. Other symbols and operations are the same with Eq. 4 except TCM representing the operation of our TCM layer. Its implementation is as follows:

$$
TCM(S^l_{\leq t}) = MA \begin{pmatrix} \delta(S^l_{\leq t})_1{}^T \\ ... \\ \delta(S^l_{\leq t})_t{}^T \end{pmatrix}^T \tag{7}
$$

Specifically, $\delta(S^l_{\leq t})$ is firstly transposed. We swap the dimension of M and N_{words} of the output of the multi-head masked attention, thus turning the size of $\delta(S^l_{\leq t})$ to $R^{N_{words}\times M\times d_{emb}}$. Latterly, a multi-head self-attention operation is implemented on the dimension of M and d_{emb} according to Eq. 2 to attend the textual information from other surrounding RoIs. After this multi-head self-attention, finally, we resize the shape to the original size again to $R^{M\times N_{words}\times d_{emb}}$ so that it can be further input into the cross-module attention as shown in Eq. 6. In this way, each word feature successfully encodes information from the language information of other paralleled RoIs. For example, the feature of n^{th} word of RoI_M comes from the word feature of first to $(n-1)^{th}$ word feature of RoI_1 to RoI_M not only feature of first to $(n-1)^{th}$ word feature of RoI_M any more and thus enlarging its vision during word inference.

D. Dynamic Vocabulary Frequency Histogram

In this section, we introduce our proposed component Dynamic Vocabulary Frequency Histogram and corresponding re-sampling strategy. We design DVFH to dynamically record the used frequency of words in the dictionary during training. The main idea of this memory module is to fully take advantage of object context and class information to relatively increase the number of training samples with infrequently-used words. The inputs of DVFH are the merged feature class information previously denoted as $\mathbf{cls} = \{cls_1, cls_2, ...cls_{obj_N}\}$, object features $\mathbf{of_{obj}} = \{of_1, of_2, ...of_{obj_N}\}$ associated with their confidence scores $\mathbf{conf_{obj}} = \{conf_1, conf_2, ...conf_{obj_N}\}$, the merged visual features $F^0 = (f^0_1, ...f^0_T) \in R^{M\times T\times d}$ that concatenate $\mathbf{Imgf} = \{Imgf_1, Imgf_2, ...Imgf_N\}$, $\mathbf{R} = \{r_1, r_2, ...r_M\}$ the object context $\mathbf{C_{align}}$ in III-B and the corresponding sentence set defined as $\mathbf{S} = \{s_1, s_2, ...s_M\}$ for each detected RoI.

At each training iteration, firstly, all the sentences of detected RoIs are marked as frequently-used or infrequently-used RoIs by our designed function dubbed $DVFH_{RoI}$ given $r_i \in \mathbf{R} = \{r_1, r_2, ...r_M\}$, as follows:

$$
DVFH_{RoI}(r_i) = \begin{cases} 1 & \text{if } \frac{1}{N_{word}} \sum_j \frac{\log f_{ij}}{\log F_{max}} \leq \frac{\log F_m}{\log F_{max}} \\ 0 & \text{otherwise} \end{cases}, \tag{8}
$$

where N_{word} is the total word number of s_i, and j is the word index of s_i, f_{ij} is the up-to-the-minute word frequency of the jth word of s_i in the DVFH. F_{max} is the biggest word frequency in the DVFH vocabulary bank, denoted as $DVFB$ and F_m is the median frequency in $DVFB$. If $DVFH_{RoI}(r_i)$ is 1, r_i is an infrequently-used RoI; otherwise, it is a frequently-used RoI. Moreover, similar to RoIs, all the detected objects are marked as frequently-used or infrequently-used objects by our designed function $DVFH_{obj}$ given obj_i, its confidence $conf_i$, and its class cls_i , as follows:

$$
DVFH_{obj}(obj_i) = \begin{cases} 1 & \text{if } \frac{1}{N_{cls}\times conf_i} \sum_k \frac{\log f^o_{ij}}{\log F_{max}} \leq \frac{\log F_m}{\log F_{max}} \\ 0 & \text{otherwise} \end{cases}, \tag{9}
$$

where N_{cls} is the total word number of the class label of obj_i. $conf_i$ is the detection confidence score of obj_i, and k

is the word index of cls_i, f^o_{ik} is the up-to-the-minute word frequency of the kth word of the class label in $DVFB$. F_{max} and F_m are the same as in Eq. 8. If $DVFH_{obj}$ value is 1, obj_i is an infrequently-used object and a frequently-used object if $DVFH_{obj}$ value is 0.

According to Eq. 8 and Eq. 9, we can gain a list of infrequently-used objects and a list of frequently-used RoIs. We then randomly replace frequently-used RoIs with infrequently-used objects. Specifically, given r_m in the list of frequently-used RoIs and obj_n in the list of infrequently-used object information. The corresponding feature of_{obj_n} is leveraged to replace rf_m concatenated in $F^0 = (f^0_1, ...f^0_T)$. Finally, the corresponding sentence of r_m, s_m, is replaced by the index of cls_n in the dictionary, and thus getting the re-sampled visual feature $F'^0 = (f'^0_1, ...f'^0_T)$ and corresponding sentence batch $\mathbf{S}' = \{s'_1, s'_2, ...s'_M\}$. In the end of each training step, $DVFB$ is updated by adding the word frequency of the words appears in $\mathbf{S}'$ into itself.

E. Training and Optimization Details

In this section, we show our training and optimization details. In order to enforce both the localization of detected RoIs and descriptive captions to be as close as training examples in an end-to-end manner, multiple loss function items are leveraged during the Stochastic Gradient Descent [36] (SGD) at each training step in a training batch as follows:

$$L = L_{cls} + L_{reg} + L_{caption}, \tag{10}$$

where L_{cls} is the classification binary cross entropy loss function of Faster R-CNN RPN [17] for RoI detection, L_{reg} is the smooth l_1 loss [37] for coordinate regression of the location of detected RoIs. It is notable that $L_{caption}$ is the cross entropy loss of $P = \{p(w_i|F^0; \theta), i \in [1, max]$, which is the probability distribution of descriptive sentences for RoIs in the RoI batch, and their groundtruth sentences word by word.

IV. Experiment

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

A. Datasets and Evaluation Metrics

We adopt the Visual Genome dataset (VG) [38] and the VG-COCO dataset [14], which is the intersection of VG V1.2 and MS COCO [35], as the evaluation benchmarks. The selection of datasets is the same as the state-of-the-art methods [14], [15] to attain a fair comparison with them. The detailed descriptions of each dataset, as well as the main evaluation metrics, are elaborated below:

1) Visual Genome (VG): For fair comparisons, we also conduct our experiments on VG V1.0 and VG V1.2, which are same with the state-of-the-art methods. The training, validation and test splits are chosen similarly as [14]–[16]. There are 77,398 images in the training split and 5,000 images in validation and test split, respectively.

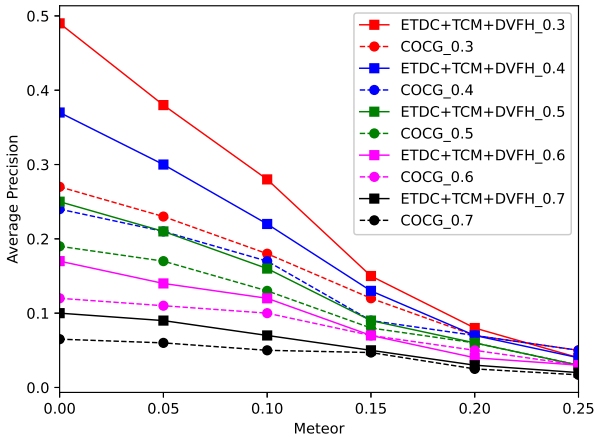


Fig. 5: Average precision with different Meteor scores and different IoU thresholds on the VG-COCO dataset.

2) VG-COCO: As elaborated 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 [39]. To achieve proper object bounding boxes and caption region bounding boxes for each image, following the configuration in [14], the intersection set 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]–[16], [20] called mean Average Precision (mAP). It measures the precision of both localizations and captions of RoIs. Following the threshold configurations in [16], 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 [40] thresholds (0, 0.05, 0.10, 0.15, 0.20, 0.25) for the evaluation of language similarity with the ground truth. With each group of thresholds, the Average Precision (AP) can be calculated. Finally, the mean value of these APs is the mAP score. For each test image, top boxes with high confidence after non-maximum suppression [41] (NMS) with an IoU threshold of 0.7 are generated. In the end, the results are generated by the second round of NMS under the IoU threshold of 0.5.

B. Implementation Details

These experiments are carried out on an NVIDIA GTX 2080 Ti GPU with a memory of 11GM. For the proposed method, all the image features, RoI features, and object bounding box features have a dimension of 2048. 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 m on the VG-COCO dataset, and 2m on VG V1.0 and VG V1.2 datasets. The initial learning rate is 0.001 and the decrease factor is 0.1 at the steps of $480k, 640k, 800k$ on VG-COCO, and $1.2m, 1.5m, 1.8m$ on

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

Method	mAP(%)
FCLN [16]	4.23
JIVC [19]	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 [20]	11.58
ETDC(VGG16)	12.28
ETDC+TCM+DVFH	**14.30**

TABLE II: The mAP (%) performance of dense captioning algorithms on VG V1.0 dataset and VG V1.2 dataset

Method	VG V1.0 mAP(%)	VG V1.2 mAP(%)
FCLN [16]	5.39	5.16
JIVC [19]	9.31	9.96
ImgG [14]	9.25	9.68
COCD [14]	9.36	9.75
COCG [14]	9.82	10.39
CAG-Net [15]	10.51	—
ETDC(VGG16)	11.31	10.60
TDC+ROCSU [20]	11.49	11.90
ETDC+TCM+DVFH	**13.24**	**12.60**

VG V1.0 and VG V1.2. The momentum factor is set to 0.9, and weight decay is 0.0005.

Furthermore, the RoI detector and object detector are trained separately. The RPN based RoI detector is trained in an online manner as a part of the whole architecture, whilst the object detection network is pre-trained offline. They cannot be trained jointly because they are designed for different tasks. Specifically, RPN is trained for selecting potential RoIs, which is a binary classification and regression problem. While, the object detector is used to provide more comprehensive contextual information to guide the entire dense captioning framework.

C. Quantitative Results and Analysis

In this section, we first display quantitative results and discussions on VG-COCO, VG V1.0 and VG1.2 respectively. Then, we show the effectiveness of our proposed components TCM and DFVH through ablation studies.

1) Results and analysis on VG-COCO Dataset: On the VG-COCO dataset, we conduct extensive experiments to compare our ETDC+TCM+DVFH approach and other baseline methods as shown in Table I. We can make an obvious observation of a significant improvement in mAP for ETDC+TCM+DVGH, reaching 14.30%. Our proposed method yields a 2.72 gain of mAP against the TDC+ROCSU method in [20]. Also, compared with the state-of-the-art LSTM method, i.e. COCG, the mAP of ETDC+TCM+DVFH method increases dramatically by more than 60%. Specifically, to conduct fair comparison with state-of-the-art methods in [14], we also implement our ETDC+TCM+DVFH method under VGG16 backbone, which is the same with [14], denoted as ETDC(VGG16). It can be easily observed that due to the lack of shortcut in ResNet-101 , the quality of visual features dropped, thus causing a slight decrease of mAP (12.28%). However, it still outperforms COCG by 38% and COCG with groundtruth by 25.4%. The better performance of our method against other methods is even more obvious, with the mAP reaching more than three times of the FCLN method. The results demonstrate the superiority of ETDC+TCM+DVFH, which stems from the broadened horizon from TCM during decoding and the DVFH, which resampled infrequently-used training sentences to diversify the captioning training. 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, ETDC+TCM+DVGH still outperforms it by a 46.07% mAP increase.

2) Results and analysis on VG V1.0 and VG V1.2 Dataset: ETDC+TCM+DVFH is also evaluated on the VG V1.0 dataset. The mAP results are shown in the second column of Table II. It can be seen that ETDC+TCM+DVFH achieves an mAP of 13.24 and also outperforms all sorts of prior works by a significant margin on this dataset. To be specific, our method significantly outperforms TDC+ROCSU [20] by 15.23% and the COCG method [14] by around 30%. Furthermore, the comparison with CAG-Net in [7] also shows the superiority of ETDC+TCM+DVFH, with 2.73 mAP improvement, which is, to a large extent, due to the TCM module in ETDC+TCM+DVFH that can supply a broad vision with the help of textual context from other neighboring RoIs during captioning. Moreover, DVFH can balance the frequently-used words and infrequently-used words, thus yielding the aforementioned better result. In addition, due to the replacement of the VGG16 backbone without the shortcut structure in ResNet-101, the performance of our proposed ETDC(VGG16) reduces to 11.31. It is slightly lower than TDC+ROCSU method (11.49) with ResNet-101 backbone. However, it still surpasses the counterpart COCG method by over 15%. It is noticeable that on VG V1.0, the performance gap between our proposed method and the state-of-the-art methods is smaller than VG-COCO. This is possibly because VG V1.0 is much bigger than VG-COCO, thus containing more images with complex scenes and captions, which is more difficult to caption even with the guidance of DVFH and TCM.

We also evaluate our ETDC+TCM+DVFH approach on the VG V1.2 dataset. The mAP results are shown in the third column of Table II. It can be observed that the proposed method ETDC+TCM+DVFH obtains a relative gain of 0.7 and 2.21 against TDC+ROCSU method (11.90) and COCG method (10.39) on VG V1.2 with an mAP of 12.60. It is worth noting that the mAP achieved by our ETDC+TCM+DVFH is more than twice the mAP of the FCLN method, which again shows the effectiveness of TCM and DVFH. Moreover, due to the missing shortcut structure in VGG16 network, the mAP of ETDC(VGG16) on the VG V1.2 drops to 10.60, but still outperforms COCG method (10.39). It is also observed that on VG V1.2, similar to VG V1.0, the advantage of the proposed method decreases a little. This is as a result of similar data distributions of VG V1.0 and VG V1.2 (same image set, with slightly different corresponding captions), thus leading

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

DVFH module	TCM module	mAP(%)
✗	✗	11.47
✗	✓	13.44
✓	✗	13.64
SVFB	✓	13.98
✓	✓	**14.30**

to more images with complex scenes and captions, which is more difficult to caption even with the guidance of DVFH and TCM.

3) AP values comparison with different threshold combinations: Fig. 5 shows the comparisons of average precision between the COCG method in [14] and ETDC+TCM+DVFH. It is easily observed that our proposed method outperforms the COCG method under every group of threshold combinations due to the extra guidance information by the TCM module and the balanced training data from the DVFH re-sampling module. In addition, with small threshold combinations, our ETDC+TCM+DVFH method achieves a significant improvement whereas the improvement shrinks with the increase of threshold combination. This is mainly because of the complicated scenes of this task, thus causing the diversity of potential descriptive ways for a given RoI. However, even though the machine can learn a correct sentence with key information subjectively, it may not literally ensemble the ground truth due to the interference of other samples with similar scenes. Therefore, it may still gain a low objective Meteor score.

4) Ablation Studies: To validate the impact of our proposed ETCM module and DVFH module, we also conduct a wide range of ablation studies. We begin with the very basic model which only maintains ETDC without any modification denoted as the ETDC method. Furthermore, we subsequently show results with the TCM module added to ETDC, denoted as ETDC+TCM. Finally, we integrate the TCM module and DVFH module denoted as ETDC+TCM+DVFH. The results of the ablation studies are shown in Table III. A significant metric increase from 11.47% (ETDC only) to 13.44% by almost 2% is seen when the TCM module is added due to more language context clues absorbed, leading to a broader horizon during captioning. In contrast, the mAP rises to 13.64% (a rise of 2.17%) when only the DVFH is used. The improvement of the metric comes from the alleviation of the unbalanced usage of the words in the dictionary by DVFH. There is a further 0.86% improvement in mAP with the DVFH module integrated, which originates from more balanced training and diverse training samples via re-sampling. To better clarify why ETDC+TCM+DVFH can achieve better dense captioning ability, we also show an example and analyze the reason in depth in the next section.

To validate the impact of our proposed ETCM module and DVFH module, we also conduct a wide range of ablation studies. We begin with the very basic model which only maintains ETDC without any modification denoted as the ETDC method. Furthermore, we subsequently show results with the TCM module added to ETDC, denoted as ETDC+TCM. Finally, we integrate the TCM module and DVFH module denoted as ETDC+TCM+DVFH. The results of the ablation studies are shown in Table III. A significant metric increase from 11.47% (ETDC only) to 13.44% by almost 2% is seen when the TCM module is added due to more language context clues absorbed, leading to a broader horizon during captioning. There is a further 0.86% improvement in mAP with the DVFH module integrated, which originates from more balanced training and diverse training samples via re-sampling. To better clarify why ETDC+TCM+DVFH can achieve better dense captioning ability, we also show an example and analyze the reason in depth in the next section.

To further validate the effectiveness of the DVFH module, we additionally propose a Static Vocabulary Frequency Histogram (SVFH) module with a static vocabulary frequency bank, denoted as SVFB. To build up SVFB, we simply calculate the frequency of every word in the groudtruth captions in the whole training set, which is formulated as follows:

$$SVFH_{RoI}(r_i) = \begin{cases} 1 & \text{if } \frac{1}{N_{word}} \sum_j \frac{\log f^s_{ij}}{\log F^s_{max}} \leq \frac{\log F^s_m}{\log F^s_{max}} \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

where N_{word} is the total word number of s_i, and j is the word index of s_i, f_{ij} is the word frequency of the jth word of s_i in the SVFB, which consists of all the words in the dataset and their frequencies. F^s_{max} is the maximum word frequency in SVFB and F^s_m is the median frequency in the SVFB. If $SVFH_{RoI}$ value is 1, r_i is an infrequently-used RoI and a frequently-used RoI if $SVFH_{RoI}$ value is 0. Moreover, similar to RoIs, all the a detected objects are marked as frequently-used or infrequently-used objects by our designed function $SVFH_{obj}$ given obj_i, its confidence $conf_i$, and its class cls_i , as follows:

$$SVFH_{obj}(obj_i) = \begin{cases} 1 & \text{if } \frac{1}{N_{cls} \times conf_i} \sum_k \frac{\log f^o_{ik}}{\log F^s_{max}} \leq \frac{\log F^s_m}{\log F^s_{max}} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where N_{cls} is the total word number of the class label of obj_i. $conf_i$ is the detection confidence of obj_i, and k is the word index of cls_i, f^o_{ik} is the up-to-the-minute word frequency of the kth word of the class label in the SVFH. F^s_{max} and F^s_m are the same as in Eq. 11. If $SVFH_{obj}$ value is 1, obj_i is an infrequently-used object and a frequently-used object if $SVFH_{obj}$ value is 0. Note that since the SVFH is static and all the word frequencies are calculated before the training, it means no update during the training stage.

According to the experimental results in Table III, when SVFH is used in conjunction with the ETDC and TCM modules, the mAP performance rises from 11.47% to 13.98%, but is still inferior to our proposed ETDC+TCM+DVFH. This is because in DVFH, the vocabulary histogram is dynamic and it is adaptive in accordance to the ongoing training process, whereas in SVFH, it reuses the same pre-defined static vocabulary frequency histogram at each training step. However, at each training step, the RoI training samples are selected randomly. As a result of this, the SVFH only reflects the distribution of word frequency distribution of the whole

dataset, not the sampled data distribution. Therefore, it can't resample and balance the infrequently-used words effectively.

D. Qualitative Results and Analysis

In this section, we demonstrate qualitative results and analysis to gain the subjective evaluation of our ETDC+TCM+DFVH method. In the first subsection, we present two examples from the VG-COCO and VG V1.0 datasets via the visualisation of all RoIs and their descriptions. In the second subsection, we present the qualitative results of ablation studies, which is the comparative result between our proposed ETDC+TCM+DFVH method and ETDC method. Thirdly, we also make a comparison of ETDC+TCM+DFVH, the state-of-the-art method COCG, and the corresponding ground truth provided. Finally, to explore the performance yielded by TCM in depth, we recreated the captioning steps of a given RoI and the corresponding TCM attention weights on the texts of other surrounding RoIs.

1) Visual examples of dense captioning by ETDC+TCM+DFVH: Two examples of dense captioning results by ETDC+TCM+DVFH method are shown in Fig. 6. The example in Fig. 6a is from VG-COCO dataset and another example in Fig. 6b is from VG V1.0 dataset. From these visualization samples, we can clearly observe the decent quality of RoI localizations and RoI captions attained by our proposed ETDC+TCM+DVFH. To start with, it is obvious that our proposed model is able to have a good command of the correct grammar. A large majority of the generated sentences are readable, following the correct plain English grammar and complying with the human understanding. This stems from the guidance of TCM, which is capable of providing language context clues that are from surrounding RoIs when captioning a given RoI.

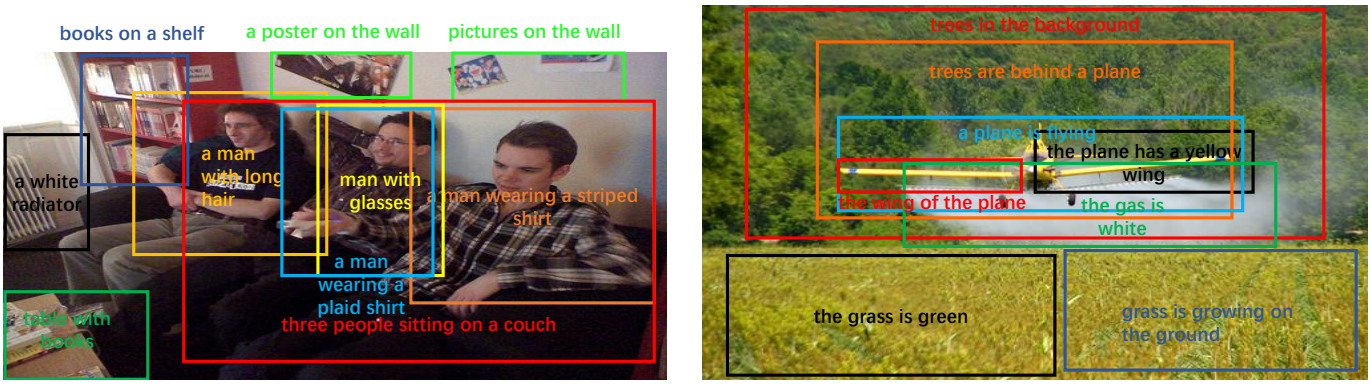
Furthermore, it is clear that the proposed method is proficient with some commonly used means of description, e.g., in the first example. Here, the structure is correctly used three times with correct grammar, and therefore, creating more informative captions with the aid of the TCM module. It can oversee the words of the surrounding context that includes different entities, thus bridging multiple entities together by properly using a with structure.

Finally, ETDC+TCM+DVFH method can adapt to different scenes, no matter whether it is an indoor scene as in Fig. 6a with multiple persons or an outdoor scene with objects as in Fig. 6b. This is because of the re-sampling mechanism of DVFH, which can balance the infrequently-used and frequently-used words in the training samples. Therefore, more infrequently-used words can be exposed more frequently in the training samples, and it is more adaptive to diverse scenes. Specifically, the model successfully describes 'radiator' (in the black RoI box), 'couch' (in the red RoI box) in Fig. 6a, and 'gas' in Fig. 6b as a result of the aforementioned mechanism.

2) Ablation studies: To analyze the experimental results of ETDC+TCM+DVFH and ETDC in depth, in this section, we discuss the importance of each part of our contributions, e.g., TCM and DVFH separately. To be specific, we provide the top-10 results sorted by RoI confidence of both ETDC+TCM+DVFH and ETDC approaches in the same image from VG-COCO as shown in Fig. 7 although we have given quantitative analysis in the last section.

Fig. 7 shows the comparative visualization results between our proposed ETDC+TCM+DVFH method and the method only keeps ETDC but removes TCM and DVFH. We choose the top-10 results according to the confidence scores of RoIs. Generally, the ETDC+TCM+DVFH can create more accurate details and interactions between different entities inside the given image. Firstly, three green boxes show the different extents of interactions with other RoIs. For example, the ETDC method only generates 'blue and white helmet' without any link with the man whereas the ETDC+TCM+DVFH method can identify the connection with the man by generating 'a man wearing a helmet', which is more well-rounded. This is due to the TCM module that can help the model to attend to the words from other RoIs (e.g. the blue one with a caption 'man wearing green jacket' can provide additional language clues for captioning). In addition, from the orange boxes in two graphs, we can see the ETDC+TCM+DVFH method is more capable of creating generalized and precise description that reveals the theme of the image whilst ETDC can only focus on the local description outputting captions including only two people due to the lack of language interactions of different RoIs. Last but not least, ETDC tends to commit a mistake if the appearance of an entity is quite similar to an alternative thing: In Fig. 7b, the model incorrectly outputs 'skateboard' instead of 'snowboard'. This is because 'snowboard' is an infrequently-used word in the dictionary and the training for using this word is inadequate while TDC+TCM+DVFH can gain extra experience of using this word by DVFH re-sampling, thus giving a satisfying result.

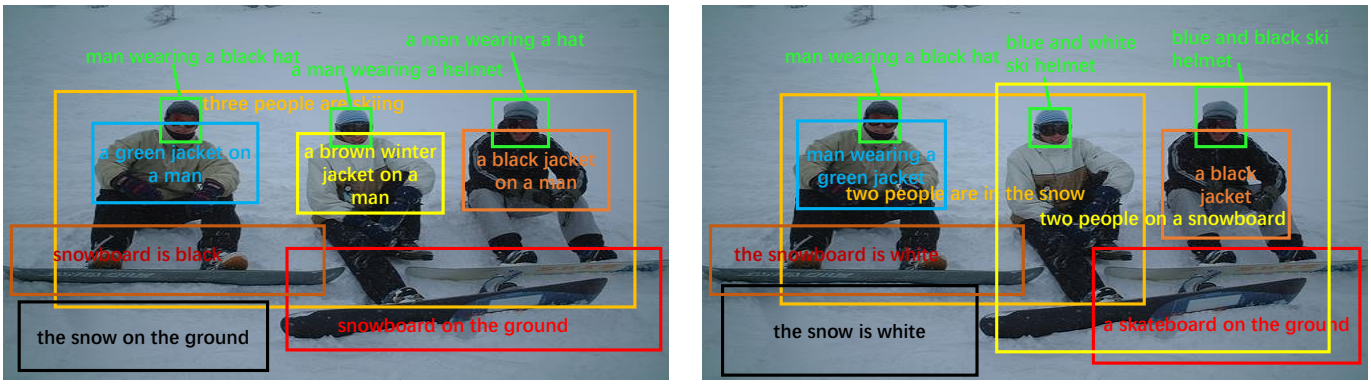
3) Comparative results with COCG method and ground truth: Fig. 8 shows some comparative qualitative results of our ETDC+TCM+DVFH method, the state-of-the-art method COCG and the ground truth as a reference to measure their performances. It is clear that the ETDC+TCM+DVFH method attains better performance in both localisation and description of RoIs due to higher IoUs and Meteor scores shown in the graph. It should be noted that ETDC+TCM+DVFH is likely to outperform the COCG method by a large margin in terms of Meteor language score. Both of our proposed modules TCM and DVFH contribute to this. On one hand, TCM can bring a wider vision during the captioning process of each word in a given RoI by interacting the attention feature with textual context from its surrounding RoIs. For instance, in Fig. 8a, the ETDC+TCM+DVFH method benefits from TCM, thus taking advantage of the textual context of other RoIs and attaining precise caption starts with two men. On the contrary, COCG method can only deduce the caption by visual contextual information and the previous captioned own texts of the given RoI, leading to a relatively bad result. Furthermore, DVFH also gives rise to the better captioning result by the re-sampling mechanism that balances the word frequency in training. For example, in Fig. 8b, since 'pine' is an infrequently-used word, without DVFH, COCG cannot think of it in this given scene, let alone captioning this RoI in a better way with this word. As a result, it can only use a frequently-used word 'trees'



(a) Dense captioning example from VG-COCO.

(b) Dense captioning example from VG V1.0.

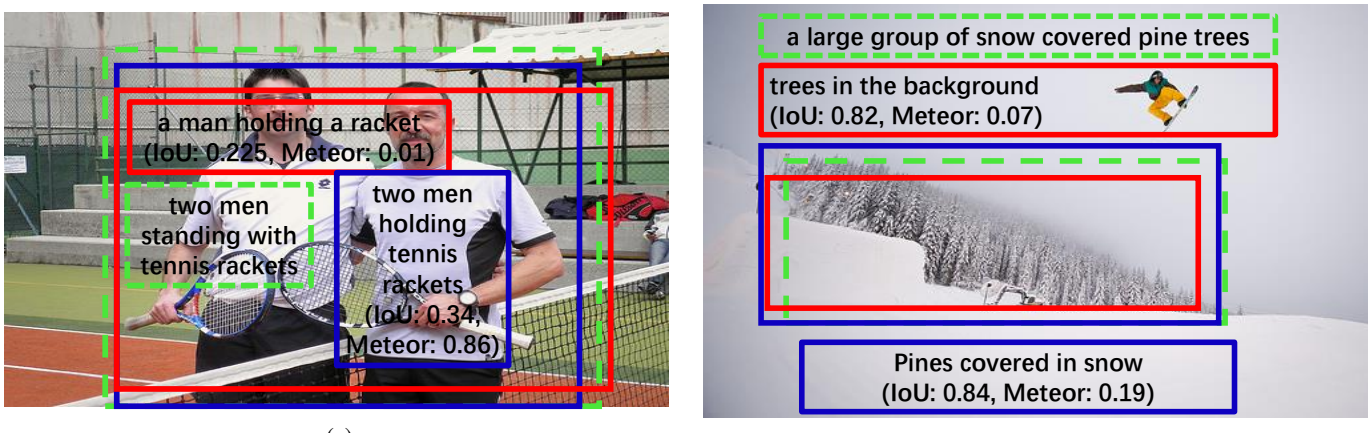
Fig. 6: Two examples of detected RoIs and their captions by ETDC+TCM+DVFH.



(a) The visualization results of ETDC+TCM+DVFH method.

(b) The visualization results of ETDC method.

Fig. 7: The comparative qualitative top-10 results of ETDC+TCM+DVFH and the model that only keeps ETDC according to the RoI confidence scores.



(a)

(b)

Fig. 8: Qualitative results of baseline (COCG) and our proposed method (ETDC+TCM+DVFH). The green dotted box represents the ground truth localization and caption, while the red box and the blue box are the prediction results of COCG and ETDC+TCM+DVFH (Best viewed in color).

to caption it. However, with DVFH, the proposed method succeeds in properly captioning this RoI with the word 'pine'.

4) Captioning Process With and Without TCM: To explore the function of TCM more in depth, we recreate the captioning process for a given RoI with and without the TCM module as shown in Fig. 9. Generally, it is observed that compared with the ETDC method in Fig. 9b, the addition of the TCM module can always select the proper textual context to help with the generation of the word at each step, which shows its effectiveness. To be specific, at step $t = 3$, the created correct word 'two' is originated from the high attention weights on 'a person' in the purple and 'a man' in the light blue surrounding RoIs, which helps the caption decoder to know there are two persons in aggregation in the RoI. In contrast, in Fig. 9b, the wrong word 'a' is generated only by the isolated decoding feature without TCM. We notice that on this occasion, the RoI box is imperfect (fails to include two people together). Therefore, only with the use of the decoding feature, it must be visually confusing, causing a wrong inferring 'a'. Furthermore, in the third column, at the step $t = 5$, the model with TCM successfully attends to the previous captioned word 'skateboard' in the surrounding RoIs indicated by high attention weights whereas the counterpart method is misled by a visual similarity between sitting and skateboarding. Without the TCM module, it cannot see the context clue of other surrounding RoIs and cannot calibrate this mistake, thus creating the wrong word 'sitting'.

V. Conclusion

In this paper, a novel trainable end-to-end Enhanced Transformer-based Dense Captioner (ETDC) was designed to boost the dense captioning performance. To this end, we proposed the Textual Context Module (TCM), to capture surrounding textual context. In addition, we presented a Dynamic Vocabulary Frequency Histogram (DVFH) re-sampling strategy during training to balance words with different frequencies by fully taking advantage of the class information of object context as the alternative infrequently-used words in the dictionary. We tested this plug-and-play method on three different standard dense captioning datasets and the results turn out that our method outperformed the state-of-the-art method by a wide margin in terms of mean Average Precision. Due to its plug-and-play property, in our future work, we may apply it to different tasks such as action recognition [42], image segmentation [3], [43],event detection [44], visual relationship detection [45] and magnetic resonance image reconstruction [10], [46] though there might be some changes on ETDC, DVFH and TCM module according to the downstream task.

References

[1] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 2117–2130, 2019.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[3] S. Qiu, Y. Zhao, J. Jiao, Y. Wei, and S. Wei, "Referring image segmentation by generative adversarial learning," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1333–1344, 2019.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 35, no. 3, 2021, pp. 2286–2293.

[11] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "Rstnet: Captioning with adaptive attention on visual and non-visual words," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 465–15 474.

[12] S. ahajan and S. Roth, "Diverse image captioning with context-object split latent spaces," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] J. Wang, W. Xu, Q. Wang, and A. B. Chan, "Compare and reweight: Distinctive image captioning using similar images sets," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 370–386.

[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] 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.

[16] 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.

[17] 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.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] 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.

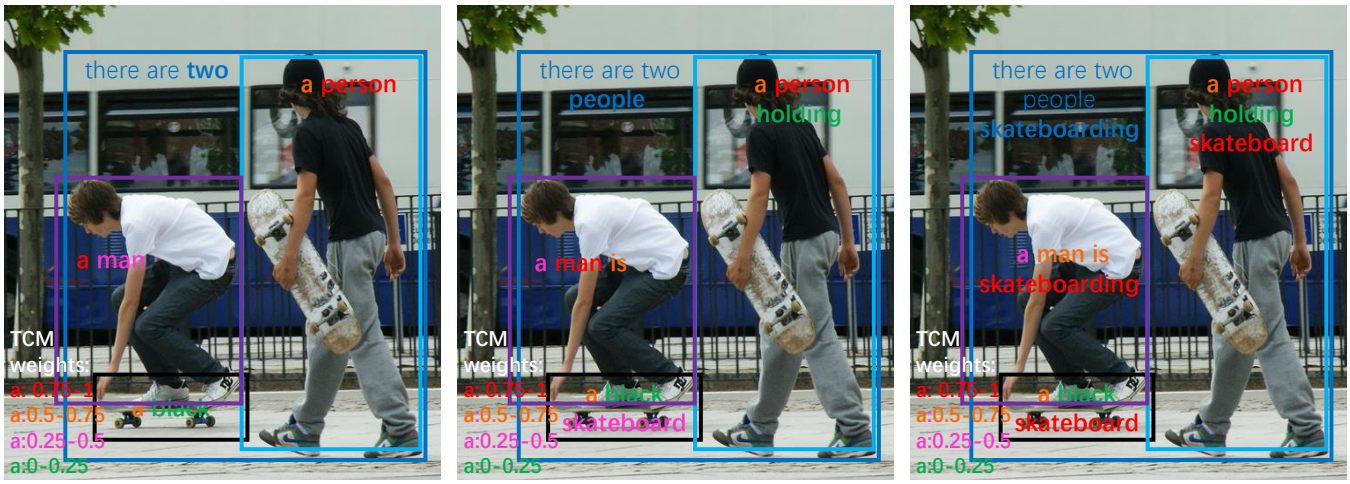
[20] Z. Shao, J. Han, D. Marnerides, and K. Debattista, "Region-object relation-aware dense captioning via transformer," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[21] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image captioning: a comprehensive survey," in *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*. IEEE, 2020, pp. 325–328.

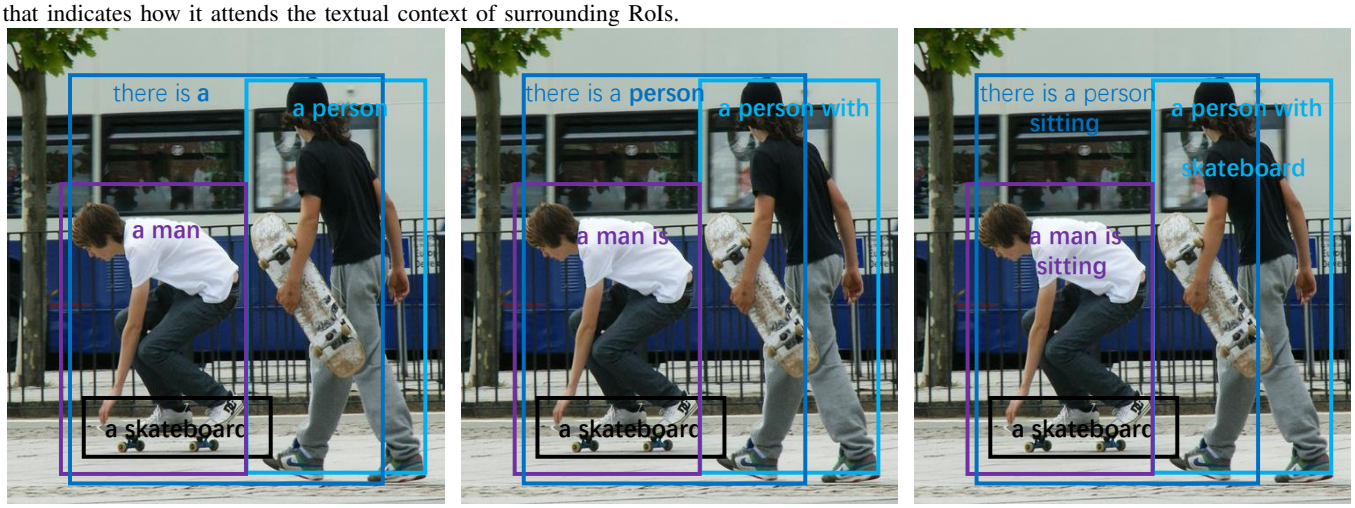
[22] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning (ICML)*, 2014, pp. 595–603.

[23] N. Xu, H. Zhang, A.-A. Liu, W. Nie, Y. Su, J. Nie, and Y. Zhang, "Multi-level policy and reward-based deep reinforcement learning framework for image captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1372–1383, 2019.

[24] 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.



(a) The captioning process for a given RoI (The dark blue one begins with 'there') of ETDC+TCM method at the step of $t = 3$, $t = 4$ and $t = 5$ respectively. The left bottom legend shows the value range of the average multi-head attention weights in TCM (MA in Eq. 6 that indicates how it attends the textual context of surrounding RoIs.



(b) The captioning process for a given RoI (The dark blue one begins with 'there') of ETDC method at the step of $t = 3$, $t = 4$ and $t = 5$.

Fig. 9: The captioning process for a given RoI (The dark blue one begins with 'there') of ETDC+TCM and ETDC, and the word in bold blue is the word generated at the corresponding step.

[25] 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.

[26] 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.

[27] 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*, 2018, pp. 8739–8748.

[28] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-end dense video captioning with parallel decoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6847–6857.

[29] J. Cao, Y. Pang, R. M. Anwer, H. Cholakkal, J. Xie, M. Shah, and F. S. Khan, "Pstr: End-to-end one-step person search with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9458–9467.

[30] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[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. 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.

[33] 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.

[34] 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.

[35] 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.

[36] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[37] 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.

[38] 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.

[39] 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.
[40] 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.
[41] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *International Conference on Pattern Recognition (ICPR)*, vol. 3, 2006, pp. 850–855.
[42] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 102–114, 2016.
[43] M. Gao, F. Zheng, J. J. Yu, C. Shan, G. Ding, and J. Han, "Deep learning for video object segmentation: a review," *Artificial Intelligence Review*, pp. 1–75, 2022.
[44] A.-A. Liu, Z. Shao, Y. Wong, J. Li, Y.-T. Su, and M. Kankanhalli, "Lstm-based multi-label video event detection," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 677–695, 2019.
[45] A.-A. Liu, Y. Wang, N. Xu, W. Nie, J. Nie, and Y. Zhang, "Adaptively clustering-driven learning for visual relationship detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 4515–4525, 2020.
[46] Y. Liu, Y. Pang, X. Liu, Y. Liu, and J. Nie, "Diik-net: A full-resolution cross-domain deep interaction convolutional neural network for mr image reconstruction," *Neurocomputing*, 2022.

Zhuang Shao is currently a Ph.D candidate with Warwick Manufacturing Group at University of Warwick, Coventry, UK. His research interests include image captioning, video captioning and machine learning.

Jungong Han is Chair Professor in Computer Vision at the Department of Computer Science, University of Sheffield, 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. He is the Fellow of the International Association of Pattern Recognition, and serves as the Associate Editor for several prestigious journals, such as IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Circuits and Systems for Video Technology, and Pattern Recognition.

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.

Yanwei Pang (M'07-SM'09) received the Ph.D. degree in electronic engineering from the University of Science and Technology of China in 2004. Currently, he is a chair professor at the Tianjin University, China, and the Founding Director of the Tianjin Key Laboratory of Brain-Inspired Intelligence Technology (BIIT lab), China. His research interests include computer vision, pattern recognition, medical imaging, magnetic resonance imaging, and image reconstruction, in which he has published 150 scientific papers, including 40 articles in IEEE TRANSACTIONS and 30 papers in top conferences (e.g., CVPR, ICCV, and ECCV).